

Stat 246 Homework I - Solution

1 [10 points]

The probability of error is

$$\begin{aligned}P(\text{error}) &= P(Y \neq \hat{Y}) \\&= P(Y = 1, \hat{Y} = 2) + P(Y = 2, \hat{Y} = 1) \\&= P(Y = 1)P(\hat{Y} = 2|Y = 1) + P(Y = 2)P(\hat{Y} = 1|Y = 2) \\&= P(Y = 1)P(X \leq \tau|Y = 1) + P(Y = 2)P(X > \tau|Y = 2) \\&= P(Y = 1) \int_{-\infty}^{\tau} f(x|Y = 1)dx + P(Y = 2) \int_{\tau}^{\infty} f(x|Y = 2)dx.\end{aligned}$$

Find critical points (necessary condition):

$$\begin{aligned}0 &= \frac{d}{d\tau}P(\text{error}) \\&= P(Y = 1)\frac{d}{d\tau} \int_{-\infty}^{\tau} f(x|Y = 1)dx + P(Y = 2)\frac{d}{d\tau} \int_{-\infty}^{\tau} f(x|Y = 2)dx \\&= P(Y = 1)f(\tau|Y = 1) - P(Y = 2)f(\tau|Y = 2) \\&\Leftrightarrow P(Y = 1)f(\tau|Y = 1) = P(Y = 2)f(\tau|Y = 2).\end{aligned}$$

This doesn't necessarily define τ uniquely; there can be more than one solution to this equation. Also, they include (local and global) maxima as well as (local and global) minima. For example, suppose $f(x|Y = i) \sim N(\mu_i, \sigma^2)$, $i = 1, 2$ with $P(Y = 1) = P(Y = 2) = 1/2$ as in 3. If $\mu_1 < \mu_2$, then $\tau = (\mu_1 + \mu_2)/2$ satisfies the above criterion, but it maximizes the probability of error.

2 [10 points]

For any non-negative a and b ,

$$\begin{aligned}\min(a, b) &= \{\min(\sqrt{a}, \sqrt{b})\}^2 \\&\leq \min(\sqrt{a}, \sqrt{b}) \cdot \max(\sqrt{a}, \sqrt{b}) \\&= \sqrt{ab}.\end{aligned}$$

A two category Bayes classifier h satisfies

$$h(x) = \operatorname{argmax}_{i=1,2} P(Y = i|x).$$

Then for this classifier, we have

$$\begin{aligned} P(\text{error}|x) &= 1 - P(Y = h(x)|x) \\ &= 1 - \max_{i=1,2} P(Y = i|x) \\ &= \min_{i=1,2} P(Y = i|x) \quad (\text{since there are only two categories}) \\ &= \min_{i=1,2} \frac{f(x|Y = i)P(Y = i)}{f(x)} \quad (\text{Bayes' theorem}). \end{aligned}$$

Therefore,

$$\begin{aligned} P(\text{error}) &= \int P(\text{error}|x)f(x)dx \\ &= \int \min_{i=1,2} f(x|Y = i)P(Y = i)dx \\ &\leq \int \sqrt{f(x|Y = 1)P(Y = 1) \cdot f(x|Y = 2)P(Y = 2)}dx \\ &= \sqrt{P(Y = 1)P(Y = 2)} \int \sqrt{f(x|Y = 1)f(x|Y = 2)}dx \\ &= \sqrt{P(Y = 1)(1 - P(Y = 1))} \int \sqrt{f(x|Y = 1)f(x|Y = 2)}dx \\ &\leq \frac{1}{2} \int \sqrt{f(x|Y = 1)f(x|Y = 2)}dx. \end{aligned}$$

3 [10 points]

Note that

$$\begin{aligned} P(\text{error}) &= \int \min_{i=1,2} f(x|Y = i)P(Y = i)dx \quad (\text{as in Problem 2}) \\ &= \frac{1}{2} \int \min_{i=1,2} f(x|Y = i)dx \end{aligned}$$

and

$$f(x|Y = i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}, \quad i = 1, 2.$$

