# Lesson 1 · SDS 383D
# Exercises 1: Preliminaries

Yunshan Duan

## 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]

$\star$ The posterior distribution

$$\begin{aligned}
p(w|x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N|w)p(w) \\
&= \prod_{i=1}^{N} w^{1(x_i=1)}(1-w)^{1-1(x_i=1)} \cdot w^{a-1}(1-w)^{b-1} \\
&= w^k(1-w)^{N-k}w^{a-1}(1-w)^{b-1} \\
&= w^{a+k-1}(1-w)^{b+N-k-1} \\
&\equiv Beta(a+k, b+N-k) \\
&= \frac{\Gamma(a+b+N)}{\Gamma(a+k)\Gamma(b+N-k)} w^{a+k-1}(1-w)^{b+N-k-1},
\end{aligned}$$

where $k = \sum_{i=1}^{N} 1(x_i = 1)$.

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

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

[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.

Suppose that $x_1 \sim \mathrm{Ga}(a_1, 1)$ and that $x_2 \sim \mathrm{Ga}(a_2, 1)$. 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.

⋆ The PDF transformation

$$p(\boldsymbol{y}) = p(g^{-1}(\boldsymbol{x}))\,|J(\boldsymbol{y})|\,,$$

when $\boldsymbol{y} = g(\boldsymbol{x})$.

$$y_1 = \frac{x_1}{x_1+x_2}, \quad y_2 = x_1 + x_2.$$

Then,

$$x_1 = y_1 y_2, \quad x_2 = y_2 - y_1 y_2.$$

The Jacobian

$$\begin{aligned} |J(\boldsymbol{y})| &= \left| \begin{pmatrix} y_2 & y_1 \\ -y_2 & 1-y_1 \end{pmatrix} \right| \\ &= |y_2(1-y_1) + y_1 y_2| \\ &= |y_2| = y_2 \end{aligned}$$

Therefore, the joint density for $(y_1, y_2)$ is

$$\begin{aligned} p_{\boldsymbol{y}}(y_1, y_2) &= p_{\boldsymbol{x}}(x_1, x_2)\,|J(\boldsymbol{y})| \\ &= \frac{1}{\Gamma(a_1)} x_1^{a_1-1} \exp(-x_1) \frac{1}{\Gamma(a_2)} x_2^{a_2-1} \exp(-x_2) y_2 \\ &= \frac{1}{\Gamma(a_1)} (y_1 y_2)^{a_1-1} \exp(-y_1 y_2) \frac{1}{\Gamma(a_2)} y_2^{a_2-1} (1-y_2)^{a_2-1} \exp(-y_2 + y_1 y_2) y_2 \\ &= \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)} (y_1)^{a_1-1} (1-y_1)^{a_2-1} \frac{1}{\Gamma(a_1+a_2)} y_2^{a_1+a_2} \exp(-y_2) \\ &\equiv \mathit{Beta}(y_1; a_1, a_2) \cdot \mathit{Ga}(y_2; a_1+a_2, 1) \end{aligned}$$

The marginals are

$$p(y_1) \equiv \mathit{Beta}(a_1, a_2),$$

$$p(y_2) \equiv \mathit{Ga}(a_1 + a_2, 1).$$

If we want to simulate $\mathit{Beta}(a_1, a_2)$, we can generate $x_1 \sim \mathit{Ga}(a_1, 1)$, $x_2 \sim \mathit{Ga}(a_2, 1)$, then $y = \frac{x_1}{x_1+x_2} \sim \mathit{Beta}(a_1, a_2)$.

[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`.

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

⋆

$$x_1, \ldots x_N \overset{\text{iid}}{\sim} N(\theta, \sigma^2), \quad \sigma^2 \text{ known},$$
$$\theta \sim N(m, v).$$

The posterior is

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

(D) Suppose that we take independent observations $x_1, \ldots, x_N$ from a normal sampling model with *known* mean $\theta$ but *unknown* variance $\sigma^2$. (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 $\omega$ has a gamma prior with parameters $a$ and $b$, implying that $\sigma^2$ 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 $\sigma^2$, the variance.

