Statistical Modeling SDS 383D: Excercise 1

Khai Nguyen - kbn397
April 24, 2022

Bayesian inference in simple conjugate families :
(A) We have

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

which is the Beta kernel of Beta$(a+\sum_{i=1}^N x_i, b+\sum_{i=1}^N(1-x_i))$.
(B) We have $y_2 = x_1+x_2$ and $y_1 = \frac{x_1}{x_1+x_2}$, hence $x_1 = y_1y_2$ and $x_2 = y_2 - y_1y_2$. We have the determination of the Jacobian is

$$|J| = \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1-y_1 \end{vmatrix} = y_2$$

Therefore,

$$\begin{aligned} p(y_1,y_2) &= \frac{1}{\Gamma(a_1)}(y_1y_2)^{a_1-1}\exp(y_1y_2)\frac{1}{\Gamma(a_2)}(y_2-y_1y_2)^{a_2-1}\exp(y_2-y_1y_2)y_2 \\ &= \frac{1}{\Gamma(a_1)\Gamma(a_2)}(y_1y_2)^{a_1-1}(y_2-y_1y_2)^{a_2-1}y_2\exp(y_2) \\ &= \frac{1}{\Gamma(a_1)\Gamma(a_2)}y_2^{a_1+a_2-1}\exp(y_2)y_1^{a_1-1}(1-y_1)^{a_2-1} \end{aligned}$$

By the factorization theorem, y_1 and y_2 is independent, and their pdfs are

$$\begin{aligned} p(y_1) &= \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)}y_1^{a_1-1}(1-y_1)^{a_2-1} := Beta(a_1,a_2) \\ p(y_2) &= \frac{1}{\Gamma(a_1+a_2)}y_2^{a_1+a_2-1}\exp(y_2) := Ga(a_1+a_2,1) \end{aligned}$$

So, we can create a Beta(a_1,a_2) (y_1) random variable by sampling two random variables from Ga$(a_1,1)$ (x_1), Ga$(a_2,1)$ (x_2), then calculate $\frac{x_1}{x_1+x_2}$.
(C) We have

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

which is the kernel of $Gausian\left((\frac{1}{v}+\frac{n}{\sigma^2})^{-1}(\frac{m}{v}+\frac{\sum_{i=1}^N x_i}{\sigma^2}),(\frac{1}{v}+\frac{n}{\sigma^2})^{-1}\right)$.

(D) We have:

$$
\begin{aligned}
p(w|x_1,\dots,x_N) &\propto p(w)p(x_1,\dots,x_N|w) \\
&\propto w^{a-1}\exp(-bw)\prod_{i=1}^N w^{1/2}\exp\left(-\frac{w}{2}(x_i-\theta)^2\right) \\
&= w^{a+n/2-1}\exp\left(-(b+\sum_{i=1}^N(x_i-\theta)^2/2)w\right)
\end{aligned}
$$

Which is the kernel of Gamma$(a+n/2, b+\sum_{i=1}^N(x_i-\theta)^2/2)$.

For σ^2

$$
\begin{aligned}
p(\sigma^2|x_1,\dots,x_N) &\propto p(\sigma^2)p(x_1,\dots,x_N|\sigma^2) \\
&\propto (1/\sigma^2)^{a+1}\exp(-b/\sigma^2)\prod_{i=1}^N\frac{1}{\sigma^2}\exp\left(-\frac{1}{2\sigma^2}(x_i-\theta)^2\right) \\
&= (1/\sigma^2)^{a+n/2-1}\exp\left(-\frac{1}{\sigma^2}(b+\frac{1}{2}\sum_{i=1}^N(x_i-\theta)^2)\right)
\end{aligned}
$$

Which is the kernel of Inv-Gamma$(a+n/2, b+\sum_{i=1}^N(x_i-\theta)^2/2)$.

(E) We have

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

The posterior mean is $(\frac{1}{v}+\sum_{i=1}^n\frac{1}{\sigma_i^2})^{-1}(\frac{m}{v}+\frac{\sum_{i=1}^N x_i}{\sigma^2})$

(F) We have

$$
\begin{aligned}
p(x,w) &= p(w)p(x|w) \\
&= \frac{(b/2)^{a/2}}{\Gamma(a/2)}w^{a/2-1}\exp(-b/2w)\left(\frac{w}{2\pi}\right)^{1/2}\exp\left\{-\frac{w}{2}(x-m)^2\right\} \\
&= \frac{(b/2)^{a/2}}{\Gamma(a/2)(2\pi)^{1/2}}w^{a/2+1/2-1}\exp\left\{-\frac{w}{2}(b+(x-m)^2)\right\}
\end{aligned}
$$

