Preliminaries

Bayesian inference in simple conjugate families

(A)

We have N independent observations x_i from a Bernoulli sampling model with unknown probability w. Moreover, a $Beta(a, b)$ prior distribution is chosen for w. The posterior distribution will be obtained with the following passages:

$$\begin{aligned} p(w|x_1, \dots x_N) &= \frac{p(x_1 \dots x_N \mid w) \cdot p(w)}{p(x_1 \dots x_N)} \\ &\propto \prod_{i=1}^{N} w^{x_i}(1-w)^{1-x_i} \cdot w^{a-1}(1-w)^{b-1} \\ &= w^{\sum^N xi}(1-w)^{N-\sum x_i} w^{a-1}(1-w)^{b-1} \\ &= w^{\sum x_i + a - 1}(1-w)^{N-\sum x_i + b - 1} \end{aligned}$$

$\Rightarrow$ kernel of a $Beta\left(\sum x_i + a, N - \sum x_i + b\right)$.

(B)

Suppose that x1 $\sim Ga(a_1, 1)$ and that $x2 \sim Ga(a_2, 1)$. Define two new random variables $y_1 = x_1/(x_1 + x_2)$ and $y_2 = x_1 + x_2$. I have

$$\begin{cases} y_1 = \dfrac{x_1}{x_1 + x_2} \\ y_2 = x_1 + x_2 \end{cases} \rightarrow \begin{cases} y_1 = y_1 \cdot y_2 \\ x_2 = y_2 - y_1 \cdot y_2 \end{cases}$$

with support:

$$y_1 \in (0, 1) \qquad y_2 \in (0, +\infty)$$

and the Jacobian and related determinant are:

$$\begin{bmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{bmatrix} \Rightarrow |det| = |y_2 (1 - y_1)| = |y_2 (1 - y_1) + y_2 y_1|$$

so that applying the formula for the joint distribution of the transformed variables, we have:

$$\begin{aligned} f(y_1, y_2) &= \frac{1}{\Gamma(a_1)} (y_1 \cdot y_2)^{a_1 - 1} \exp(-y_1 \cdot y_2) \cdot \frac{1}{\Gamma(a_2)} (y_2 (1 - y_1))^{a_2 - 1} \exp(y_2 (y_1 - 1)) \cdot |y_2| \\ &= \frac{1}{\Gamma(a_1) \cdot \Gamma(a_2)} (y_1 \cdot y_2)^{a_1 - 1} \cdot (y_2 (1 - y_1))^{a_2 - 1} \cdot y_2 \cdot \exp(-y_2) \\ &= \frac{1}{\Gamma(a_1) \cdot \Gamma(a_2)} \cdot (y_1 \cdot y_2) \cdot (-y_2 \cdot y_1)^{a_2 - 1} \cdot y_2^{a_2 - 1} \cdot y_2 \cdot \exp(-y_2) \\ &= \frac{1}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1 - 1} (1 - y_1)^{a_2 - 1} (y_2)^{a_1 + a_2 - 1} \exp(-y_2) \end{aligned}$$

from which we can obtain the marginals:

$$\begin{aligned} p(y_1) &= \int_0^\infty \frac{1}{\Gamma(a_1)\cdot\Gamma(a_2)} \cdot y_1^{q-1}(1-y_1)^{a_2-1} \cdot y_2^{a_1-1+a_2} \cdot \exp(-y_2)\, dy_2 \\ &= \frac{1}{\Gamma(a_1)\cdot\Gamma(a_2)} \cdot y_1^{a_1-1}(-1-y_1)^{a_2-1} \cdot \int_0^\infty \underbrace{y_2^{a_1+a_2-1}\cdot\exp(-y_2)}_{\text{kernel of } Gamma(a_1+a_2,1)}\ dy_2 \\ &= \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1-1}(1-y_1)^{a_2-1} \end{aligned}$$

and recalling that $y_1 \in (0,1)$, we have $y_1 \sim \text{Beta}(a_1, a_2)$ and

$$\begin{aligned} p(y_2) &= \int_0^\infty \frac{1}{\Gamma(a_1)\Gamma(a_2)} (y_2)^{a_2+a_1-1} \cdot \exp((-y_2)\, y_1^{a_1-1} \cdot (1-y_1)^{a_2-1}\, dy_1 \\ &= \left(\frac{1}{\Gamma(a_1)\Gamma(a_1)}\right) \cdot y_2^{a_1+a_2-1} \cdot \exp(-y_2) \int_0^\infty y_1^{a-1} \cdot (1-y_2)^{a_2-1}\, dy_1 \end{aligned}$$

and recalling that $y_2 \in (0,\infty)$, we have $y_2 \sim \text{Gamma}(a_1+a_2, 1)$. **Proposed method**: If I have a source of $_1$ and $_2$ gamma, I can simulate gamma and then take the transformation $x_1 + x_2$ to obtain betas.

