Homework 3

Li Donghao 20653877

Problem 1

1.1

$$
\begin{aligned}
l_n(\mu,\Sigma) &= \log \Pi_{i=1}^n \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)} \\
&= -\frac{1}{2}\sum_{i=1}^n (X_i-\mu)^T\Sigma^{-1}(X_i-\mu) - \frac{n}{2}\log\det(\Sigma) + C \\
&= -\frac{1}{2}\sum_{i=1}^n \operatorname{trace}\left[(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right] - \frac{n}{2}\log\det(\Sigma) + C \\
&= -\frac{1}{2}\sum_{i=1}^n \operatorname{trace}\left[\Sigma^{-1}(X_i-\mu)(X_i-\mu)^T\right] - \frac{n}{2}\log\det(\Sigma) + C \\
&= -\frac{n}{2}\operatorname{trace}\left(\Sigma^{-1}S_n\right) - \frac{n}{2}\log\det(\Sigma) + C
\end{aligned}
$$

1.2

$$
\begin{aligned}
f(X+\Delta) &= \operatorname{tr}\left(A(X+\Delta)^{-1}\right) \\
&= \operatorname{tr}\left(AX^{-1}\left(I+\Delta X^{-1}\right)^{-1}\right) \\
&\approx \operatorname{tr}\left(AX^{-1}\left(I-\Delta X^{-1}\right)\right) \\
&= \operatorname{tr}\left(AX^{-1}\right) - \operatorname{tr}\left(AX^{-1}\Delta X^{-1}\right) \\
&= f(X) - \operatorname{tr}\left(X^{-1}AX^{-1}\Delta\right)
\end{aligned}
$$

1.3

$$
\begin{aligned}
g(X+\Delta) &= \log\det(X+\Delta) \\
&= \log\det\left(X^{\frac{1}{2}}X^{\frac{1}{2}} + X^{\frac{1}{2}}X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}X^{\frac{1}{2}}\right) \\
&= \log\det\left(X^{\frac{1}{2}}\left(I + X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}\right)X^{\frac{1}{2}}\right) \\
&= \log\det\left(X^{\frac{1}{2}}Q(I+\Lambda)Q^{-1}X^{\frac{1}{2}}\right) \\
&= \log\det(X) + \log\det(I+\Lambda) \\
&\approx g(X) + \operatorname{tr}(\Lambda) \\
&= g(X) + \operatorname{tr}\left(X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}\right) \\
&= g(X) + \operatorname{tr}\left(X^{-1}\Delta\right)
\end{aligned}
$$

Where we define the eigenvalue decomposition of $X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}} = Q\Lambda Q^{-1}$

1.4

We need to take partial derivative w.r.t. Σ and set it to 0. So, we have :

$$
\frac{\partial l_n(\mu,\Sigma)}{\partial\Sigma} = \frac{n}{2}(\Sigma^{-1}S_n\Sigma^{-1} + \Sigma^{-1}\Sigma\Sigma^{-1}) = 0
$$

That is :

$$
\hat{\Sigma}^{MLE} = S_n
$$

2.1

Since we can split the objective function, we can just consider the $i-th$ term and do minimizaton. Take partial derivative w.r.t μ_i and set it to zero, we can have:

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

Then we can compute the risk:

$$
\begin{aligned}
R = E\|\mu - \hat{\mu}^{ridge}\|^2 &= \Sigma_{i=1}^p E\left(\mu_i - \frac{1}{1+\lambda}y_i\right)^2 \\
&= \frac{\lambda^2}{(1+\lambda)^2}||\mu||_2^2 + \frac{p}{(1+\lambda)^2}
\end{aligned}
$$

2.2

We just consider one dimension.

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

Then consider the derivative $\frac{df}{d\mu_i} = \mu_i - y_i + \lambda\frac{|\mu_i|}{\mu_i}$. When $y_i > lambda$ and $\mu_i > 0$, we have $\hat{\mu}_i^{soft} = y_i - \lambda$.

When $y_i < lambda$ and $\mu_i < 0$,we have $\hat{\mu}_i^{soft} = y_i + \lambda$.

Otherwise, we have $\hat{\mu}_i^{soft} = 0$. That finish the proof.

$$
E(\hat{\mu}_i^{soft} - \mu_i)^2 = \int_{-\inf}^{-\lambda}(x+\lambda-\mu)^2 f(x)dx + \int_{-\lambda}^{\lambda}(\mu)^2 f(x)dx + \int_{\lambda}^{\inf}(x-\lambda-\mu)^2 f(x)dx
$$

And I don't know how to do next.

2.3

We just consider one dimension:

$$
\min_{mu_i}(y_i-\mu_i)^2 + \lambda^2 1(\mu_i \neq 0)
$$

So the minimal value should be either λ^2 or y_i^2. That means when $\lambda > y_i$,$\mu_i = 0$,when $\lambda < y_i$,$\mu_i = y_i$.

$g(x) = 1$ when $y_i < \lambda$, $g(x) = 0$ when $y_i > \lambda$. Then g is weakly differentiable.

2.4

Using SURE with $g(Y) = -\frac{p-2}{||Y^2||}Y$ Then we have :

$$
\begin{aligned}
U(Y) &= p + 2\nabla^T g(Y) + ||g(Y)||^2 \\
R(\mu, \mu^{JS}) &= EU(Y) = p - E\frac{(p-2)^2}{||Y||^2}
\end{aligned}
$$

Comparing with MLE, when $p > 2$, the risk is smaller.

2.5

All the conditions are shrinkage rules.

3

In this problem, I use conter-examples to show we can improve the risk.

$$
r(\hat{\mu}_C, \mu) = \sigma^2\operatorname{tr}\left(C^TC\right) + \|(I-C)\mu\|^2
$$

(a): Let A=I-((I-C)^T(I-C))^(\frac{1}{2}) which is a symmetric matrix.

$$
\begin{aligned}
||(1-A)\mu||^2 &= \mu^T(1-C)^T(1-C)\mu = ||(1-A)\mu||^2 \\
trace(A^TA) &= n - 2trace(I-A) + trace(1-A)^T(1-A)
\end{aligned}
$$

By last equation, we have:

$$
\begin{aligned}
&trace(A^TA) = n - 2trace((I-C)^T(I-C))^{(}\frac{1}{2})) + trace(1-C)^T(1-C) \\
&\leq n - 2trace((I-C)^T(I-C)) + trace((1-C)^T(1-C)) = trace(C^TC)
\end{aligned}
$$

So the risk could be improved, so it is not admissible.

(b): Assume that the eigenvalues of C can be more than 1, $C = U\Lambda U^T$. Then we define $B = U\Lambda_{new}U^T$ where $\Lambda_{new_{ii}} = max(0, min(1, \Lambda_{ii}))$ Then

$$
\begin{aligned}
r(\hat{\mu}, \mu) &= \sigma^2 tr(C^TC) + ||(1-C)\mu||^2 \\
&= \sum_{i=1}^p(\sigma^2\rho_i^2 + (1-\rho_i)^2\mu_i^2) \\
&> \sum_{i=1}^p(\sigma^2\,\Lambda_{new_{ii}}^2 + (1-\Lambda_{new_{ii}})^2\mu_i^2) \\
&= r(\mu_{new}, \mu)
\end{aligned}
$$

(c): If there are $d \geq 3$ eigenvalues are one, we can design a JS estimator to these dimensions. Since JS estimator has smaller risk than MLE when $d \geq 3$, the new risk can be smaller.

In [ ]:

In [ ]:

In [ ]: