Amber Day

SDS 383D

Statistical Modeling II

Exercises 1: Preliminaries

Due: January 31st

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 $x_1, \ldots, x_N$ 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 that w is given a Beta(a,b) prior distribution:

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

where $\Gamma(\cdot)$ denotes the Gamma function. Derive the posterior distribution $p(w \mid x_1, \ldots, x_N)$.[1]

[1] I offer two tips here that are quite general. (1) Your final expression will be cleaner if you reduce the data to a sufficient statistic. (2) Start off by ignoring normalization constants (that is, factors in the density function that do not depend upon the unknown parameter, and are only there to make the density integrate to 1.) At the end, re-instate these normalization constants based on the functional form of the density.

Since $x_1, \ldots, x_N \sim$ Bernoulli(w) we have the following pmf

$$\Rightarrow p(x_i|w) = w^{x_i}(1-w)^{1-x_i}, \quad x_i = 0,1, \quad w \in (0,1)$$

We begin by calculating the likelihood* $L(w|x_1, \ldots, x_N)$

Since $x_1, \ldots, x_N$ are independent observations the likelihood function is the joint density calculated as follows.

$$L(w|x_1, \ldots, x_N) = p(x_1, \ldots, x_N|w) = p(x_1|w) \cdots p(x_N|w) = \prod p(x_i|w)$$

$$= w^{x_1}(1-w)^{1-x_1} \cdots w^{x_N}(1-w)^{1-x_N} = w^{\sum x_i}(1-w)^{N-\sum x_i} = L(w|x_1, \ldots, x_N)$$

The prior distribution given is: $p(w) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} w^{a-1}(1-w)^{b-1}$

The posterior distribution is proportional to prior $\cdot$ likelihood*

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

Note that this is proportional to a Beta$(\sum x_i + a, N+b-\sum x_i)$ distribution.

Thus the posterior distribution is

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

* The likelihood function of a sample is the joint density $p(x_1, \ldots, x_N|w)$ of the random variables involved but viewed as a function of the unknown parameters given a specific sample of realizations from these random variables.

* Prior distribution of θ, $\pi(\theta) = p(\theta)$

Posterior distribution

$\pi(\theta|x) = p(\theta|x)$

$= \frac{p(x|\theta)p(\theta)}{c} = \frac{f(x|\theta)p(\theta)}{c}$

$= \frac{\prod f(x_i|\theta)p(\theta)}{c} = \frac{L(\theta|x)p(\theta)}{c}$

$\propto L(\theta|x)p(\theta)$

x_i's must be independent

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

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)\,.$$

Suppose that $x_1 \sim \text{Ga}(a_1, 1)$ and that $x_2 \sim \text{Ga}(a_2, 1)$. 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).[2] 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.

[2]Take care that you apply the important change-of-variable formula from basic probability. See, e.g., Section 1.2 of http://www.stat.umn.edu/geyer/old/5102/n.pdf.

Theorem 4.3.2 pg 158 of Casella-Berger statistical inference states:

If (X,Y) is a continuous random vector with joint pdf $f_{X,Y}(x,y)$ then the joint pdf of (U,V) can be expressed in terms of $f_{X,Y}(x,y)$ in a manner analogous to (2.1.8). As before, $A = \{(x,y): f_{X,Y}(x,y) > 0\}$ and $B = \{(u,v): u = g_1(x,y), v = g_2(x,y) \text{ for some } x,y \in A\}$. The joint pdf $f_{U,V}(u,v)$ will be positive on the set B. For the simplest version of this result we assume that the transformation $u = g_1(x,y)$ and $v = g_2(x,y)$ defines a one-to-one tranformation of A onto B. The transformation is onto because of the definition of B. We are assuming that for each $(u,v) \in B$ there is only one $(x,y) \in A$ such that $(u,v) = (g_1(x,y), g_2(x,y))$. For such a one-to-one, onto transformation we can solve the equations $u = g_1(x,y)$ and $v = g_2(x,y)$ for x and y in terms of u and v. We will denote this inverse transformation by $x = h_1(u,v)$ and $y = h_2(u,v)$. The role played by the derivative in the univariate case is now played by the Jacobian of the transformation, defined:

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} = \frac{\partial h_1(u,v)}{\partial u}\frac{\partial h_2(u,v)}{\partial v} - \frac{\partial h_1(u,v)}{\partial v}\frac{\partial h_2(u,v)}{\partial u}$$

we assume that J is not identically 0 on B. Then the joint pdf of (U,V) is 0 outside the set B and on the set B is given by $f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v))|J|$

Lemma 4.2.7: Let (X,Y) be a bivariate random vector with joint pdf or pmf $f(x,y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$ $f(x,y) = g(x)h(y)$.

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

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)\,.$$

Suppose that $x_1 \sim \text{Ga}(a_1, 1)$ and that $x_2 \sim \text{Ga}(a_2, 1)$. 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).[2] 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.

[2]Take care that you apply the important change-of-variable formula from basic probability. See, e.g., Section 1.2 of http://www.stat.umn.edu/geyer/old/5102/n.pdf.

Following Theorem 4.3.2 of Casella Berger

$X_1 \sim Ga(a_1, 1)$ $X_2 \sim Ga(a_2, 1)$

The joint pdf of (X_1, X_2) where $x_1, x_2, a_1, a_2 > 0$ is

$$f_{X_1,X_2}(x_1,x_2) = f(x_1) f(x_2) = \frac{1}{\Gamma(a_1)} x_1^{a_1-1} e^{-x_1} \frac{1}{\Gamma(a_2)} x_2^{a_2-1} e^{-x_2}$$

Let $U = g_1(x_1, x_2) = \frac{x_1}{x_1 + x_2}$ and $V = g_2(x_1, x_2) = x_1 + x_2$ Note: This is a one-to-one transformation.

We have $x_1 = h_1(u, v) = uv$ and $x_2 = h_2(u, v) = v - uv$

$$\Rightarrow f_{X,V}(h_1(u,v), h_2(u,v)) = \frac{1}{\Gamma(a_1)\Gamma(a_2)} (uv)^{a_1-1} (v-uv)^{a_2-1} e^{-v}$$

Now we can find

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial h_1(u,v)}{\partial u} \frac{\partial h_2(u,v)}{\partial v} - \frac{\partial h_1(u,v)}{\partial v} \frac{\partial h_2(u,v)}{\partial u} = (v)(1-u) - (u)(-v)$$

$= v - uv + uv = v$ Note: $v = x_1 + x_2 > 0$ so $|J| = |v| = v$

Thus $f_{U,V}(u,v) = f_{X,V}(h_1(u,v), h_2(u,v))|J| = \frac{v}{\Gamma(a_1)\Gamma(a_2)} (uv)^{a_1-1} (v-uv)^{a_2-1} e^{-v}$

To find the marginal densities of U and V we need to show that we can re-arrange $f_{U,V}(u,v)$ into two factors: one of u only and one of v only.

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

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

Suppose that $x_1 \sim \text{Ga}(a_1, 1)$ and that $x_2 \sim \text{Ga}(a_2, 1)$. 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).[2] 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.