(C)

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

$$\begin{aligned} p(\theta \mid x_1 \dots x_N) &\propto \prod_1^n \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\theta)^2\right\} \cdot \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{v}}\exp\left\{-\frac{1}{2v}(\theta-m)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum(xi-\theta)^2 - \frac{1}{2v}(\theta-m)^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum x_i^2 - \sum 2xi\theta + N\theta^2\right) - \frac{1}{2v}\left(\theta^2 - 2\theta m + m^2\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(N\theta^2 - \sum^N 2x_i\theta\right) - \frac{1}{2v}\left(\theta^2 - 2\theta m\right)\right] \\ &= \exp\left\{-\frac{1}{2}\left(\frac{N\theta^2}{\sigma^2} + \frac{1}{v}\theta^2\right) + \frac{1}{2}\left(\frac{1}{\sigma^2}\sum^N 2x_i\theta + \frac{2\theta m}{v}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\theta^2\left(\frac{N}{\sigma^2}+\frac{1}{v}\right) + \theta\left(\frac{N\bar{x}}{\sigma^2}+\frac{m}{v}\right)\right\} \end{aligned}$$

from which, making use of well-known results for the normal distribution, we can say that

$$\theta|x_1,...,x_n \sim N(\frac{vN}{vN+\sigma^2}\bar{x} + \frac{\sigma^2}{vN+\sigma^2}m; \frac{\sigma^2 v}{vN+\sigma^2})$$

(D)

Now assume known mean θ but unknown variance σ^2. Also, re-express things in terms of the precision, or inverse variance$w = 1/\sigma^2$. Moreover, Suppose that w has a gamma prior with parameters a and b, implying that σ^2 has an inverse-gamma prior. Derive the posterior for w and σ^2.
We have

$$p(xi \mid \rho, \omega) = \frac{\sqrt{\omega}}{\sqrt{2\pi}} \exp\left\{-\frac{\omega}{2}(xi-\theta)^2\right\}$$

and

$$p(w) \propto (w)^{a-1} exp(-wb)$$

so that

$$\begin{aligned} p\left(\omega \mid x_1 \ldots x_N\right) &\propto \left(\prod_{i=1}^{N} \frac{\sqrt{w}}{\sqrt{2\pi}} \exp\left\{-\frac{w}{2}(xi-\theta)^2\right\}\right) \cdot w^{a-1} \cdot \exp(-wb) \\ &\propto (\sqrt{w})^N \cdot w^{a-1} \cdot \exp(-wb) \cdot \exp\left\{-\frac{w}{2}\sum (x_i-\theta)^2\right\} \\ &= w^{a-1+\frac{N}{2}} \exp\left(-w\left(b + \frac{(x_i-\theta)^2}{2}\right)\right\} \end{aligned}$$

which is the kernel of a gamma, so that:

$$w \sim \text{ga}\left(a+\frac{1}{2}, b + \frac{\sum(xi-\theta)^2}{2}\right) \qquad \sigma^2 \sim Inv-ga\left(a+\frac{1}{2}, b + \frac{\sum(xi-\theta)^2}{2}\right)$$

(E)

$$\begin{aligned} p\left(\theta \mid x_1 \ldots x_N\right) &\propto \prod^{N} \frac{1}{\sqrt{2\pi}}\frac{1}{\delta_i} \exp\left\{-\frac{1}{2\delta_i}\left(x_i-\theta\right)^2\right\} \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{v}} \exp\left\{-\frac{1}{2v}(\theta-m)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\sum \frac{\left(x_i-\theta\right)^2}{\delta_i^2}\right\} \cdot \exp\left\{-\frac{1}{2v}(\theta-m)^2\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\sum^{N} \frac{(x_i-\theta)^2}{\delta i^2}\right) - \frac{1}{2}\frac{(\theta-m)^2}{v}\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\sum \frac{x^2-2x_i\theta+\theta^2}{\delta_i^2}\right) - \frac{1}{2}\left(\frac{\theta^2-2\theta m+m^2}{v}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\theta^2\left(\sum \frac{1}{\delta_i^2}+\frac{1}{v}\right) + \frac{1}{2}\theta\left(2\sum^{N}\frac{x_i}{\delta_i^2}+\frac{2m}{v}\right)\right\} \end{aligned}$$