⋆

$$x_1, \ldots x_N \overset{\text{iid}}{\sim} N(\theta, \sigma^2), \quad \theta \text{ known},$$
$$p(x_1|\theta, w) = (\frac{w}{2\pi}) \exp\{-\frac{w}{2}(x_i - \theta)^2\},$$
$$p(w) = Ga(a, b) \propto w^{a-1} \exp(-bw).$$

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

$$\begin{aligned}
p(w|x_1,\ldots,x_N) &\propto p(x_1,\ldots,x_N|w)p(w)\\
&\propto w^{\frac{N}{2}}\exp\{-\frac{w}{2}\sum_{i=1}^{N}(x_i-\theta)^2\}\cdot w^{a-1}\exp(-bw)\\
&= w^{a+\frac{N}{2}-1}\exp\{-w[b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2]\}\\
&\equiv Ga(a+\frac{N}{2},b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2)\\
&= \frac{[b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2]^{a+\frac{N}{2}}}{\Gamma(a+\frac{N}{2})}w^{a+\frac{N}{2}-1}\exp\{-w[b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2]\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
p(\sigma^2|x_1,\ldots,x_N) &\propto p(x_1,\ldots,x_N|\sigma^2)p(\sigma^2)\\
&\equiv \text{Inv-Ga}(a+\frac{N}{2},b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2)\\
&= \frac{[b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2]^{a+\frac{N}{2}}}{\Gamma(a+\frac{N}{2})}(\sigma^2)^{-a-\frac{N}{2}-1}\exp\{-\frac{1}{\sigma^2}[b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2]\}.
\end{aligned}$$

(E) Suppose that, as above, we take independent observations $x_1,\ldots,x_N$ from a normal sampling model with unknown, common mean $\theta$. This time, however, each observation has its own idiosyncratic (but known) variance: $x_i \sim \mathrm{N}(\theta,\sigma_i^2)$. Suppose that $\theta$ 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.

⋆

$$x_1,\ldots x_N \overset{\text{iid}}{\sim} N(\theta,\sigma_i^2),\quad \sigma^2 \text{ known},$$
$$\theta \sim N(m,v).$$

$$\begin{aligned}
p(\theta|x_1,\dots,x_N) &\propto p(x_1,\dots,x_N|\theta)p(\theta) \\
&\propto \prod_{i=1}^{N} \exp\{-\frac{1}{2\sigma_i^2}(x_i-\theta)^2\}\cdot\exp\{-\frac{1}{2v^2}(\theta-m)^2\} \\
&\propto \exp\{-\frac{1}{2}[\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(x_i-\theta)^2+\frac{1}{v}(\theta-m)^2]\} \\
&\propto \exp\{-\frac{1}{2}[(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}+\frac{1}{v})\theta^2-2(\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}+\frac{m}{v})\theta]\} \\
&\equiv N(\mu,\sigma^2),
\end{aligned}$$

where $\sigma^2 = (\sum_{i=1}^{N}\frac{1}{\sigma_i^2}+\frac{1}{v})^{-1}$, $\mu = \sigma^2(\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}+\frac{m}{v})$. Therefore, the posterior mean in a form that is clearly interpretable as a weighted average of the observations and the prior mean.