[2]Take care that you apply the important change-of-variable formula from basic probability. See, e.g., Section 1.2 of http://www.stat.umn.edu/geyer/old/5102/n.pdf.

$$f_{U,V}(u,v) = \frac{1}{\Gamma(a_1)\Gamma(a_2)} u^{a_1-1}(1-u)^{a_2-1} v^{a_1+a_2-1} e^{-v}$$

Note: U and V appear to have Beta(a_1, a_2) and Gamma$(a_1+a_2, 1)$ pdfs

$$f_{U,V}(u,v) = \left[\frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)} u^{a_1-1}(1-u)^{a_2-1}\right]\left[\frac{1}{\Gamma(a_1+a_2)} v^{a_1+a_2-1} e^{-v}\right]$$

Thus by Lemma 4.2.7 of Casella Berger

$y_1 \sim$ Beta(a_1, a_2) and $y_2 \sim$ Gamma$(a_1+a_2, 1)$

Furthermore if you have a way to simulate random variables such that $X_1 \sim Ga(a_1, 1)$ and $X_2 \sim Ga(a_2, 1)$ where X_1 and X_2 are independent you can use them to calculate $Y_1 = \frac{X_1}{X_1+X_2}$ and $Y_1 \sim$ Beta(a_1, a_2) allowing the possibility to simulate Beta random variables with your choice of parameters for α and β.

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

We begin by calculating the likelihood $L(\theta \mid x_1, \ldots, x_N)$

Since $x_1, \ldots, x_N$ are independent observations the likelihood function is the joint density calculated as follows.

$$L(\theta \mid x_1, \ldots, x_N) = p(x_1, \ldots, x_N \mid \theta) = p(x_1 \mid \theta) \cdots p(x_N \mid \theta) = \prod p(x_i \mid \theta)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\theta)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-N/2} e^{-(\sum x_i^2 - 2\theta\sum x_i + N\theta^2)/2\sigma^2}$$

The prior distribution given is: $p(\theta) = \frac{1}{\sqrt{2\pi v}} e^{-(\theta-m)^2/2v}$

The posterior distribution is proportional to prior $\cdot$ likelihood

$$p(\theta \mid x_1, \ldots, x_N) \propto (2\pi\sigma^2)^{-N/2} e^{-(\sum x_i^2 - 2\theta\sum x_i + N\theta^2)/2\sigma^2} \frac{1}{\sqrt{2\pi v}} e^{-(\theta-m)^2/2v}$$

$$\propto \exp\left(\frac{\sum x_i}{\sigma^2}\theta - \frac{N}{2\sigma^2}\theta^2 - \frac{1}{2v}\theta^2 + \frac{m}{v}\theta\right) = \exp\left(\theta^2\left(\frac{-Nv-\sigma^2}{2\sigma^2 v}\right) + \theta\left(\frac{\sum x_i}{\sigma^2} + \frac{m}{v}\right)\right)$$

$$\propto \exp\left(\frac{-Nv-\sigma^2}{2\sigma^2 v}\left(\theta^2 - 2\left(\frac{v\sum x_i + \sigma^2 m}{Nv+\sigma^2}\right)\theta + \left(\frac{v\sum x_i + \sigma^2 m}{Nv+\sigma^2}\right)^2\right)\right)$$

$$= \exp\left\{-\left(\theta - \frac{v\sum x_i + \sigma^2 m}{Nv+\sigma^2}\right)^2 \Big/ \left(2\left(\frac{\sigma^2 v}{Nv+\sigma^2}\right)\right)\right\}$$

In this form we can see that the posterior distribution of θ is $N\left(\frac{v\sum x_i + \sigma^2 m}{Nv+\sigma^2}, \frac{\sigma^2 v}{Nv+\sigma^2}\right)$.

(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. (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 $\omega = 1/\sigma^2$:

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

Suppose that ω has a gamma prior with parameters a and b, implying that σ^2 has what is called an inverse-gamma prior.[3] Derive the posterior distribution $p(\omega \mid x_1, \ldots, x_N)$. Re-express this as a posterior for σ^2, the variance.

[3] Written $\sigma^2 \sim \text{IG}(a, b)$.

We begin by calculating the likelihood $L(\omega | x_1, \ldots, x_N)$

Since $x_1, \ldots, x_N$ are independent observations the likelihood function is the joint density calculated as follows.

$$L(\omega | x_1, \ldots, x_N) = p(x_1, \ldots, x_N | \omega) = p(x_1|\omega) \cdots p(x_N|\omega) = \prod p(x_i|\omega)$$

$$= \prod_{i=1}^{N} \sqrt{\frac{\omega}{2\pi}}\, e^{\frac{-\omega(x_i-\theta)^2}{2}} = \left(\frac{\omega}{2\pi}\right)^{N/2} e^{-\frac{\omega}{2}\sum(x_i-\theta)^2}$$

The prior distribution given is: $p(\omega) = \frac{b^a}{\Gamma(a)}\, \omega^{a-1} e^{-\omega b}$

The posterior distribution is proportional to prior $\cdot$ likelihood

$$p(\omega | x_1, \ldots, x_N) \propto \left(\frac{\omega}{2\pi}\right)^{N/2} e^{-\frac{\omega}{2}\sum(x_i-\theta)^2} \frac{b^a}{\Gamma(a)}\, \omega^{a-1} e^{-\omega b}$$

$$\propto \omega^{\frac{N}{2}+a-1} \exp\left(-\omega\left(\tfrac{1}{2}\sum(x_i-\theta)^2 + b\right)\right)$$

In this form we can see that the posterior distribution of ω is $\text{Gamma}\left(\frac{N}{2} + a,\ b + \frac{1}{2}\sum(x_i-\theta)^2\right)$

Since $\omega = \frac{1}{\sigma^2} \sim \text{Gamma}\left(\frac{N}{2} + a, b + \frac{1}{2}\sum(x_i-\theta)^2\right)$ we have that

$$\frac{1}{\omega} = \sigma^2 \sim \text{IG}\left(\frac{N}{2} + a, b + \frac{1}{2}\sum(x_i-\theta)^2\right).$$

don't use these properties show w/ transformation of variables.

$$\left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}+a-1} \exp\left(-\left(\frac{1}{\sigma^2}\right)\left(\tfrac{1}{2}\sum(x_i-\theta)^2 + b\right)\right)\left|-\left(\frac{1}{\sigma^2}\right)^2\right|$$

$$(\sigma^2)^{-\frac{N}{2}-a-1} \exp\left(-(\sigma^2)^{-1}\left(\tfrac{1}{2}\sum(x_i-\theta)^2 + b\right)\right)$$

↓ kernel of IG

★ Ignoring the normalizing constant doesn't work if the kernel of the known density integrates to ∞ ★

(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: $x_i \sim \mathrm{N}(\theta, \sigma_i^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta \mid x_1, \ldots, x_N)$. Express the posterior mean in a form that is clearly interpretable as a weighted average of the observations and the prior mean.