Let $m = \min(\mu_1, \mu_2)$ and $M = \max(\mu_1, \mu_2)$. Since

$$\begin{cases} |x - m| \leq |x - M| & \text{for } x \leq \frac{m+M}{2}, \\ |x - m| > |x - M| & \text{otherwise,} \end{cases}$$

we have

$$\min_{i=1,2} f(x|Y = i) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-M)^2}{2\sigma^2}} & \text{if } x \leq \frac{m+M}{2}, \\ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned}
P(\text{error}) &= \frac{1}{2} \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{\frac{m+M}{2}} e^{-\frac{(x-M)^2}{2\sigma^2}} dx + \int_{\frac{m+M}{2}}^{\infty} e^{-\frac{(x-m)^2}{2\sigma^2}} dx \right) \\
&= \frac{1}{2} \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{-\frac{M-m}{2}} e^{-\frac{u^2}{2}} du + \int_{\frac{M-m}{2}}^{\infty} e^{-\frac{v^2}{2}} dv \right) \quad (\text{integration by substitution}) \\
&= \frac{1}{\sqrt{2\pi}} \int_{\frac{M-m}{2}}^{\infty} e^{-\frac{u^2}{2}} du \\
&= \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{-\frac{u^2}{2}} du \\
&= 1 - \Phi(a) \\
&\rightarrow 0 \text{ as } a \rightarrow \infty.
\end{aligned}$$

4 [15 points]

(a)

Solve the following equation for α :

$$\begin{aligned}
0 &= \text{cov}(Z, X_1) \\
&= \text{cov}(X_2 - \alpha X_1, X_1) \\
&= \text{cov}(X_2, X_1) - \alpha \text{var}(X_1) \\
&= a - \alpha v_1.
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
\alpha &= \frac{a}{v_1}, \\
Z &= X_2 - \frac{a}{v_1} X_1.
\end{aligned}$$

(b)

Since $Z = X_2 - \frac{a}{v_1} X_1$ is a linear combination of two normal distributions, Z is also normal. Since Z and X_1 are jointly normal, $\text{cov}(Z, X_1) = 0$ implies that they are independent. The variance of Z is

$$\begin{aligned}
\text{var}(Z) &= \text{var}\left(X_2 - \frac{a}{v_1} X_1\right) \\
&= \text{var}(X_2) + \left(\frac{a}{v_1}\right)^2 \text{var}(X_1) - 2\frac{a}{v_1} \text{cov}(X_2, X_1) \\
&= v_2 + \left(\frac{a}{v_1}\right)^2 v_1 - 2\frac{a}{v_1} a \\
&= v_2 - \frac{a^2}{v_1}.
\end{aligned}$$

(c)

The conditional mean and variance of X_2 given $X_1 = x$ are

$$\begin{aligned} E(X_2|X_1 = x) &= E\left(\frac{a}{v_1}x + Z|X_1 = x\right) \\ &= \frac{a}{v_1}x + E(Z|X_1 = x) \\ &= \frac{a}{v_1}x + E(Z) \quad (\text{by independence}) \\ &= \frac{a}{v_1}x, \\ \text{var}(X_2|X_1 = x) &= \text{var}\left(\frac{a}{v_1}x + Z|X_1 = x\right) \\ &= \text{var}(Z|X_1 = x) \\ &= \text{var}(Z) \quad (\text{by independence}) \\ &= v_2 - \frac{a^2}{v_1}. \end{aligned}$$

5 [15 points]

(a)

Let $\xi = (\xi_1, \xi_2)^t$ and $X = (X_1, X_2)^t$. Note that if $Y = (Y_1, Y_2)^t \sim N(\mu, \Sigma)$, then its moment generating function is given by

$$E(e^{\xi^t Y}) = e^{\xi^t \mu + \frac{1}{2} \xi^t \Sigma \xi}.$$

The moment generating function of X is

$$\begin{aligned} E(e^{\xi^t X}) &= E(E(e^{\xi^t X}|X_1)) \\ &= E(E(e^{\xi_1 X_1 + \xi_2 X_2}|X_1)) \\ &= E(e^{\xi_1 X_1} E(e^{\xi_2 X_2}|X_1)) \\ &= E(e^{(\xi_1 + \xi_2)X_1 + \frac{1}{2}\xi_2^2\sigma_2^2}) \\ &= e^{\frac{1}{2}\xi_2^2\sigma_2^2} E(e^{(\xi_1 + \xi_2)X_1}) \\ &= e^{\frac{1}{2}\xi_2^2\sigma_2^2 + (\xi_1 + \xi_2)\mu + \frac{1}{2}(\xi_1 + \xi_2)^2\sigma_1^2} \\ &= e^{\xi^t \mu^* + \frac{1}{2} \xi^t \Sigma^* \xi}, \end{aligned}$$

where $\mu^* = \begin{pmatrix} \mu \\ \mu \end{pmatrix}$ and $\Sigma^* = \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix}$. Therefore, by the uniqueness property of moment generating functions,

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix}\right).$$

(b)

Since $X'_2 = X_1 + Z$ is a linear combination of normal distributions, X'_2 is normal and X_1, X'_2 are jointly normal. Therefore, the distribution of $(X_1, X'_2)^t$ is determined by its mean vector and covariance matrix. Since

$$\begin{aligned} E(X'_2) &= E(X_1 + Z) = EX_1 + EZ = \mu, \\ \text{var}(X'_2) &= \text{var}(X_1 + Z) = \text{var}(X_1) + \text{var}(Z) = \sigma_1^2 + \sigma_2^2, \\ \text{cov}(X_1, X'_2) &= \text{cov}(X_1, X_1 + Z) = \text{var}(X_1) + \text{cov}(X_1, Z) = \sigma_1^2 + 0 = \sigma_1^2, \end{aligned}$$

we have

$$\begin{pmatrix} X_1 \\ X'_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix} \right).$$

(c)

X_2 has the same marginal distribution as $X'_2 \sim N(\mu, \sigma_1^2 + \sigma_2^2)$.

6 [10 points]

The variance of \hat{a} is

$$\begin{aligned} \text{var}(\hat{a}) &= \text{var} \left(\frac{1}{n} \sum_i X_{i1} X_{i2} \right) \\ &= \frac{1}{n^2} \left(\sum_i \text{var}(X_{i1} X_{i2}) + \sum_{i \neq j} \text{cov}(X_{i1} X_{i2}, X_{j1} X_{j2}) \right) \\ &= \frac{1}{n^2} (n \text{var}(X_1 X_2) + 0) \\ &= \frac{1}{n} \left(E((X_1 X_2)^2) - \{E(X_1 X_2)\}^2 \right). \end{aligned}$$

Note that $E(X_1 X_2) = \text{cov}(X_1, X_2) + E(X_1)E(X_2) = a$ and

$$\begin{aligned} E((X_1 X_2)^2) &= E(E(X_1^2 X_2^2 | X_1)) \\ &= E(X_1^2 E(X_2^2 | X_1)) \\ &= E \left(X_1^2 \left(\text{var}(X_2 | X_1) + \{E(X_2 | X_1)\}^2 \right) \right) \\ &= E \left(X_1^2 \left(v_2 - \frac{a^2}{v_1} + \left(\frac{a}{v_1} \right)^2 X_1^2 \right) \right) \\ &= \left(v_2 - \frac{a^2}{v_1} \right) E(X_1^2) + \left(\frac{a}{v_1} \right)^2 E(X_1^4) \\ &\stackrel{(*)}{=} \left(v_2 - \frac{a^2}{v_1} \right) v_1 + \left(\frac{a}{v_1} \right)^2 3v_1^4 \\ &= v_1 v_2 + 2a^2, \end{aligned}$$

where (*) is from $E(X_1^4) = \frac{\partial^4}{\partial t^4} E(e^{tX_1}) \Big|_{t=0} = \frac{\partial^4}{\partial t^4} e^{\frac{1}{2}t^2v_1} \Big|_{t=0} = 3v_1^2$.