(F) Suppose that $(x \mid \omega) \sim \mathrm{N}(m, \omega^{-1})$, and that $\omega$ 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.*

$\star$

$$\begin{aligned}
x|w &\sim N(m, w^{-1}), \\
w &\sim Ga(\frac{a}{2},\frac{b}{2}).
\end{aligned}$$

The marginal distribution of $x$ is

$$\begin{aligned}
p(x) &= \int p(x|w)p(w)dw \\
&\propto \int (2\pi w^{-1})^{\frac{1}{2}}\exp\{-\frac{1}{2w^{-1}}(x-m)^2\}w^{\frac{a}{2}-1}\exp\{-\frac{b}{2}w\}dw \\
&\propto \int w^{\frac{a}{2}+\frac{1}{2}-1}\exp\{-(\frac{b}{2}+\frac{(x-m)^2}{2})w\}dw \\
&\propto \int \frac{\Gamma(\frac{a}{2}+\frac{1}{2})}{(\frac{b}{2}+\frac{(x-m)^2}{2})^{\frac{a}{2}+\frac{1}{2}}}w^{\frac{a}{2}+\frac{1}{2}-1}\exp\{-(\frac{b}{2}+\frac{(x-m)^2}{2})w\}dw\cdot\frac{(\frac{b}{2}+\frac{(x-m)^2}{2})^{\frac{a}{2}+\frac{1}{2}}}{\Gamma(\frac{a}{2}+\frac{1}{2})} \\
&\propto [b+(x-m)^2]^{-\frac{a+1}{2}} \\
&\propto [1+\frac{1}{a}(x-m)^2/(\frac{b}{a})]^{-\frac{a+1}{2}} \\
&\equiv t(a, m, (\frac{b}{a})^{1/2}).
\end{aligned}$$

# 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 $x \sim \mathrm{N}(m, v)$.

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 $E\left\{\exp(tx)\right\} = \exp(mt + vt^2/2)$ 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 $\mathrm{cov}(x)$ 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, $\mathrm{cov}(x) = E\{(x-\mu)(x-\mu)^T\}$, where $\mu$ is the mean vector whose $i$th component is $E(x_i)$. Prove the following: (1) $\mathrm{cov}(x) = E(xx^T) - \mu\mu^T$; and (2) $\mathrm{cov}(Ax + b) = A\mathrm{cov}(x)A^T$ for matrix $A$ and vector $b$.

⋆

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

$$\begin{aligned}
cov(Ax+b) &= E\{(Ax+b)(Ax+b)^T\} - (A\mu+b)(A\mu+b)^T \\
&= E\{Axx^TA^T + Axb^T + bx^TA^T + bb^T\} - (A\mu\mu^TA^T + b\mu^TA^T + A\mu b^T + bb^T) \\
&= AE\{xx^T\}A^T + A\mu b^T + b\mu^TA^T + bb^T - (A\mu\mu^TA^T + b\mu^TA^T + A\mu b^T + bb^T) \\
&= AE\{xx^T\}A^T - A\mu\mu^TA^T \\
&= Acov(x)A^T
\end{aligned}$$

(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.

[4] Laplace transforms to everybody but statisticians.

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

⋆ The probability density function

$$\begin{aligned} p(z) &= \prod_{i=1}^{p} (2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2} z_i^2\} \\ &= (2\pi)^{-\frac{p}{2}} \exp\{-\frac{1}{2} zz^T\} \end{aligned}$$

The moment generating function is

$$\begin{aligned} E\{\exp(t^T z)\} &= E\{\prod_{i=1}^{p} \exp(t_i z_i)\} \\ &= \prod_{i=1}^{p} \exp(\frac{t_i^2}{2}) \\ &= \exp(\frac{1}{2} \sum_{i=1}^{p} t_i^2) \\ &= \exp(\frac{tt^T}{2}) \end{aligned}$$

(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 $x \sim \mathrm{N}(\mu, \Sigma)$, 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 $x$?

⋆ If $x \sim N(\mu, \Sigma)$, then $\forall a \neq 0$, $a^T x$ is a univariate normal distribution.

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

$$Var(a^T x) = a^T cov(x) a = a^T \Sigma a.$$

Then, $a^T x \sim N(a^T \mu, a^T \Sigma a)$, it's moment generating function is

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

Let $t = ta$, therefore, the MGF is

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

If the MGF is

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

then for all $a \neq 0$, $t \in \mathbb{R}$,

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

Therefore, $a^T x$ is univariate normal.