We begin by calculating the likelihood $L(\theta \mid x_1, \ldots, x_N)$

Since $x_1, \ldots, x_N$ are independent observations the likelihood function is the joint density calculated as follows.

$$L(\theta \mid x_1, \ldots, x_N) = p(x_1, \ldots, x_N \mid \theta) = p(x_1 \mid \theta) \cdots p(x_N \mid \theta) = \prod p(x_i \mid \theta)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x_i-\theta)^2}{2\sigma_i^2}} = \left[2\pi \prod \sigma_i^2\right]^{-N/2} e^{-\frac{1}{2}\sum \frac{x_i^2 - 2\theta x_i + \theta^2}{\sigma_i^2}}$$

The prior distribution given is: $p(\theta) = \frac{1}{\sqrt{2\pi v}} e^{-(\theta-m)^2/2v}$

The posterior distribution is proportional to prior · likelihood

$$p(\theta \mid x_1, \ldots, x_N) \propto \left[2\pi \prod \sigma_i^2\right]^{-N/2} e^{-\frac{1}{2}\left\{\sum \frac{x_i^2 - 2\theta x_i + \theta^2}{\sigma_i^2}\right\}} \frac{1}{\sqrt{2\pi v}} e^{-(\theta-m)^2/2v}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{v} + \sum \frac{1}{\sigma_i^2}\right) - 2\theta\left(\frac{m}{v} + \sum \frac{x_i}{\sigma_i^2}\right)\right]\right\} \frac{\frac{m}{v} + \sum \frac{x_i}{\sigma_i^2}}{\frac{1}{v} + \sum \frac{1}{\sigma_i^2}}$$

$$\propto \exp\left\{-\frac{1 + v\sum \frac{1}{\sigma_i^2}}{2v}\left[\theta^2 - 2\theta\left(\frac{m + v\sum \frac{x_i}{\sigma_i^2}}{1 + v\sum \frac{1}{\sigma_i^2}}\right) + \left(\frac{m + v\sum \frac{x_i}{\sigma_i^2}}{1 + v\sum \frac{1}{\sigma_i^2}}\right)^2\right]\right\}$$

$$= \exp\left\{-\frac{1 + v\sum \frac{1}{\sigma_i^2}}{2v}\left(\theta - \frac{m + v\sum \frac{x_i}{\sigma_i^2}}{1 + v\sum \frac{1}{\sigma_i^2}}\right)^2\right\}$$

Lemma for $\exp\{-\frac{1}{2}(A\theta^2 - 2b\theta)\} \sim N(A^{-1}b, A^{-1})$

In this form we can see that the posterior distribution $P(\theta \mid x_1, \ldots, x_N)$ is $N\left(\frac{\frac{m}{v} + \sum \frac{x_i}{\sigma_i^2}}{\frac{1}{v} + \sum \frac{1}{\sigma_i^2}}, \left[\frac{1}{v} + \sum \frac{1}{\sigma_i^2}\right]^{-1}\right)$. Here we can see that the mean is a weighted average of the observations and the prior mean.

$$N\left(\frac{\sum \frac{1}{\sigma_i^2}(x_i) + \frac{1}{v}m}{\sum \frac{1}{\sigma_i^2} + \frac{1}{v}}\right) \begin{cases} \text{weight } \left(\sum \frac{1}{\sigma_i^2} + \frac{1}{v}\right) \\ \text{observations } x_i \\ \text{prior mean } m \end{cases}$$

This is a nice result of the linearity of data and prior

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

Given $x|w \sim N(m, w^{-1})$ and $w \sim \text{Gamma}(\frac{a}{2}, \frac{b}{2})$ we can find the marginal density of x as follows

$$p(x) = \int_{w>0} p(x|w)\,p(w)\,dw = \int_{w>0} \sqrt{\frac{w}{2\pi}}\, e^{-\frac{w}{2}(x-m)^2} \left[\frac{(\frac{b}{2})^{a/2}}{\Gamma(\frac{a}{2})}\right] w^{\frac{a}{2}-1} e^{-\frac{b}{2}w}\, dw$$

$$= (2\pi)^{-1/2} \left(\frac{b}{2}\right)^{a/2} \left[\Gamma\left(\frac{a}{2}\right)\right]^{-1} \int_{w>0} \underbrace{w^{(\frac{1}{2}a+\frac{1}{2})-1}\, e^{-\left[\frac{1}{2}(x-m)^2+\frac{b}{2}\right]w}}_{\text{kernel of Gamma}\left(\frac{a}{2}+\frac{1}{2},\ \frac{1}{2}(x-m)^2+\frac{b}{2}\right)} dw = \frac{(\frac{b}{2})^{a/2}\ \Gamma(\frac{a}{2}+\frac{1}{2})}{\sqrt{2\pi}\,\Gamma(\frac{a}{2})\left[\frac{1}{2}(x-m)^2+\frac{b}{2}\right]^{\frac{a}{2}+\frac{1}{2}}}$$

$$= \frac{\Gamma\left(\frac{a+1}{2}\right)}{\Gamma\left(\frac{a}{2}\right)\sqrt{\pi a}\,\left(\frac{b}{a}\right)^{1/2}} \left(1 + \frac{1}{a}\left(\frac{x-m}{(b/a)^{1/2}}\right)^2\right)^{-\frac{(a+1)}{2}}$$

Notice this is the pdf of a non-standardized students t-distribution with location parameter m, scale parameter $(\frac{b}{a})^{1/2}$ and $d = a$ degrees of freedom.

Note: the scale parameter $\left(\frac{b}{a}\right)^{1/2}$ is a product of a, so a lot of times people will use Gamma$(\frac{a}{2}, \frac{ab}{2})$ as a prior so that the scale parameter would only be a function of b.

2 The multivariate normal distribution

2.1 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 $x \sim \mathcal{N}(m, v)$.

Here's an alternative characterization of the univariate normal distribution in terms of moment-generating functions:[4] 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.

[4] Laplace transforms to everybody but statisticians.

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

(1) WTS $\text{Cov}(x) = E(xx^T) - \mu\mu^T$

We start with

$$\text{Cov}(x) = E\{(x-\mu)(x-\mu)^T\} = E\{(x-\mu)(x^T-\mu^T)\}$$

$$= E\{xx^T - x\mu^T - \mu x^T + \mu\mu^T\}$$

$$= E(xx^T) - E(x\mu^T) - E(\mu x^T) + E(\mu\mu^T)$$

Note: $E(x) = \mu$, $E(\mu) = \mu$

$$= E(xx^T) - \mu\mu^T - \mu\mu^T + \mu\mu^T = E(xx^T) - \mu\mu^T$$

(2) WTS $\text{Cov}(Ax+b) = A\,\text{Cov}(x)A^T$

We start with

$$\text{Cov}(Ax+b) = E\{(Ax+b - A\mu - b)(Ax+b - A\mu - b)^T\}$$

Note: mean of $Ax+b$ is $A\mu+b$

$$= E\{[A(x-\mu)][A(x-\mu)]^T\} = E\{A(x-\mu)(x-\mu)^T A^T\}$$

$$= A\,E\{(x-\mu)(x-\mu)^T\}A^T = A\,\text{Cov}(x)\,A^T$$

(B) Consider the random vector $z = (z_1, \ldots, z_p)^T$, 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 z, expressed in vector notation.[5] We say that z has a standard multivariate normal distribution.