Hence

$$
\begin{aligned}
p(x) &= \int_{-\infty}^{\infty} p(x,w)dw \\
&= \int_{-\infty}^{\infty} \frac{(b/2)^{a/2}}{\Gamma(a/2)(2\pi)^{1/2}} w^{a/2+1/2-1} \exp\left\{-\frac{w}{2}(b+(x-m)^2)\right\} dw \\
&= \int_{-\infty}^{\infty} \frac{(b/2)^{a/2}\Gamma(a/2+1/2)}{\Gamma(a/2)(2\pi)^{1/2}((b+(x-m)^2)/2)^{a/2+1/2}} \\
&\quad \frac{((b+(x-m)^2)/2)^{a/2+1/2}}{\Gamma(a/2+1/2)} w^{a/2+1/2-1} \exp\left\{-\frac{w}{2}(b+(x-m)^2)\right\} dw \\
&= \frac{(b/2)^{a/2}\Gamma(a/2+1/2)}{\Gamma(a/2)(2\pi)^{1/2}((b+(x-m)^2)/2)^{a/2+1/2}}
\end{aligned}
$$

The multivariate normal distribution :

(A) We have

$$
\begin{aligned}
Cov(x) &= E[(x-\mu)(x-\mu)^\top] = E[xx^\top - x\mu^\top - \mu x^\top + \mu\mu^\top] = E[xx^\top] - \mu\mu^\top \\
Cov(Ax+b) &= E[(Ax+b-E[Ax+b])(Ax+b-E[Ax+b])^\top] \\
&= AE[(x-\mu)(x-\mu)^\top]A^\top = ACov(x)A^\top
\end{aligned}
$$

(B) We have the pdf:

$$
p(z) = \prod_{i=1}^{p} p(z_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) = \frac{1}{\sqrt{(2\pi)^p}} \exp\left(-\frac{1}{2} z^\top z\right)
$$

we have the mgf:

$$
M_z(t) = E[e^{t^\top z}] = \prod_{i=1}^{p} E[e^{t_i z_i}] = \prod_{i=1}^{p} \exp\left(t_i^2/2\right) = \exp\left(t^\top t/2\right)
$$

(C)

1. If $x \sim N(\mu, \Sigma)$, we have $a^\top x$ is a Gaussian. We have $E[z] = a^\top \mu$, $Cov[z] = a\Sigma a^\top$, so $E[\exp(tz)] = \exp(a^\top \mu t + a\Sigma a^\top t^2/2) = \exp(\mu^\top at + (at)\top \Sigma at/2)$. Let $at = y$, we have

$$
M(y) = \exp(\mu^\top y + y^\top \Sigma y/2)
$$

2. If $M_x(t) = \exp(\mu^\top t + t^\top \Sigma t/2)$, then

$$
M_z(t) = \exp(\mu^\top at + (at)\top \Sigma at/2) \tag{1}
$$

(D) We have

$$
\begin{aligned}
E[\exp(t^\top x)] &= E[\exp(t^\top(Lz+\mu))] = \exp(t^\top \mu)E[\exp(t^\top Lz)] \\
&= \exp(t^\top \mu)\exp(t^\top LL^\top t/2) = \exp(t^\top \mu + t^\top \Sigma t/2)
\end{aligned}
$$

Now we calculate the mean and the variance

$$\frac{\partial \exp(t^\top \mu + t^\top \Sigma t/2)}{\partial t}(0) = \mu$$
$$\frac{\partial^2 \exp(t^\top \mu + t^\top \Sigma t/2)}{\partial t^2}(0) = \Sigma$$

(E) Using spectral decomposition: $\Sigma = P\Lambda P^\top$. Let $L = P(\Lambda)^{1/2}$, using previous problem we have $Lz + \mu \sim N(\mu, \Sigma)$.
(F) $x = Lz + \mu$, hence $Z = L^{-1}(x - \mu)$. Using Multivariate transformation with the Jacobian matrix is L^{-1}.

$$p(x) = \frac{1}{\sqrt{(2\pi)^p}} \exp(-\frac{1}{2}(L^{-1}(x-\mu))^\top (L^{-1}(x-\mu)))) \det(L^{-1})$$

(G) Using mgf:

$$\begin{aligned} E[\exp(t^\top y)] &= E[\exp^{t^\top (Ax_1 + Bx_2)}] = E[\exp(t^\top Ax_1)] + E[\exp(t^\top Bx_2)] \\ &= \exp(t^\top A\mu_1 + t^\top A\Sigma_1 A^\top t) + \exp(t^\top B\mu_2 + t^\top B\Sigma_2 B^\top t) \\ &= \exp(t^\top (A\mu_1 + B\mu_2) + t^\top (A\Sigma_1 A^\top + B\Sigma_2 B^\top) t) \end{aligned}$$

