HW3

PENG, Han 07/03/2024

1. Maximum Likelihood Method

(a)

$$p(X_i) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left[-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right]$$

$$p(\mathcal{D}|\mu,\Sigma) = [(2\pi)^p|\Sigma|]^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right]$$

$$\begin{aligned} l_n(\mu,\Sigma) = \log p(\mathcal{D}|\mu,\Sigma) &= -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu) \\ &= -\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu) - \frac{n}{2}\log|\Sigma| + C \end{aligned}$$

Since $l_n(\mu,\Sigma)$ is a constant, $l_n(\mu,\Sigma) = trace\big(l_n(\mu,\Sigma)\big)$, which leads to

$$l_n(\mu,\Sigma) = trace\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right] - trace\left[\frac{n}{2}\log|\Sigma|\right] + Trace(C)$$

Using the property of $Trace(ABC) = Trace(BCA)$

$$\begin{aligned} trace\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right] &= -\frac{1}{2}\sum_{i=1}^{n} trace[\Sigma^{-1}(X_i-\mu)(X_i-\mu)^T] \\ &= -\frac{n}{2}trace[\Sigma^{-1}S_n] \end{aligned}$$

Where $S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^T(X_i-\mu)$

Consequently,

$$l_n(\mu,\Sigma) = -\frac{n}{2}trace[\Sigma^{-1}S_n] - \frac{n}{2}\log|\Sigma| + C$$

(b)

Denote

$$(X+\Delta)^{-1} = X^{-1} + dX^{-1}$$

I have that

$$(X+\Delta)^{-1}(X+\Delta) = I = (X^{-1}+dX^{-1})(X+\Delta) = I$$

Notice $\Delta dX^{-1} \approx 0$, expand the above equation we have

$$\begin{aligned} &X^{-1}X + dX^{-1}X + X^{-1}\Delta \approx I \\ &\to dX^{-1}X + X^{-1}\Delta = 0 \\ &\to dX^{-1} = -X^{-1}\Delta X^{-1} \end{aligned}$$

Therefore

$$
\begin{aligned}
&trace(A(X+\Delta)^{-1}) = trace\big(A(X^{-1}+dX^{-1})\big)\\
&= \ trace\big(A(X^{-1}-X^{-1}\Delta X^{-1})\big)\\
&= \ trace(AX^{-1}) - trace(\ AX^{-1}\Delta X^{-1})\\
&= trace(AX^{-1}) - trace(\ X^{-1}AX^{-1}\Delta)
\end{aligned}
$$

Recall that if $dy = trace(AdX)$ then $\frac{dy}{dX} = A$. Then we know the derivative is

$$
\frac{df(X)}{dX} = -\ X^{-1}AX^{-1}
$$

(c)

Let $X + \Delta = Z$

$$
\begin{aligned}
&\log\det Z = \log\det(X+\Delta)\\
&= \log\det\left(X^{\frac{1}{2}}\left(I + X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}\right)X^{\frac{1}{2}}\right)\\
&= \log\det X + \log\det\left(I + X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}\right)\\
&= \log\det X + \sum_{i=1}^{n}\log(1+\lambda_i)
\end{aligned}
$$

where λ_i is the ith eigenvalue of $X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}$. Since ΔX is small, this indicates that λ_i is small. So that $\log(1+\lambda_i) \approx \lambda_i$.Then

$$
\begin{aligned}
&\log\det Z \approx \log\det X + \sum_{i=1}^{n}\lambda_i\\
&= \log\det X + trace\left(\left(X^{-\frac{1}{2}}\Delta X\right)\ X^{-\frac{1}{2}}\right)\\
&= \log\det X + trace(\ X^{-1}\Delta X)\\
&= \log\det X + trace(\ X^{-1}(Z-X))
\end{aligned}
$$

Therefore

$$
g(X+\Delta) \approx g(X) + trace(X^{-1}\Delta)
$$

Recall that if $dy = trace(AdX)$ then $\frac{dy}{dX} = A$. Then we know

$$
\frac{dg(X)}{dX} = X^{-1}
$$

(d)

$$
l_n(\mu,\Sigma) = -\frac{n}{2}trace[\Sigma^{-1}S_n] - \frac{n}{2}\log\det\Sigma + C
$$

From (b) and (c) we know that

$$
\frac{dtrace[\Sigma^{-1}S_n]}{d\Sigma} = \frac{dtrace[S_n\Sigma^{-1}]}{d\Sigma} = -\Sigma^{-1}S_n\Sigma^{-1}
$$

$$
\frac{d\log\det\Sigma}{d\Sigma} = \Sigma^{-1}
$$

Consequently

$$\frac{dl_n(\mu,\Sigma)}{d\Sigma} = \frac{n}{2}\Sigma^{-1}S_n\Sigma^{-1} - \frac{n}{2}\Sigma^{-1} = 0 \rightarrow \hat{\Sigma}_n^{MLE} = S_n$$

2. Shrinkage

(a) Ridge regression

Let

$$J = \frac{1}{2}\|y-\mu\|_2^2 + \frac{\lambda}{2}\|\mu\|_2^2$$

We can write J in the vector form as

$$J = \frac{1}{2}(\mu-y)^T(\mu-y) + \frac{\lambda}{2}\mu^T\mu$$

By matrix calculus we have

$$\frac{\partial J}{\partial \mu} = (\mu-y) + \lambda\mu = 0$$
$$\rightarrow \mu = \frac{1}{1+\lambda}y$$

In element form we have

$$\hat{\mu}_i^{ridge} = \frac{1}{1+\lambda}y_i$$

We can use bias-variance decomposition to estimate the risk.