[5]Remember that the MGF of a vector-valued random variable x is the expected value of the quantity $\exp(t^T x)$, as a function of the vector argument t.

$z = \begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix}$ where $z_i \sim N(0,1)$ $i = 1, \ldots, p$ meaning the pdf of z_i is $f(z_i) = \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2}$ $i = 1, \ldots, p$. Since each z_i is independent the joint pdf of the z_i's is the product of their individual pdfs.

$$f_z(z) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = (2\pi)^{-p/2} e^{-\frac{1}{2}\sum z_i^2} = (2\pi)^{-p/2} e^{-\frac{1}{2} z^T z}$$

$$\Longrightarrow f_z(z) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2} z^T z}$$

Now to find the mgf

$$m_z(t) = E(e^{t^T z}) = E(e^{\sum_{i=1}^{p} t_i z_i}) = \prod_{i=1}^{p} E(e^{t_i z_i}) = \prod_{i=1}^{p} e^{t_i^2/2}$$

$$= e^{\sum t_i^2/2} = e^{t^T t/2}$$

$$\Longrightarrow m_z(t) = e^{t^T t/2}$$

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

Part 1 $\Rightarrow$ let $x \sim N(\mu, \Sigma)$ WTS $M_x(t) = E(e^{t^T x}) = \exp\{t^T\mu + t^T\Sigma t/2\}$

We know that the MGF of a vector-valued random variable x is the expected value of $e^{t^T x}$ as a function of the vector argument t.

Let $a \neq 0$ be a $p \times 1$ vector. Since a is not identically zero we have $z = a^T x$ is normally distributed with mean $a^T\mu$ and variance $a^T\Sigma a$

$E(a^T x) = a^T E(x) = a^T \mu$

$Var(a^T x) = a^T Cov(x) a = a^T \Sigma a$

MGF for $N(\mu, \sigma^2)$: $M_x(t) = e^{\mu t + \sigma^2 t^2/2}$

$$E(e^{t a^T x}) = \exp\{ t a^T \mu + t\, a^T \Sigma a\, t/2 \}$$

Let $at = t$ now we have

$$M_x(t) = E(e^{t^T x}) = \exp\{t^T\mu + t^T\Sigma t/2\}$$

Part 2 $\Leftarrow$ let x have mgf of the form

$$M_x(t) = E(e^{t^T x}) = \exp\{t^T\mu + t^T\Sigma t/2\}$$ WTS $x \sim N(\mu, \Sigma)$

Now let x Not follow $N(\mu, \Sigma)$

Let $y \sim N(\mu, \Sigma)$ then by part 1

$$M_y(t) = E(e^{t^T y}) = \exp\{t^T\mu + t^T\Sigma t/2\}$$

By properties of MGFs since $M_x(t) = M_y(t)$ that means $F_x(x) = F_y(y)$. Thus $x \sim N(\mu, \Sigma)$ $\longrightarrow\ \longleftarrow$ contradiction

Thus we can't have x with a distribution other than MVN and $x \sim N(\mu, \Sigma)$.

A geometric transformation that preserves lines and parallelism.

(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 $x = Lz + \mu$ for some $p \times p$ matrix L of full column rank.[6] Prove that x is multivariate normal. In addition, use the moment identities you proved above to compute the expected value and covariance matrix of x.

[6]The full rank restriction turns out to be unnecessary; relaxing it leads to what is called the *singular normal distribution*.

Let z have a standard multivariate normal distribution. $f_z(z) = (2\pi)^{\frac{-p}{2}} e^{\frac{-1}{2} z^T z}$

Let $X = LZ + \mu$ where $\begin{cases} X \text{ is } p \times 1 \\ L \text{ is } p \times p \text{ full rank} \\ \mu \text{ is } p \times 1 \end{cases}$

We will show $X \sim MVN$ using transformations of variables

From pg 53 of Casella Berger we have

Theorem 2.1.5 Let z have pdf $f_z(z)$ and let $x = g(z)$, where g is a monotone function. Let Z and X be defined by (2.1.7). Suppose that $f_z(z)$ is continuous on Z and that $g^{-1}(x)$ has a continuous derivative on X. Then the pdf of X is given by

$$f_X(x) = \begin{cases} f_z(g^{-1}(x)) \left| \frac{d}{dx} g^{-1}(x) \right| & x \in X \\ 0 & \text{otherwise} \end{cases}$$

Notice that $X = g(z) = Lz + \mu$ is a monotone function.

We also notice that $f_z(z)$ is continuous on it's domain Z

$X = L g^{-1}(x) + \mu \Rightarrow X - \mu = L g^{-1}(x) \Rightarrow L^{-1}X - L^{-1}\mu = g^{-1}(x)$

Note L is full rank so L^{-1} exists

Now consider $\frac{d}{dx} g^{-1}(x) = \frac{d}{dx} L^{-1}x - L^{-1}\mu = L^{-1}$ this shows us that the derivative of $g^{-1}(x)$ is continuous on the domain of X, X. Thus by 2.1.5

$$f_X(x) = (2\pi)^{-p/2} e^{-1/2 [L^{-1}(x-\mu)]^T L^{-1}(x-\mu)} |L^{-1}| = (2\pi)^{\frac{-p}{2}} e^{\frac{-1}{2}(x-\mu)^T (LL^T)^{-1}(x-\mu)} |L^{-1}|$$

(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 $x = Lz + \mu$ for some $p \times p$ matrix L of full column rank.[6] Prove that x is multivariate normal. In addition, use the moment identities you proved above to compute the expected value and covariance matrix of x.

[6]The full rank restriction turns out to be unnecessary; relaxing it leads to what is called the *singular normal distribution*.

Looking at the definition of MVN we can see this is the pdf of $N(\mu, LL^T)$ also note since L is full rank, $\Sigma = LL^T$ is positive definite.

A vector-valued random variable $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ is said to have a **multivariate normal (or Gaussian) distribution** with mean $\mu \in \mathbf{R}^n$ and covariance matrix $\Sigma \in \mathbf{S}^n_{++}$ if its probability density function[2] is given by

symmetric positive definite

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

We write this as $X \sim \mathcal{N}(\mu, \Sigma)$. In these notes, we describe multivariate Gaussians and some of their basic properties.

Thus $E(X) = \mu$ and $\text{cov}(X) = LL^T$

MGF

$$E\left[\exp\{t^T(Lz+\mu)\}\right]$$

$$E\left[\exp\{t^TLz\}\exp\{t^T\mu\}\right]$$

$$\exp\{t^T\mu\}\, E\left[\exp\{t^TLz\}\right]$$

$$\exp\{t^T\mu\}\, M_z(L^Tt)$$

$$\exp\{t^T\mu\}\exp\{t^TLIL^Tt/2\}$$

$$\exp\{t^T\mu + t^T(LL^T)t/2\}$$

(E) Now for the "only if." Suppose that x has a multivariate normal distribution. Prove that x can be written as an affine transformation of standard normal random variables. (Note: a good way to prove that something can be done is to do it! Think about a matrix A such that $AA^T = \Sigma$.) Use this insight to propose an algorithm for simulating multivariate normal random variables with a specified mean and covariance matrix.

let $x \sim N(\mu, \Sigma)$ WTS $x = Lz + \mu$ where L is full rank

Remember that a property of Σ is that it is symmetric and positive definite.