(D) Another basic theorem is that a random vector is multivariate normal if and only if it is an affine transformation of independent univariate normals. You will first prove the "if" statement. Let $z$ have a standard multivariate normal distribution, and define the random vector $x = Lz + \mu$ 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 $x$.

$\star$ If $z \sim N(0, I)$, let $x = Lz + \mu$. Then, the moment generating function of $x$ is

$$\begin{aligned} E\{exp(t^T x)\} &= E\{exp(t^T Lz + t^T \mu)\} \\ &= E\{exp((L^T t)^T z)\} \cdot exp(t^T \mu) \\ &= exp\{\frac{1}{2}(L^T t)^T (L^T t)\} \cdot exp(t^T \mu) \\ &= exp\{\frac{1}{2} t^T LL^T t + t^T \mu\} \end{aligned}$$

Based on the conclusion in (C) and the uniqueness of the MGF, therefore, $x = Lz\mu \sim N(\mu, \Sigma = LL^T)$.

(E) Now for the "only if." Suppose that $x$ has a multivariate normal distribution. Prove that $x$ can be written as an affine transformation of standard normal random variables. (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.

$\star$ If $x \sim N(\mu, \Sigma)$, $\Sigma$ is the covariance matrix which is positive-semi-definite, then it can be decomposed as $\Sigma = AA^T$. The MGF of $x$ is

$$\begin{aligned} E\{exp(t^T x)\} &= exp\{\frac{1}{2} t^T \Sigma t + t^T \mu\} \\ &= exp\{\frac{1}{2} t^T AA^T t + t^T \mu\} \end{aligned}$$

Let $z \sim N(0, I)$, $y = Az + \mu$. Then based on (D), the MGF of $y$ is

$$E\{exp(t^T y)\} = exp\{\frac{1}{2} t^T AA^T t + t^T \mu\} = E\{exp(t^T x)\}$$

Because the uniqueness of MGF, $x = y = Az + \mu$.

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

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

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

$\star$ If $x \sim N(\mu, \Sigma)$, then $x = Az + \mu$, where $z \sim N(0, I)$, and $\Sigma = AA^T$. Therefore, the pdf of $x$ is

$$\begin{aligned}
p(x) &= p(x = Az + \mu)|A^{-1}| \\
&= (2\pi)^{-\frac{p}{2}} \exp\{-\frac{1}{2}(A^{-1}(x-\mu))^T(A^{-1}(x-\mu))\}|A^{-1}| \\
&= (2\pi)^{-\frac{p}{2}}|A^{-1}| \exp\{-\frac{1}{2}(x-\mu)^T(A^TA)^{-1}(x-\mu)\} \\
&= (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}
\end{aligned}$$

(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 $y$.

$\star$ The MGF of $y$ is

$$\begin{aligned}
E\{exp(t^Y y)\} &= E\{exp[t^T(Ax_1 + Bx_2)]\} \\
&= E\{exp((A^Tx)^Tx_1)\} \cdot E\{exp((B^Tt)^Tx_2)\} \\
&= exp\{\frac{1}{2}(A^Tt)^T\Sigma_1(A^Tt) + (A^Tt)^T\mu_1\} \cdot exp\{\frac{1}{2}(B^Tt)^T\Sigma_2(B^Tt) + (B^Tt)^T\mu_2\} \\
&= exp\{\frac{1}{2}t^T(A\Sigma_1A^T + B\Sigma_2B^T)t + t^T(A\mu_1 + B\mu_2)\}
\end{aligned}$$

Therefore,

$$y \sim N(A\mu_1 + B\mu_2, A\Sigma_1A^T + B\Sigma_2B^T).$$

## 2.2 Conditionals and marginals

Suppose that $x \sim \mathrm{N}(\mu, \Sigma)$ 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 $\mu$ and $\Sigma$ have been partitioned conformably with $x$:

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

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

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

$\star$

$$x_1 = [I|0]_{k \times p}\, x.$$

Let $A = [I|0]_{k\times p}$, then $x_1 = Ax$. Then,

$$x_1 \sim N(A\mu, A\Sigma A^T),$$

$$A\mu = \mu_1,$$

$$A\Sigma A^T = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} = \Sigma_{11}.$$

Therefore,

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

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

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

Using (or deriving!) identities for the inverse of a partitioned matrix, express each block of $\Omega$ in terms of blocks of $\Sigma$.