$\Rightarrow \theta|x_1, ..., x_N \sim N(M, V)$ with

$$M = \frac{\sum(\frac{1}{\sigma_i^2}x_i) + \frac{1}{v}m}{\sum \frac{1}{\sigma^2} + \frac{1}{v}}$$

which is a weighted averge of the observations and the prior mean.

$$V = (\sum \frac{1}{\sigma_i^2} + \frac{m}{v})^{-1}$$

(F)

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

$\Rightarrow x$ follows a Student's t distribution with a degrees of freedom, center m, and scale parameter $(b/a)^{1/2}$

The multivariate normal distribution

(A)

$$\begin{aligned} cov(x) &= E\{(x-\mu)(x-\mu)^T\} \\ &= E\{(x-\mu)(x^T-\mu^T)\} \\ &= E\{xx^T - x\mu^T - \mu x^T + \mu\mu^T\} \\ &= E\{xx^T\} - E\{x\}\mu^T - \mu E\{x^T\} + \mu\mu^T\} \\ &= E\{xx^T\} - 2\mu\mu^T + \mu\mu^T \\ &= E\{xx^T\} - \mu\mu^T \end{aligned}$$

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

(B)

Derive the PDF of $z = (z_1, ..., z_p)$

$$\begin{aligned} p(z) &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_i^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^{p/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{p} z_i^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^{p/2} \exp\left(-\frac{1}{2}z^Tz\right) \end{aligned}$$

so that

$$\begin{aligned} M_z(t) &= \prod_{i=1}^{p} M_{zi}(t_i) \\ &= \prod_{i=1}^{p} \exp\left(\frac{1}{2}t_i^2\right) \\ &= \exp\left(\frac{1}{2}\sum_{i=1}^{n} t_i^2\right) \\ &= \exp\left(\frac{1}{2}t^Tt\right) \end{aligned}$$

(C)

We have that for a univariate normal distribution

$$M_z = E\left\{\exp\left(mt + \frac{1}{2}vt^2\right)\right\}$$

so that
$m = E(z) = E(a^T x) = a^T \mu$ and
$v = var(z) = var(a^T x) = a^T var(x) a = a^T \Sigma a$
Now plug m and v into the M_z equation

$$M_z = E\{\exp(ta^T x)\} = \exp\left(a^T \mu t + \frac{1}{2} a^T \Sigma a t^2\right)$$

let $t = 1$

$$M_z = E\{\exp(a^T x)\} = \exp\left(a^T \mu + \frac{1}{2} a^T \Sigma a\right)$$

If we replace a with t, this has the same form as in the question:

$$M_z = \exp(t^T x)\} = \exp\left(t^T \mu + \frac{1}{2} t^T \Sigma t\right)$$

(D)

Let $x = Lz + \mu$

$$\begin{aligned}
M_x &= E\{\exp(t^T z)\} \\
&= E\{\exp(t^T (Lz + \mu))\} \\
&= E\{\exp(t^T Lz \exp\{t^T \mu)\} \\
&= \exp\{t^T \mu) E\{\exp(t^T Lz\} \\
&= \exp(t^T \mu) E\{\exp[(L^T t)^T z]\} \\
&= \exp\left\{t^T \mu + \frac{(L^T t)^T (L^T t)}{2}\right\} \\
&= \exp\left\{t^T \mu + \frac{t^T L L^T t}{2}\right\}
\end{aligned}$$

This has the form of a multivariate normal MGF with mean μ and variance LL^T.

(E)

We know the MFG for z, and $x = Lz + \mu$. Hence:

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

and writing Σ as LL^T

$$\begin{aligned} &= E\{\exp(t^T z)\} = \exp\left(-t^T L^{-1}\mu\right) \exp\left[(L^{-T}t)^T\mu + \frac{1}{2}\left(L^{-T}t\right)^T LL^T \left(L^{-T}t\right)\right] \\ &= \exp\left(-t^T L^{-1}\mu\right) \exp\left[(L^{-T}t)^T\mu + \frac{1}{2}\left(t^T L^{-1} LL^T L^{-T}t\right)\right] \\ &= \exp\left(-t^T L^{-1}\mu\right) \exp\left[(L^{-T}t)^T\mu + \frac{1}{2}\left(t^T L^{-1} LL^T L^{-T}t\right)\right] \\ &= \exp\left(-t^T L^{-1}\mu\right) \exp\left[(L^{-T}t)^T\mu + \frac{1}{2}\left(t^T t\right)\right] \\ &= \exp\left(\frac{t^T t}{2}\right) \end{aligned}$$

Which is the MGF for a standard multivariate normal distribution.