Theorem B.22 A is positive definite iff there exists a square matrix Q such that $A = QQ^T$

One property of positive definite matrices is that Σ is positive definite iff there exists an invertible matrix L such that $\Sigma = LL^T$ (Note L is full rank so it is invertible). (Cholesky decomposition)

Thus $x \sim N(\mu, LL^T)$ we know that the mgf of this distribution is

$$M_x(t) = E(e^{t^T x}) = \exp\{t^T\mu + t^T\Sigma t/2\} = \exp\{t^T\mu + t^T LL^T t/2\}$$

$$= e^{t^T\mu} e^{t^T LL^T t/2} = e^{t^T\mu} E[e^{t^T Lz}] \text{ where } z \sim MVN(0,1)$$

$$= E[e^{z^T\mu + t^T Lz}] = E[e^{t^T(Lz+\mu)}] = M_x(t) \text{ where } x = Lz + \mu$$

Since the mgf of x is the same as the mgf of $Lz + \mu$ and mgfs are unique, we know that $x = Lz + \mu$ where $\Sigma = LL^T$ and $z \sim MVN(0,1)$. Thus x can be written as an affine transformation of standard normal random variables.

This means that if we wish to simulate a sample of MVN random variables with a specified mean μ and covariance matrix Σ then we can find the Cholesky decomposition of $\Sigma = LL^T$ and sample from $z \sim MVN(0,1)$ then calculate $x = Lz + \mu$ to obtain a sample of $x \sim MVN(\mu, \Sigma)$.

(F) Use this last result, together with the PDF of a standard multivariate normal, to show that the PDF of a multivariate normal $x \sim \mathcal{N}(\mu, \Sigma)$ takes the form $p(x) = C \exp\{-Q(x-\mu)/2\}$ for some constant C and quadratic form $Q(x-\mu)$.[7]

[7] A useful fact is that the Jacobian matrix of the linear map $x \to Ax$ is simply A.

From part (D) we have that for $z \sim MVN(0,1)$, and

$X = Lz + \mu$

$$f_X(x) = (2\pi)^{-p/2} e^{-1/2 [L^{-1}(x-\mu)]^T L^{-1}(x-\mu)} |L^{-1}| = (2\pi)^{-p/2} e^{-\frac{1}{2}(x-\mu)^T (L^{-1})^T L^{-1}(x-\mu)} |L^{-1}|$$

Definition 1.3.1 from Plane Answers to Complex Questions

Let Y be a p-dimensional random vector and let A be a $p \times p$ matrix. A quadratic form is a random variable defined by $Y^T A Y$ for some Y and A.

Let $C = (2\pi)^{-p/2} |L^{-1}|$ $\quad Q = (L^{-1})^T L^{-1}$ then we have

$$f_x(x) = C e^{-\frac{1}{2}(x-\mu)^T Q (x-\mu)}$$

this is the desired form.

Note $(x-\mu)^T Q (x-\mu)$ is in quadratic form.

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

Let x_1 be $n_1 \times 1$ and x_2 be $n_2 \times 1$ then let

A be $n \times n_1$ and B be $n \times n_2$ making y $n \times 1$.

We will find the mgf of y to see which distribution y follows

$$M_y(t) = E(e^{t^T y}) = E(e^{t^T(Ax_1 + Bx_2)}) = E(e^{t^T Ax_1}) E(e^{t^T Bx_2})$$

$$= M_{x_1}(A^T t) M_{x_2}(B^T t) = \exp\{t^T A\mu_1 + t^T A\Sigma_1 A^T t/2\} \exp\{t^T B\mu_2 + t^T B\Sigma_2 B^T t/2\}$$

$$= \exp\{t^T(A\mu_1 + B\mu_2) + t^T(A\Sigma_1 A^T + B\Sigma_2 B^T)t/2\}$$

This is the mgf of $MVN(A\mu_1 + B\mu_2, A\Sigma_1 A^T + B\Sigma_2 B^T)$

Since the mgf uniquely describes a distribution

we have $y \sim MVN(A\mu_1 + B\mu_2, A\Sigma_1 A^T + B\Sigma_2 B^T)$

Note $A\Sigma_1 A^T + B\Sigma_2 B^T$ needs to be positive definite we are given A and B are full column rank and Σ_1, Σ_2 are covariance matrices so they are symmetric and positive definite.

Two properties of positive definite matrices are:

1. If M and N are positive definite then $M+N$ is positive definite
2. If M is positive definite and Y has full column rank then $Y^T M Y$ is positive definite.

Thus $A\Sigma_1 A^T$ and $B\Sigma_2 B^T$ are positive definite and $A\Sigma_1 A^T + B\Sigma_2 B^T$ is positive definite

2.2 Conditionals and marginals

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

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

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

(A) Derive the marginal distribution of x_1. (Remember your result about affine transformations.)

2.2 Conditionals and marginals

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

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

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

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ \vdots \\ x_p \end{bmatrix} \begin{matrix} \} X_1 = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \\ \} X_2 = \begin{bmatrix} x_{k+1} \\ \vdots \\ x_p \end{bmatrix} \end{matrix}$$

we have $x_1 = [x_1 \dots x_k]^T$ is a $k \times 1$ vector

let $A = [\, I_{k\times k} \;\; 0_{(p-k)\times(p-k)} \,]_{p\times p}$ then we can write

$$x_1 = Ax = \underbrace{\begin{bmatrix} 1 & & 0 & 0 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & 1 & 0 & & 0 \end{bmatrix}}_{p\times p} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}}_{p\times 1} = \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{p\times 1}$$

Now we will find the mgf of x_1 in hopes of finding the distribution it's associated with

$$M_{x_1}(t) = E[e^{t^T x_1}] = E[e^{t^T Ax}] = E[e^{(A^T t)^T x}] = M_x(A^T t)$$

$$= \exp\{ t^T A\mu + t^T A \Sigma A^T t/2 \}$$

$A\mu = \mu_1$

$$A\Sigma A^T = [\, I_{k\times k} \; 0_{(p-k)\times(p-k)} \,] \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I_{k\times k} \\ 0_{(p-k)\times(p-k)} \end{bmatrix} = [\, \Sigma_{11} \; \Sigma_{12} \,] \begin{bmatrix} I_{k\times k} \\ 0_{(p-k)\times(p-k)} \end{bmatrix} = \Sigma_{11}$$

$$= \exp\{ t^T \mu_1 + t^T \Sigma_{11} t/2 \}$$

This is the mgf of a $MVN(\mu_1, \Sigma_{11})$ distribution

thus $x_1 \sim MVN(\mu_1, \Sigma_{11})$.