$\star$ Denote

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

and

$$\Omega = \Sigma^{-1} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}.$$

We have

$$\Sigma\Omega = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} = \begin{bmatrix} \Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{21} & \Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} \\ \Sigma_{21}\Omega_{11} + \Sigma_{22}\Omega_{21} & \Sigma_{12}\Omega_{12} + \Sigma_{22}\Omega_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

$$\Sigma_{11}^{-1}(\Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22}) = 0,$$
$$\Sigma_{22}^{-1}(\Sigma_{21}\Omega_{11} + \Sigma_{22}\Omega_{21}) = 0.$$

Then,

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

Then we have,

$$\Sigma_{11}\Omega_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}\Omega_{11} = I,$$
$$-\Sigma_{12}\Sigma_{11}^{-1}\Sigma_{12}\Omega_{22} + \Sigma_{22}\Omega_{22} = I.$$

Therefore,

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

(C) Derive the conditional distribution for $x_1$, given $x_2$, in terms of the partitioned elements of $x$, $\mu$, and $\Sigma$. 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.

$\star$ The conditional distribution

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{p(x)}{p(x_2)} \propto p(x).$$

The log pdf is

$$\begin{aligned}
\log p(x_1|x_2) &\propto \log p(x) \\
&\propto \log (2\pi)^{\frac{p}{2}} |\Sigma|^{-1} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\} \\
&\propto -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \\
&= -\frac{1}{2}\begin{bmatrix}(x_1-\mu_1)^T & (x_2-\mu_2)^T\end{bmatrix}\begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}\begin{bmatrix}x_1-\mu_1 \\ x_2-\mu_2\end{bmatrix} \\
&= -\frac{1}{2}\begin{bmatrix}(x_1-\mu_1)^T\Omega_{11} + (x_2-\mu_2)^T\Omega_{21} & (x_1-\mu_1)^T\Omega_{12} + (x_2-\mu_2)^T\Omega_{22}\end{bmatrix}\begin{bmatrix}x_1-\mu_1 \\ x_2-\mu_2\end{bmatrix} \\
&= -\frac{1}{2}\{(x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + (x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) \\
&+ (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2) + (x_2-\mu_2)^T\Omega_{22}(x_2-\mu_2)\} \\
&= -\frac{1}{2}\{x_1^T\Omega_{11}x_1 - 2x_1^T(\Omega_{11}\mu_1 + \frac{1}{2}\Omega_{21}^T\mu_2 + \frac{1}{2}\Omega_{12}\mu_2 - \frac{1}{2}\Omega_{21}^T x_2 - \frac{1}{2}\Omega_{12}x_2)\} \\
&= -\frac{1}{2}\{x_1^T\Omega_{11}x_1 - 2x_1^T(\Omega_{11}\mu_1 + \Omega_{12}\mu_2 - \Omega_{12}x_2)\}
\end{aligned}$$

Therefore,

$$\begin{aligned}
p(x_1|x_2) &\equiv N(\Omega_{11}^{-1}(\Omega_{11}\mu_1 + \Omega_{12}\mu_2 - \Omega_{12}x_2), \Omega_{11}^{-1}) \\
&\equiv N(\mu_1 + \Omega_{11}^{-1}\Omega_{12}\mu_2 - \Omega_{11}^{-1}\Omega_{12}x_2, \Omega_{11}^{-1})
\end{aligned}$$

## 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;

[8] In scalar form:

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

and the $\epsilon_i$ 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 $\beta$.[9]

Consider three classic inferential principles that are widely used to estimate $\beta$, the vector of regression coefficients. In this context we will let $\hat{\beta}$ denote an estimate of $\beta$, $y = (y_1, \ldots, y_n)^T$ the vector of outcomes, $X$ the matrix of predictors whose ith row is $x_i^T$, and $\epsilon$ 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 $\sigma^2$. 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 $\beta$ and $\sigma^2$. 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 $\epsilon$ and each column of $X$ is 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. What is the variance of this estimator under the assumption that each $\epsilon_i$ is independent and identically distribution with variance $\sigma^2$?