Consider $y \sim \mathcal{N}(\mu, \sigma^2 I_p)$

$$Var(\hat{\mu}^{ridge}) = \mathbb{E}\left[\left(\frac{1}{1+\lambda}y - \frac{1}{1+\lambda}\mu\right)^T\left(\frac{1}{1+\lambda}y - \frac{1}{1+\lambda}\mu\right)\right]$$
$$= \left(\frac{1}{1+\lambda}\right)^2 \mathbb{E}[y^Ty - 2\mu^Ty + \mu^T\mu] = \left(\frac{1}{1+\lambda}\right)^2 \sum_{i=1}^{p} \mathbb{E}[y_i^2 - \mu_i^2] = p\left(\frac{1}{1+\lambda}\right)^2 \sigma^2$$

$$Bias(\hat{\mu}^{ridge}) = \mathbb{E}\left[\left(\frac{1}{1+\lambda}\mu - \mu\right)^T\left(\frac{1}{1+\lambda}\mu - \mu\right)\right]$$
$$= \mathbb{E}\left[\left(\frac{\lambda}{1+\lambda}\right)^2 \mu^T\mu\right] = \left(\frac{\lambda}{1+\lambda}\right)^2 \sum_{i=1}^{p}\mu_i^2$$

Therefore, the risk of the estimator is

$$\mathbb{E}\left\|\hat{\mu}^{ridge} - \mu\right\| = p\left(\frac{1}{1+\lambda}\right)^2\sigma^2 + \left(\frac{\lambda}{1+\lambda}\right)^2\sum_{i=1}^{p}\mu_i^2$$

(b) LASSO problem

Let

$$J = \frac{1}{2}\|y-\mu\|_2^2 + \lambda\|\mu\|_1$$

We can write J in the vector form as

$$
\begin{aligned}
J &= \frac{1}{2}(\mu - y)^T(\mu - y) + \lambda|\mu|^T 1 \\
&= \frac{1}{2}\left(\sum_{i=1}^{p}\mu_i^2 - \sum_{i=1}^{p}2\mu_i y_i + \sum_{i=1}^{p}y_i^2\right) + \lambda\sum_{i=1}^{p}|\mu_i|
\end{aligned}
$$

The above equation can be minimized if each component J_i is minimized individually, which is

$$
J_i = \frac{1}{2}\ (\mu_i^2 - 2\mu_i y_i + y_i^2) + \lambda|\mu_i|
$$

It is obvious that μ_i and y_i should have the same sign for J_i to be minimized.

If $y_i \leq 0$, we take $\mu_i \leq 0$, and then

$$
J_i = \frac{1}{2}\ (\mu_i^2 - 2\mu_i y_i + y_i^2) - \lambda\mu_i
$$

$$
\begin{aligned}
&\frac{\partial J_i}{\partial \mu_i} = \mu_i - y_i - \lambda = 0 \\
&\rightarrow\ \mu_i = y_i + \lambda
\end{aligned}
$$

If $y_i \geq 0$, we take $\mu_i \geq 0$, and then

$$
J_i = \frac{1}{2}\ (\mu_i^2 - 2\mu_i y_i + y_i^2) + \lambda\mu_i
$$

$$
\begin{aligned}
&\frac{\partial J_i}{\partial \mu_i} = \mu_i - y_i + \lambda = 0 \\
&\rightarrow\ \mu_i = y_i - \lambda
\end{aligned}
$$

Together we have

$$
\mu_i = \begin{cases} y_i + \lambda & if\ y_i + \lambda < 0\ and\ y_i < 0 \\ y_i - \lambda & if\ y_i - \lambda > 0\ and\ y_i > 0 \\ 0 & otherwise \end{cases}
$$

The result can be written in a compact form

$$
\hat{\mu}_i^{soft} = sign(y_i)(|y_i| - \lambda)_+
$$

Using Stein's unbiased risk estimate, we have soft-thresholding in the form of

$$
\hat{\mu}(y_i) = y_i + g(y_i)
$$

$$
\frac{\partial g(y_i)}{\partial y_i} = -I(|y_i| \leq \lambda)
$$

$$
\begin{aligned}
&\mathbb{E}\left\|\hat{\mu}^{soft}(y) - \mu\right\|^2 = \mathbb{E}\left(p - 2\sum_{i=1}^{p} I(|y_i| \leq \lambda) + \sum_{i=1}^{p} y_i^2 \wedge \lambda^2\right) \\
&\leq 1 + (2\log p + 1)\sum_{i=1}^{p}\mu_i^2 \wedge 1
\end{aligned}
$$

if we take $\lambda = \sqrt{2\log p}$.