Proposed strategy to simulate a multivariate normal distribution with known covariance matrix and means:

1. Generate n standard univariate distributions
2. Generate vector of length n containing means
3. Generate covariance matrix of size $n \times n$
4. Get target distribution using $x = Lz + \mu$

(F)

From above, we know that

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

with $L = P\Lambda^{1/2}$ and $L^{-1} = \left(P\Lambda^{1/2}\right)^{-1} = \Lambda^{-1/2}P^T$. Hence, $z = \Lambda^{-1/2}P^T(x- \mu)$. We can write:

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

since $|\Sigma| = \left|P\Lambda P^T\right| = |P||\Lambda||P| = |\Lambda|$.
So that

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

(G)

We have $x_1 \sim N(\mu_1, \Sigma_1)$ and $x_2 \sim N(\mu_2, \Sigma_2)$ with $y = Ax_1 + Bx_2$

$$M_{Ax_1}(t) = \exp\left[t^TA\mu_1 + \frac{1}{2}t^TA\Sigma_1A^Tt\right]$$

so that, from above, we have:

$$M_{Bx_2}(t) = \exp\left[t^TB\mu_2 + \frac{1}{2}t^TB\Sigma_2B^Tt\right]$$

Now we can get the mgf of y, taking advantage of independence: By independence,

$$\begin{aligned} M_{Ax_1+Bx_2}(t) &= M_{Ax_1}(t)\cdot M_{Bx_2}(t) \\ &= \exp\left[t^T\left(A\mu_1 + B\mu_2\right) + \frac{1}{2}t^T\left(A\Sigma_1A^T + B\Sigma_2B^T\right)t\right] \end{aligned}$$

from bove, we know that this has the form of a multivariate normal mgf. In oter words,

$$y = Ax_1 + Bx_2 \sim \mathrm{N}\left(A\mu_1 + B\mu_2, A\Sigma_1A^T + B\Sigma_2B^T\right)$$

Multiple regression: three classical principles for inference

A

From the previous question, for $y = Ax_1 + Bx_2$

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

Here let $x_1 = Ax$, where $A = [I_{k\times k}|0_{k\times(p-k)}]$, so x_1 has is multivariate normal with mean $A\mu$ and covariance $A\Sigma A^T$

We have:

$$E(x_1) = A\mu = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mu_1$$

and

$$cov(x_1) = A\Sigma A^T = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{11} \\ \Sigma_{21} \end{bmatrix} = \Sigma_{11}$$

so that he marginal distribution of x_1 is a multivariate normal with parameters μ_1 and $cov(x_1)$.

(B)

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

(C)

We know

$$f(x_1, x_2) \propto \exp\left(-\frac{1}{2}(x-\mu)^T\Omega(x-\mu)\right)$$

Focusing on $(x-\mu)^T\Omega(x-\mu)$, we have:

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

Now, we only consider terms containing x_1. Thus,

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

now we can use the complete the square trick and see that:

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

Multiple regression: three classical principles for inference

(A)

- Least squares

Find β such that $(y - X\beta)^T(y - X\beta)$ is minimized.

$$(y - X\hat{\beta})^T(y - X\hat{\beta})$$
$$= y^Ty - y^TX\hat{\beta} - \hat{\beta}^TX^Ty + \hat{\beta}^TX^TX\hat{\beta}$$
$$= y^Ty - (X\hat{\beta})^Ty - (X\hat{\beta})^Ty + \hat{\beta}^TX^TX\hat{\beta}$$
$$= y^Ty - 2(X\hat{\beta})^Ty + \hat{\beta}^TX^TX\hat{\beta}$$

Let

$$\frac{\partial}{\partial\hat{\beta}}[y^Ty - 2\hat{\beta}^TX^Ty + \hat{\beta}^TX^TX\hat{\beta}] = 0$$
$$-2y^TX + 2X^TX\hat{\beta} = 0$$
$$2X^TX\hat{\beta} = 2y^TX$$
$$\hat{\beta} = (X^TX)^{-1}y^TX$$

- Maximum Likelihood

We assume $\epsilon_i \sim N(0, \sigma^2)$

$$P(y_i|\hat{\beta}, \sigma^2) \propto \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i^T\hat{\beta})^2\right\}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2\right\}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}(y - X\hat{\beta})^T(y - X\hat{\beta})\right\}$$

We can take the log and maximize

$$\frac{\partial}{\partial\hat{\beta}}[y^Ty - 2\hat{\beta}^TX^Ty + \hat{\beta}^TX^TX\hat{\beta}] = 0$$

We end up with the same thing as before.