### $\star$ Maximum likelihood

$$\begin{aligned}
\hat{\beta} &= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^n p(y_i \mid \beta, \sigma^2) \right\} \\
&= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \log \prod_{i=1}^n p(y_i \mid \beta, \sigma^2) \right\} \\
&= \arg\max_{\beta \in \mathcal{R}^p} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i \beta)^2 \right\} \\
&= \arg\max_{\beta \in \mathcal{R}^p} \left\{ -\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta) \right\} \\
&= \arg\min_{\beta \in \mathcal{R}^p} \left\{ (Y - X\beta)^T (Y - X\beta) \right\}
\end{aligned}$$

[9] 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 $\beta_1$ will be an intercept term.

**Method of moments** Denote $X = (x_1, \dots, x_j, \dots, x_p)$, where $x_j$ is the $j$th column of $X$. The sample covariance of $\epsilon$ and each column of $X$ is zero, that is, $\forall j = 1, \dots, p$

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

We have $\bar{e} = 0$, therefore, $\forall j = 1, \dots, p$,

$$e^T x_j = 0,$$

which is

$$e^T X = 0.$$

Then,

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

This is the solution to

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

The variance of the estimator is

$$\begin{aligned}
Cov(\hat{\beta}) &= Cov((X^T X)^{-1} X^T Y) \\
&= (X^T X)^{-1} X^T Cov(Y)((X^T X)^{-1} X^T)^T \\
&= (X^T X)^{-1} X^T (\sigma^2 I) X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

(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 $\Sigma$ is an arbitrary known covariance matrix, not necessarily proportional to the identity? What is the maximum likelihood estimate for $\beta$ now, and what is the variance of this estimator?

$$\begin{aligned}
\hat{\beta} &= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^{n} p(y_i \mid \beta, \sigma^2) \right\} \\
&= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \log \prod_{i=1}^{n} p(y_i \mid \beta, \sigma^2) \right\} \\
&= \arg\max_{\beta \in \mathcal{R}^p} \left\{ -\frac{1}{2}(Y - X\beta)^T \Sigma^{-1} (Y - X\beta) \right\} \\
&= \arg\min_{\beta \in \mathcal{R}^p} \left\{ (Y - X\beta)^T \Sigma^{-1} (Y - X\beta) \right\} \\
&= \arg\min_{\beta \in \mathcal{R}^p} \left\{ \beta^T X^T \Sigma^{-1} X\beta - 2\beta^T X^T \Sigma^{-1} Y \right\}
\end{aligned}$$

Then,

$$\begin{aligned}
2X^T\Sigma^{-1}X\hat{\beta} - 2X^T\Sigma^{-1}Y &= 0 \\
\hat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y
\end{aligned}$$

The variance is

$$\begin{aligned}
Cov(\hat{\beta}) &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Cov(Y)((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1})^T \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\Sigma\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\
&= (X^T\Sigma^{-1}X)^{-1}
\end{aligned}$$

(C) Show that in the special case where $\Sigma$ 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. That is, show that $\hat{\beta}$ is the solution to the following linear system of $P$ equations in $P$ unknowns:

$$(X^TWX)\hat{\beta} = X^TWy\,,$$

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

$$\begin{aligned}
\hat{\beta} &= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^{n} p(y_i \mid \beta, \sigma^2) \right\} \\
&= \arg\min_{\beta \in \mathcal{R}^p} \left\{ (Y - X\beta)^T \Sigma^{-1} (Y - X\beta) \right\} \\
&= \arg\min_{\beta \in \mathcal{R}^p} \left\{ \frac{1}{\sigma_i^2}(y_i - x_i^T\beta)^2 \right\}
\end{aligned}$$