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

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix}$$

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

Since $\Omega = \Sigma^{-1}$ and we know $\Sigma\Sigma^{-1} = I$ we have

$$\Sigma\Sigma^{-1} = \Sigma\Omega = I \Rightarrow \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{bmatrix} = I$$

$$\Rightarrow \begin{bmatrix} \Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{12}^T & \Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} \\ \Sigma_{12}^T\Omega_{11} + \Sigma_{22}\Omega_{12}^T & \Sigma_{12}^T\Omega_{12} + \Sigma_{22}\Omega_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

① $\Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{21} = I$

② $\Sigma_{21}\Omega_{11} + \Sigma_{22}\Omega_{21} = 0$

③ $\Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} = 0$

④ $\Sigma_{21}\Omega_{12} + \Sigma_{22}\Omega_{22} = I$

③ $\Sigma_{12}\Omega_{22} = -\Sigma_{11}\Omega_{12}$

$\Omega_{22} = -\Sigma_{12}^{-1}\Sigma_{11}\Omega_{12}$

④ $\Sigma_{21}\Omega_{12} + \Sigma_{22}\Omega_{22} = I$

$\Sigma_{21}\Omega_{12} - \Sigma_{22}\Sigma_{12}^{-1}\Sigma_{11}\Omega_{12} = I$

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

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

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

② $\Sigma_{22}\Omega_{21} = -\Sigma_{21}\Omega_{11}$

$\Omega_{21} = -\Sigma_{22}^{-1}\Sigma_{21}\Omega_{11}$

① $\Sigma_{11}\Omega_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Omega_{11} = I$

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

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

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

(C) Derive the conditional distribution for x_1, given x_2, in terms of the partitioned elements of x, μ, and Σ. There are several keys to inner peace: work with densities on a log scale, ignore constants that don't affect x_1, and remember the cute trick of completing the square from basic algebra.[8] Explain briefly how one may interpret this conditional distribution as a linear regression on x_2, where the regression matrix can be read off the precision matrix.

[8] In scalar form:

$$\begin{aligned} x^2 - 2bx + c &= x^2 - 2bx + b^2 - b^2 + c \\ &= (x-b)^2 - b^2 + c. \end{aligned}$$

We know that $p(x) = p(x_1, x_2) = p(x_1|x_2)\,p(x_2)$

We will focus on terms involving x_1 and match the kernel of $p(x_1|x_2)$ to its distribution. We will also use the advice of the problem and work with the logarithm

$$\ln(p(x_1|x_2)) \propto \ln(p(x_1, x_2)) = \ln(p(x))$$

$$\propto -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) = -\frac{1}{2}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^T \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)$$

$$= -\frac{1}{2}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} = -\frac{1}{2}\left[(x_1-\mu_1)^T (x_2-\mu_2)^T\right] \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

$$= -\frac{1}{2}\begin{bmatrix} (x_1-\mu_1)^T\Omega_{11} + (x_2-\mu_2)^T\Omega_{21} & (x_1-\mu_1)^T\Omega_{12} + (x_2-\mu_2)^T\Omega_{22} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

$$= -\frac{1}{2}\left[(x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + (x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) + (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2) + (x_2-\mu_2)^T\Omega_{22}(x_2-\mu_2)\right]$$

$$\propto -\frac{1}{2}\left[(x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + (x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) + (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2)\right]$$

$$= -\frac{1}{2}\left[(x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + \left[(x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2)\right]^T + (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2)\right]$$

Note this is a scalar, let's say c, so $c^T = c$

$$= -\frac{1}{2}\left[(x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + 2(x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2)\right]$$

(C) Derive the conditional distribution for x_1, given x_2, in terms of the partitioned elements of x, μ, and Σ. There are several keys to inner peace: work with densities on a log scale, ignore constants that don't affect x_1, and remember the cute trick of completing the square from basic algebra.[8] Explain briefly how one may interpret this conditional distribution as a linear regression on x_2, where the regression matrix can be read off the precision matrix.

[8] In scalar form:

$$\begin{aligned} x^2 - 2bx + c &= x^2 - 2bx + b^2 - b^2 + c \\ &= (x-b)^2 - b^2 + c. \end{aligned}$$

$$= -\frac{1}{2}\left[(x_1-\mu_1)^T \Omega_{11} (x_1-\mu_1) + 2 (x_1-\mu_1)^T \Omega_{12} (x_2-\mu_2) \right]$$

$$\propto -\frac{1}{2}\left[x_1^T \Omega_{11} x_1 - x_1^T \Omega_{11} \mu_1 - \mu_1^T \Omega_{11} x_1 + 2\left(x_1^T \Omega_{12} x_2 - x_1^T \Omega_{12} \mu_2 \right) \right]$$

$$= -\frac{1}{2}\left[x_1^T \Omega_{11} x_1 - 2 x_1^T \Omega_{11} \mu_1 + 2 x_1^T \Omega_{12} x_2 - 2 x_1^T \Omega_{12} \mu_2 \right]$$

$$= -\frac{1}{2}\left[x_1^T \Omega_{11} x_1 - 2 x_1^T \left(\Omega_{11}\mu_1 - \Omega_{12} x_2 + \Omega_{12}\mu_2 \right) \right]$$

Completing the square (multivariate)

For any vectors $x, b \in \mathbb{R}^d$ and symmetric invertible matrix $M \in \mathbb{R}^{d \times d}$, we have

$$x^T M x - 2b^T x = (x - M^{-1}b)^T M (x - M^{-1}b) - b^T M b$$

Note $x = x_1$ and $b = \Omega_{11}\mu_1 - \Omega_{12} x_2 + \Omega_{12}\mu_2$ are vectors and Ω_{11} is symmetric and invertible

$$\propto -\frac{1}{2}\left\{ \left[x_1 - \Omega_{11}^{-1}\left(\Omega_{11}\mu_1 - \Omega_{12} x_2 + \Omega_{12}\mu_2\right) \right]^T \Omega_{11} \left[x_1 - \Omega_{11}^{-1}\left(\Omega_{11}\mu_1 - \Omega_{12} x_2 + \Omega_{12}\mu_2\right) \right] - \left(\Omega_{11}\mu_1 - \Omega_{12} x_2 + \Omega_{12}\mu_2\right)^T \Omega_{11} \left(\Omega_{11}\mu_1 - \Omega_{12} x_2 + \Omega_{12}\mu_2\right) \right\}$$

$$\propto -\frac{1}{2}\left[x_1 - \Omega_{11}^{-1}\left(\Omega_{11}\mu_1 - \Omega_{12} x_2 + \Omega_{12}\mu_2\right) \right]^T \Omega_{11} \left[x_1 - \Omega_{11}^{-1}\left(\Omega_{11}\mu_1 - \Omega_{12} x_2 + \Omega_{12}\mu_2\right) \right]$$

In this form we can see that this has the kernel of a MVN distribution with mean and covariance

$$\mu = \mu_1 - \Omega_{11}^{-1}\Omega_{12}(x_2 - \mu_2) \qquad \Sigma = \Omega_{11}^{-1}$$

Thus $x_1 | x_2 \sim MVN\left(\mu_1 - \Omega_{11}^{-1}\Omega_{12}(x_2 - \mu_2), \Omega_{11}^{-1}\right)$

(C) Derive the conditional distribution for x_1, given x_2, in terms of the partitioned elements of x, μ, and Σ. There are several keys to inner peace: work with densities on a log scale, ignore constants that don't affect x_1, and remember the cute trick of completing the square from basic algebra.[8] Explain briefly how one may interpret this conditional distribution as a linear regression on x_2, where the regression matrix can be read off the precision matrix.