- Moment matching

Find $\hat{\beta}$ so that $cov(\epsilon, x_j) = 0$

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})\epsilon_i = 0$$

$$X^T\epsilon = 0$$

Assuming $\bar{x} = 0$.

Then,

$$X^T\epsilon = 0$$
$$X^T(y - X\hat{\beta}) = 0$$
$$X^Ty - X^TX\hat{\beta} = 0$$
$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

Note

$$y \sim N(X\beta, \sigma^2 I)$$

$$E(\hat{\beta}) = (X^TX)^{-1}X^TE(y) = (X^TX)^{-1}X^TX\beta = \beta$$

$$\begin{aligned} cov(\hat{\beta}) &= (X^TX)^{-1}X^Tcov(y)[(X^TX)^{-1}X^T]^T \\ &= (X^TX)^{-1}X^T(\sigma^2 I)[(X^TX)^{-1}X^T]^T \\ &= (X^TX)^{-1}X^T(\sigma^2 I)X(X^TX)^{-T} \\ &= (X^TX)^{-1}X^T(\sigma^2 I)X(X^TX)^{-1} \\ &= \sigma^2[(X^TX)^{-1}X^TX(X^TX)^{-1}] \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

(B)

We want to maximize

$$-\frac{1}{2}(y - X\hat{\beta})^T\Sigma^{-1}(y - X\hat{\beta})$$

$$\propto \frac{\partial}{\partial\hat{\beta}}[-y^T\Sigma^{-1}y + \hat{\beta}^TX^T\Sigma^{-1}y + y^T\Sigma^{-1}X\hat{\beta} - \hat{\beta}^TX^T\sigma^{-1}X\hat{\beta}] = 0$$

$$2y^T\Sigma^{-1}X - 2X^T\Sigma^{-1}X\hat{\beta} = 0$$

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

As before,

$$E(\hat{\beta}) = \beta$$

$$\begin{aligned} cov(\hat{\beta}) &= [(X^T\Sigma^{-1}X)^{-1}\Sigma^{-1}]^{-1}X^T cov(y)[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]^T \\ &= [(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]\Sigma[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]^T \\ &= (X^T\Sigma^{-1}X)^{-1} \end{aligned}$$

(C)

We want to maximize

$$-\frac{1}{2}(y - X\hat{\beta})^T\Sigma^{-1}(y - X\hat{\beta})$$

Where W is a diagonal precisions matrix with entries $w_i = \frac{1}{\sigma_i^2}$

We get

$$\hat{\beta} = (X^TWX)^{-1}X^TWy$$

Why Shouldn't I Invert That Matrix?

The results we have outlined so far bring us to obtain β which solves the equation:

$$(X'\Sigma X)\beta = X'y$$

If we denote $(X'\Sigma X) = A$, $X'y = b$, $\beta = X$ this problem boils down to the more general problem of solving for $\mathbf{x}$ in

$$\mathbf{Ax} = \mathbf{b}$$

where $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{x}$ and $\mathbf{b}$ are n vectors. One way to solve this equation is a matrix inversion $\mathbf{A}^{-1}$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

However, we can in practice avoid the action of computing$\mathbf{A}^{-1}$. Indeed, we can solve the system of linear equations directly. As we will see, this approach could actually outperform the former.