As is written in (B)

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^p} \left\{ \beta^T X^T \Sigma^{-1} X\beta - 2\beta^T X^T \Sigma^{-1} Y \right\}$$

$$\begin{aligned}
2X^T\Sigma^{-1}X\hat{\beta} - 2X^T\Sigma^{-1}Y &= 0 \\
X^T\Sigma^{-1}X\hat{\beta} &= X^T\Sigma^{-1}Y \\
(X^TWX)\hat{\beta} &= X^TWY
\end{aligned}$$

where $W = diag(1/\sigma_1^2, \ldots, 1/\sigma_n^2)$.

# 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

$$(X^T WX)\hat{\beta} = X^T Wy\,,$$

One way to calculate $\hat{\beta}$ is to: (1) recognize that, trivially, the solution to the above linear system must satisfy $\hat{\beta} = (X^T WX)^{-1} X^T Wy$; and (2) to calculate this directly, i.e. by inverting $X^T WX$. 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.[10]. Summarize what you find, and provide pseudo-code for at least one alternate method based on matrix factorizations—call it "your method" for short.[11]

⋆ Is the inversion method the fastest and most stable way to actually solve the above linear system?

Based on the blog, the total complexity of solving a linear system of equations using the LU decomposition is $\frac{2}{3}n^3 + 2n^2$, while using the inverse of the matrix requires $2n^3$ flops.

Solving a system of linear equations by performing a matrix inversion is typically less accurate than solving the system directly. The matrix inversion approach can have significantly worse backward error than directly solving for $x$ if the matrix $A$ is ill-conditioned.

We can use LU decomposition to make the calculation of matrix inverse easier. Denote the matrix to be inverse as $A$. LU decomposition is decomposing

$$A = LU,$$

where $L$ is a lower triangular matrix and $U$ is an upper triangular matrix. If we want to solve

$$Ax = LUx = b$$

(a) Use forward substitution to solve $Ly = b$.

(b) Use backward substitution to solve $Ux = y$.

**Algorithm 1** Solve weighted least squares estimator using LU decomposition

```
Let $A = X^T WX$, $b = X^T Wy$.
Use forward substitution to solve $Ly = b$.
Use backward substitution to solve $Ux = y$. Then $x$ is what we want for $\hat{\beta}$.
```

[10] `https://www.google.com/search?q=Why+Shouldn\%27t+I+Invert+That+Matrix`

[11] Our linear system is not a special flower; whatever you discover about general linear systems should apply here.

| N | P | solve() (s) | LU decomp (s) | inv() |
|---|---|---|---|---|
| 10 | 5 | $31.9226 \times 10^{-6}$ | $34.0464 \times 10^{-6}$ | $268.2630 \times 10^{-6}$ |
| 60 | 50 | $333.4448 \times 10^{-6}$ | $6834.8066 \times 10^{-6}$ | $91702.9042 \times 10^{-6}$ |
| 200 | 100 | $4.413174 \times 10^{-3}$ | $53.850031 \times 10^{-3}$ | $2074.851510 \times 10^{-3}$ |
| 1000 | 900 | 1.15714 | 35.65266 | |

Table 1: Performance

(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.[12]

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.[13]

⋆ Generate data

$$W = I$$

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

$$\beta_j \sim N(0,1)$$

We have the results in Table 1. Unfortunately, my computer memory is exhausted when running inv() with $P = 900, N = 1000$. And the solve() function in R is super smart. But we can still tell that the inv() function in R which uses the basic matrix inverse takes much longer time than our method using LU decomposition.

[12] Be attentive to how you multiply a matrix by a diagonal matrix, or you'll waste a lot of time multiplying stuff by zero.

[13] In R, a simple library for this purpose is microbenchmark.