Conditional and Margianls :
(A) Let $a = (1, 0)'$, hence $x_1 = a^\top x$ has the distribution $N(\mu_1, \Sigma_{11})$
(B) We have

$$\Sigma^{-1} = \begin{pmatrix} \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & -\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{pmatrix}$$

(C) we have

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

We have $p(x_2) = \frac{1}{\sqrt{(2\pi)^k|\Sigma_{22}|}} \exp\left(-\frac{1}{2}(x_2 - \mu_{22})^\top \Sigma_{22}^{-1}(x_2 - \mu_{22})\right)$, Hence

$$p(x_1|x_2) \propto \exp\left(-\frac{1}{2}\left[(x-\mu)^\top \Sigma^{-1}(x-\mu) - (x_2 - \mu_{22})^\top \Sigma_{22}^{-1}(x_2 - \mu_{22})\right]\right)$$

We now calculate

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

So $p(x_1|x_2) = N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.
We can interpret $x_1 = (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)) + (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{1/2}\epsilon$

Multiple Regression: Three classical principles for inference :

(A) For LSE, we have

$$\nabla_\beta (y - X\beta)^\top (y - X\beta) = 2X^\top (X\beta - y)$$

Set the gradient to 0, we have

$$\beta^* = (X^\top X)^{-1} X^\top y$$

For MLE, using log function which is monotonically increasing

$$\begin{aligned}\log \left(\prod_{i=1}^{n} p(y_i|\beta, \sigma^2) \right) &= \sum_{i=1}^{n} \log p(y_i|\beta, \sigma^2) \\ &= -\frac{1}{2} \sum_{i=1}^{n} ||y - X\beta||_2^2 + constant\end{aligned}$$

So maximizing the log-likelihood is equivalent to the LSE.
For method of moments, we have for all column index j

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

Suppose data is centered, hence

$$\begin{aligned}&\sum_{i=1}^{n} e_i x_{ij} = 0 \\ &e^T X be = 0 \\ &(Y - X\beta)^T X = 0\end{aligned}$$

Then

$$\beta = (X^\top X)^{-1} X^\top Y$$

(B) In this case, the problem is equivalent to

$$\arg\min_\beta (y - X\beta) \top \Sigma^{-1} (y - X\beta)$$

Taking the gradient and set to 0, we have

$$2X^\top \Sigma^{-1} (X\beta - y) = 0$$

	N=100	N=500	N=1000	N=5000
d=5	9.17911530e-05	6.41345978e-05	5.38825989e-05	2.04086304e-04
d=10	4.79221344e-05	7.29560852e-05	1.09195709e-04	6.19173050e-04
d=100	1.80816650e-03	2.22802162e-03	4.44316864e-03	1.13751888e-02
d=1000	2.17528820e-01	3.80424976e-01	2.06042051e-01	3.43861818e-01

Table 1: Time of Inv Method

	N=100	N=500	N=1000	N=5000
d=5	1.97124481e-03	8.17775726e-05	6.91413879e-05	1.01089478e-04
d=10	2.15053558e-04	8.79764557e-05	1.66893005e-04	1.08957291e-04
d=100	2.93731689e-03	3.29875946e-03	2.96974182e-03	3.29399109e-03
d=1000	4.71949577e-02	4.47421074e-02	4.03697491e-02	3.71558666e-02

Table 2: Time of Dec Method

hence the estimator is

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

Now the variance of the estimator is:

$$Var[(X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y] = X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Var[y] (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1})^\top = (X^\top \Sigma^{-1} X)^{-1}$$

(C) $W = diag(1/\sigma_i^2)$

Some practical details :

(A) Matrix inverse is computationally expensive. The fastest algorithm has the complexity of $O(n^{2.373})$ with n is the size of the square matrix. Also, if the data matrix is not full column rank, we cannot inverse the matrix.

Using LU decomposition $O(n^{2376})$ to solve the equation. However, it seesm to be slower than using matrix inverse.

Pseudo code

1. Decomposing $X^\top W X$ as LU
2. Solve for z :$Lz = X \top W y$ using forward substitution
3. Solve for $\hat{\beta}$: $U\hat{\beta} = z$ using backward substitution

(B) I set $W = I$, $d \in \{2, 5, 10, 50\}$, $N \in \{100, 500, 1000, 5000\}$, the time of Inv method is given in Table 1: Time of decomposition method is given in Table 2.