LU method

In order to say if inverting the matrix is the best method we could find, we need to first outline an alternative method and then compare their performances. One possible method is to use the LU decomposition of A. Denote by LU the LU decomposition of A such that LU=A with U being an upper triangular matrix. Then one can perform the following steps:

1. Perform LU decomposition to get $\mathbf{LU}$ such that $\mathbf{LU} = \mathbf{A}$
2. Solve the linear system $\mathbf{L}y = b$ and get y
3. Take y and solve the linear system $\mathbf{U}x = y$

Before comparing note that...

matrix inversion involves LU decomposition! Indeed, note that computing $\mathbf{A}^{-1}$ consists of finding $\mathbf{A}^{-1}$ such that:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = [\mathbf{e}_1, ..., \mathbf{e}_n]$$

this means that, denoting by x_i each column of $\mathbf{A}^{-1}$ we will be solving for x_i in

$$\mathbf{A}x_i = \mathbf{e}_i$$

which will be worked out by using

$$\mathbf{LU}x_i = \mathbf{e}_i$$

Comparison

Computational complexity

Why should this approach be more convenient than the one simply inverting the matrix A? The first answer we can give boils down to the computational complexity of the steps involving the use of the LU decomposition.

Recall that obtaining the LU decomposition consists in transforming **A** into an upper-triangular matrix **U** by repeatedly introducing zeros below the diagonal: first add zeros to the first column, then the second, and so on. Hence, on the $i-th$ transformation, the algorithm introduces $n-i$ zeros below the diagonal by subtracting multiples of the $(i-1)-th$ row from the rows beneath it $(i+1,\dots,n)$...so that one can show that for each of the $n-1$ iterations of the algorithm, we have complexity that is $(n-i)\times(ni)$ and that the work required for LU decomposition is $\sim \frac{2}{3}n^3$ flops (floating-point operations). Moreover, once the LU decomposition has been performed, one still need to solve for the second and third step. One can show that the last two steps we indicated will imply that additional $2n^2$ "flops"will be involved. Summing up:

1. Perform LU factorization $\rightarrow \frac{2}{3}n^3$ "flops"
2. solve for y and then for $x \rightarrow$ additional $2n^2$ flops for each column $\rightarrow$ additional $2n^3$ in total.

Now consider the steps we outlined for finding the inverse. We will:

1. Carry on the LU factorization $\rightarrow \frac{2}{3}n^3$ "flops"
2. solve for each $x_i \rightarrow$ additional $2n^2$ flops for each column $\rightarrow$ additional $2n^3$ in total.

Accuracy

Inversion is not more stable than factorization. Indeed, think about the case of $X'WX$ being a ill-conditioned matrix. Intuitively, this means that when applying that matrix to similar vectors but you could get solutions that are significantly different in the output. More rigorously, a ill-conditioned matrix is such that the smallest and biggest eigenvalues are considerably different. In this case, tiny changes in y would produce bug changes is β, this undermining the accuracy of matrix inversion.

Proposed method

The LU decomposition method for solving for $\hat{\beta}$ in $X'WX\hat{\beta} = X'Wy$ follows the following steps: ("My method"):

1. Factor $X'WX$ as LU using LU decomposition, where L and U are lower and upper triangular matrices.

2. solve for z in $LZ = X'Wy$ using forward substitution

3. Solve for β in $U\beta = z$ using backward substitution

Simulation

I implemented the LU decomposition method using R (see appendix for the code). Results are summarised in the following table: (values are given in microseconds). Each method was implemented 10 times, te means are reported.

N	**P**	Inversion method	Factorization method
10	2	189.63	69.35
100	50	299.74168	13.55321
500	100	4679.965	161.231height

We can conclude that the LU method outperforms the inversion method under different scenarios. Unfortunately, I wasn't able to obtain results for higher values of N and P.

Appendix

```
#Function that inverts the matrix
invert_sol <- function(A, b, S){
  beta_hat <- inv(t(A)%*%S%*%A)%*%t(A)%*%S%*%b
  return(beta_hat)
}

#Function that uses LU decomposition
lin_sistem_sol <- function(A, b, S){
  to_invert <- t(A)%*%S%*%A
  lu_decomp = lu.decomposition(to_invert)
  beta_hat = solve(lu_decomp$U, solve(lu_decomp$L, t(A)%*%b))
  return(beta_hat)
}

#repeat for a range of N and P:
N <- c(10, 100, 500)
P <- c(2, 50, 100)
#evaluate over the different pairs:
for(i in 1:length(N)){
  ## create random dataset
  n <- N[i]
  p <- P[i]
  W <- diag(n) # identity matrix for W for
  X <- matrix(rnorm(n*p), n, p) # design matrix
  y <- rnorm(n, 0.3*X[,1]+0.5*X[,2], 1)
```

```
  ## Implementation
  assign(paste0("benchmark",i),microbenchmark(
  invert_sol(X, y, W), lin_sistem_sol(X, y, W), times=10))
}
```