[8] In scalar form:

$$\begin{aligned} x^2 - 2bx + c &= x^2 - 2bx + b^2 - b^2 + c \\ &= (x-b)^2 - b^2 + c\,. \end{aligned}$$

Thus $x_1 | x_2 \sim MVN(\mu_1 - \Omega_{11}^{-1}\Omega_{12}(x_2 - \mu_2), \Omega_{11}^{-1})$

Now to write it in terms of the partitioned elements of $x, \mu,$ and Σ.

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

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

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

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

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

Consider rewriting x_1 in the form

$x_1 = \mu_1 + \Sigma_{12}\Sigma_{22}(x_2 - \mu_2) + \varepsilon$ where $\varepsilon \sim N(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

let $\beta = \Sigma_{12}\Sigma_{22}$ then we have

$x_1 = \mu_1 + \beta(x_2 - \mu_2) + \varepsilon$ where $\varepsilon \sim N(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

This can be interpreted as a linear regression on x_2 where $\Sigma_{12}\Sigma_{22}$ is the estimate for β

3 Multiple regression: three classical principles for inference

Suppose we observe data that we believe to follow a linear model, where $y_i = x_i^T\beta + \epsilon_i$ for $i = 1, \ldots, n$. To fix notation: y_i is a scalar response; 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 $x_i^T\beta$ will be our typical way of writing the inner product between the vectors x_i and β.[-4pc]Notice we have no explicit intercept. For now you can imagine that all the variables have had their sample means subtracted, making an intercept superfluous. Or you can just assume that the leading entry in every x_i is equal to 1, in which case β_1 will be an intercept term.

Consider three classic inferential principles that are widely used to estimate β, the vector of regression coefficients. In this context we will let $\hat{\beta}$ denote an estimate of β, $y = (y_1, \ldots, y_n)^T$ the vector of outcomes, X the matrix of predictors whose ith row is x_i^T, and ϵ the vector of residuals $(\epsilon_1, \ldots, \epsilon_n)^T$.

Least squares: make the sum of squared errors as small as possible. We can express this in terms of the squared Euclidean norm of the residual vector $\epsilon = y - X\beta$:

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^p} \|y - X\beta\|_2^2 = \arg\min_{\beta \in \mathcal{R}^p} (y - X\beta)^T(y - X\beta)$$

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

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

Here $p_i(y_i \mid \sigma^2)$ is the conditional probability density function of y_i, given the model parameters β and σ^2. Note that an equivalent way to write the likelihood is to say that the response vector y is multivariate normal with mean $X\beta$ and covariance matrix $\sigma^2 I$, where I is the n-dimensional identity matrix.

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

* Assuming X is full rank

(A) Show that all three of these principles lead to the same estimator. What is the variance of this estimator under the assumption that each ϵ_i is independent and identically distribution with variance σ^2?

① We begin with $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2$ we will find the min by setting the partial derivative wrt β to zero

$$\frac{\partial}{\partial \beta}(y - X\beta)^T(y - X\beta) = \frac{\partial}{\partial \beta} y^T y - \underbrace{\beta^T X^T y - y^T X \beta}_{\text{These are both scalar}} + \beta^T X^T X \beta$$

$$= \frac{\partial}{\partial \beta} y^T y - 2\beta^T X^T y + \beta^T X^T X \beta = -2X^T y + 2X^T X \beta$$

Note: second derivative is $2X^TX > 0$ so this is a minimum

set to zero $2X^TX\beta = 2X^Ty \Rightarrow X^TX\beta = X^Ty \Rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty$

② We have that $y \sim MVN(X\beta, \sigma^2 I)$, using this as an equivalent way to write the likelihood $\prod p(y_i|\beta,\sigma^2)$ we have

$$\arg\max_{\beta \in \mathbb{R}^p}\left\{\prod p(y_i|\beta,\sigma^2)\right\} = \arg\max_{\beta \in \mathbb{R}^p}\left\{(2\pi)^{-n/2}|\sigma^2 I|^{-1/2} e^{\frac{-1}{2\sigma^2}(y-X\beta)^T(y-X\beta)}\right\}$$

$$= \arg\max_{\beta \in \mathbb{R}^p}\left\{\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right\} = \arg\min_{\beta \in \mathbb{R}^p}\left\{(y-X\beta)^T(y-X\beta)\right\}$$

we showed in the previous part that this yields $\hat{\beta} = (X^TX)^{-1}X^TY$

③ We are assuming that all of the variables have had their sample means subtracted so $E(X_{ij}) = 0$ we will also assume that the residuals are centered at zero, thus $E(\varepsilon_j) = 0$.

we want to choose β s.t. $\text{Cov}(\varepsilon, x_i) = 0$ for $i = 1, \dots, p$

$$\text{Cov}(\varepsilon, x_i) = E\left[(x_i^T - E(x_i^T))(\varepsilon - E(\varepsilon))\right] = E\left[x_i^T \varepsilon\right] = E\left[\begin{bmatrix} x_{i1} & \cdots & x_{in}\end{bmatrix}\begin{bmatrix}\varepsilon_1 \\ \vdots \\ \varepsilon_n\end{bmatrix}\right]$$

$= E\left[x_i^T(y - X\beta)\right]$ for $i = 1, \dots, p \Rightarrow$ the $i = 1, \dots, p$ can be summarized to

$E\left[X^T(y - X\beta)\right] = E\left[X^Ty - X^TX\beta\right]$ If we want this to be zero then we need $\hat{\beta} = (X^TX)^{-1}X^TY$

$\Rightarrow E\left[X^Ty - X^TX\beta\right] = E\left[X^Ty - X^TX(X^TX)^{-1}X^TY\right] = E[0] = 0$

edited on prev. page

(A) Show that all three of these principles lead to the same estimator. What is the variance of this estimator under the assumption that each ϵ_i is independent and identically distribution with variance σ^2?

$$Var(\hat{\beta}) = Var(\hat{\beta} - \beta) = Var\left((X^TX)^{-1}X^TY - \beta\right)$$

$$= Var\left((X^TX)^{-1}X^T(X\beta + \varepsilon) - \beta\right) = Var\left(\beta + (X^TX)^{-1}X^T\varepsilon - \beta\right)$$

$$= Var\left((X^TX)^{-1}X^T\varepsilon\right) = (X^TX)^{-1}X^T\, Var(\varepsilon)\left[(X^TX)^{-1}X^T\right]^T$$

$$= (X^TX)^{-1}X^T(\sigma^2)\times(X^TX)^{-1} = \sigma^2\left[(X^TX)^{-1}X^T\times(X^TX)^{-1}\right]$$

$$= \sigma^2(X^TX)^{-1}$$

$$\Rightarrow Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$$

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

We have that $y \sim MVN(X\beta, \Sigma)$, using this as an equivalent way to write the likelihood $\prod p(y_i|\beta,\sigma^2)$ we have

$$\arg\max_{\beta\in\mathbb{R}^p}\left\{\prod p(y_i|\beta,\Sigma)\right\} = \arg\max_{\beta\in\mathbb{R}^p}\left\{(2\pi)^{-n/2}|\Sigma|^{-1/2} e^{\frac{-1}{2}(y-x\beta)^T\Sigma^{-1}(y-x\beta)}\right\}$$

$$= \arg\max_{\beta\in\mathbb{R}^p}\left\{\exp\left(-\frac{1}{2}(y-x\beta)^T\Sigma^{-1}(y-x\beta)\right\} = \arg\min_{\beta\in\mathbb{R}^p}\left\{(y-x\beta)^T\Sigma^{-1}(y-x\beta)\right\}$$

We find the min by setting the partial derivative wrt β to 0.

$$\frac{\partial}{\partial\beta}(y-X\beta)^T\Sigma^{-1}(y-x\beta) = \frac{\partial}{\partial\beta}(y^T-\beta^Tx^T)\Sigma^{-1}(y-x\beta)$$

$$= \frac{\partial}{\partial\beta}(y^T\Sigma^{-1}-\beta^Tx^T\Sigma^{-1})(y-x\beta) = \frac{\partial}{\partial\beta}\, y^T\Sigma^{-1}y - \underbrace{y^T\Sigma^{-1}x\beta - \beta^Tx^T\Sigma^{-1}y}_{\text{These are both scalar}} + \beta^Tx^T\Sigma^{-1}x\beta$$

$$= \frac{\partial}{\partial\beta}\, y^T\Sigma^{-1}y - 2\beta^Tx^T\Sigma^{-1}y + (x\beta)^T\Sigma^{-1}(x\beta) = -2x^T\Sigma^{-1}y + 2x^T\Sigma^{-1}x\beta$$

Setting to zero $\Rightarrow 2x^T\Sigma^{-1}x\beta = 2x^T\Sigma^{-1}y$

$$\Rightarrow x^T\Sigma^{-1}x\beta = x^T\Sigma^{-1}y \Rightarrow \boxed{\hat{\beta} = (x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}y}$$

Note: 2nd derivative is $2x^T\Sigma^{-1}x > 0$ so this is a minimum

$$Var(\hat{\beta}) = Var(\hat{\beta}-\beta) = Var\left((x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}y - \beta\right)$$

$$= Var\left((x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}(x\beta+\varepsilon) - \beta\right) = Var\left(\beta + (x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}\varepsilon - \beta\right)$$

$$= Var\left((x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}\varepsilon\right) = (x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}\,Var(\varepsilon)\left[(x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}\right]^T$$

$$= (x^T\Sigma^{-1}x)^{-1}x^T\underbrace{\Sigma^{-1}(\Sigma)}\left[\Sigma^{-1}x(x^T\Sigma^{-1}x)^{-1}\right] = (x^T\Sigma^{-1}x)^{-1}\underbrace{x^T\Sigma^{-1}x(x^T\Sigma^{-1}x)^{-1}}$$

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

$$\boxed{Var(\hat{\beta}) = (x^T\Sigma^{-1}x)^{-1}}$$

Econometrics

Huber-White

(C) Show that in the special case where Σ is a diagonal matrix, i.e. $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$, that the MLE is the familiar *weighted least squares* estimator.

We have from part (b) that $\hat{\beta} = (x^T \Sigma^{-1} x)^{-1} x^T \Sigma^{-1} y$

Now let $\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$ then let $W = \Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_n^2} \end{bmatrix}$

this means that the MLE for $\hat{\beta}$ becomes

$\hat{\beta} = (x^T W x)^{-1} x^T W y$ which we recognize as the familiar weighted least squares estimator.

least squares objective

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2$$

4 Some practical details

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

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

One way to calculate $\hat{\beta}$ is to: (1) recognize that, trivially, the solution to the above linear system must satisfy $\hat{\beta} = (X^T W X)^{-1} X^T W y$; and (2) to calculate this directly, i.e. by inverting $X^T W X$. Let's call this the "inversion method" for calculating the WLS solution.

Numerically speaking, is the inversion method the fastest and most stable way to actually solve the above linear system? Do some independent sleuthing on this question.[9]. Summarize what you find, and provide pseudo-code for at least one alternate method based on matrix factorizations—call it "your method" for short.[10]

We turn to Gunderson's blog post (http://gregorygundersen.com/blog/2020/12/09/matrix-inversion/) for insight. Gunderson suggests that solving a linear system with inversion has a much higher computation cost than using LU decomposition.

LU decomposition requires $\sim \frac{2}{3}n^3 + 2n^2 \quad \text{flops.}$

While inverting requires $\sim 2n^3 \quad \text{flops.}$

This shows us that the inversion method is roughly three times as expensive as LU decomposition when n is large, which means it is not the fastest way to solve the system.

Turning to the second part of the question, inversion is not a stable way to solve the system either since X'WX could be an ill-conditioned matrix or a singular matrix which would lead to instabilities or the need for a generalized inverse which is not unique.

Pseudo-code:

Goal: Solve $X'WX\,\hat{\beta} = X'Wy$ for $\hat{\beta}$

1. Factor $X'WX$ into lower (L) and upper (U) triangular matrices using LU decomposition.
   $LU\hat{\beta} = X'Wy$
2. Solve the following for z using forward substitution
   $Lz = X'Wy$
3. Solve the following for $\hat{\beta}$ using backward substitution
   $U\hat{\beta} = z$

(B) Code up functions that implement both the inversion method and your method for an arbitrary X, y, and set of weights W. Obviously you shouldn't write your own linear algebra routines for doing things like multiplying or decomposing matrices. But don't use a direct model-fitting function like R's "lm" either. Your actual code should look a lot like the pseudo-code you wrote for the previous part.[11]

Now simulate some silly data from the linear model for a range of values of N and P. (Feel free to assume that the weights w_i are all 1.) It doesn't matter how you do this—e.g. everything can be Gaussian if you want. (We're not concerned with statistical principles in this problem, just with algorithms, and using least squares is a pretty terrible idea for enormous linear models, anyway.) Just make sure that you explore values of P up into the thousands, and that $N > P$. Benchmark the performance of the inversion solver and your solver across a range of scenarios.[12]

For $N = c(10, 50, 100, 200)$ and $P = c(2, 10, 50, 100)$ the following initializations were used

$W = I_n$

$X_{ii} \sim N(0,1)$

$y \sim N(2 * X[,1] + 3 * X[,2], 1)$

The table below summarizes the mean execution times for the two methods

N	P	Inversion (ms)	LU Decomposition
10	2	0.267	0.128
100	50	250.50	14.262
500	100	4607.16	133.32