Thus,

$$\begin{aligned} \text{var}(\hat{a}) &= \frac{1}{n} \left(E((X_1 X_2)^2) - \{E(X_1 X_2)\}^2 \right) \\ &= \frac{1}{n} (v_1 v_2 + 2a^2 - a^2) \\ &= \frac{1}{n} (v_1 v_2 + a^2). \end{aligned}$$

Note: There is another way to get $E(X_1 X_2)$ and $E((X_1 X_2)^2)$ above. Since

$$\begin{aligned} E(e^{\xi^t X}) &= e^{\frac{1}{2}\xi^t C \xi} \\ &= e^{\frac{1}{2}\xi_1^2 v_1 + 2\xi_1 \xi_2 a + \xi_2^2 v_2}, \end{aligned}$$

$$\begin{aligned} E(X_1 X_2) &= \frac{\partial^2}{\partial \xi_2 \partial \xi_1} E(e^{\xi^t X}) \Big|_{s=t=0} = a, \\ E((X_1 X_2)^2) &= \frac{\partial^4}{\partial \xi_2^2 \partial \xi_1^2} E(e^{\xi^t X}) \Big|_{s=t=0} = 2a^2 + v_1 v_2. \end{aligned}$$

7 [10 points]

Suppose that Σ is a $d \times d$ matrix. Then, the loglikelihood is

$$l(\theta; X_1, X_2, \dots, X_n) = -\frac{nd}{2} \log(2\pi) - \frac{nd}{2} \log \theta - \frac{n}{2} \log |\Sigma| - \frac{1}{2\theta} \sum_{i=1}^n X_i^t \Sigma^{-1} X_i.$$

By setting $\frac{dl}{d\theta} = 0$, we get the MLE $\hat{\theta}_{MLE}$ of θ as

$$\hat{\theta}_{MLE} = \frac{1}{nd} \sum_{i=1}^n X_i^t \Sigma^{-1} X_i.$$

8 [20 points]

(a)

The following is an R code to generate A and compute Σ , λ_{\max} and λ_{\min} .

```
#define a function
randA=function(dim){
A=matrix(0,dim,dim)
  #generate a random matrix of i.i.d N(0,1) variables
  #and make sure it is non-singular
while(det(A)==0){
```

```

A=matrix(rnorm(dim*dim),dim,dim)
}
return(A)
}
#set the dimensions of a random matrix A
d=10
#simulate a non-singular dxd random matrix A
A=randA(d)
#compute Sigma
Sigma=t(A)%*(A)
#compute the largest and smallest eigenvalues
max.eigen=max(eigen(Sigma)$values)
min.eigen=min(eigen(Sigma)$values)

```

(b)

Note that SE_{ij} is given by Problem 6 as

$$SE_{ij} = \sqrt{\text{var}(\hat{\Sigma}_{ij})} = \sqrt{\frac{1}{n}(\Sigma_{ii}\Sigma_{jj} + \Sigma_{ij}^2)}.$$

The following is an R code to compute \mathbf{e} , \mathbf{M} and \mathbf{m} for $d = 10$, $n = 150$.

```

d=10
A=randA(d)

Sigma=t(A)%*(A)
max.eigen=max(eigen(Sigma)$values)
min.eigen=min(eigen(Sigma)$values)

T=100
n=150

# set vectors e, M and n
e=array(0,d*d*T)
M=array(0,T)
m=array(0,T)

SE=matrix(0,d,d)
for(i in 1:d){
for(j in 1:d){
SE[i,j]= sqrt((Sigma[i,i]*Sigma[j,j]+Sigma[i,j]^2)/n)
}
}
for(t in 1:T){

```

```

#sample from N(0,Sigma) and compute the MLE of Sigma
cov.mle=matrix(0,d,d)
for(i in 1:n){
  X=t(A)%*%matrix(rnorm(d),d,1)
  cov.mle=cov.mle+X%*%t(X)
}
cov.mle=cov.mle/n
e[(1+(t-1)*d*d):(t*d*d)]=as.vector((Sigma-cov.mle)/SE)
M[t]=max(eigen(cov.mle)$values)
m[t]=min(eigen(cov.mle)$values)
}

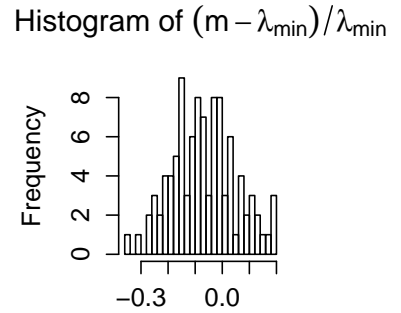
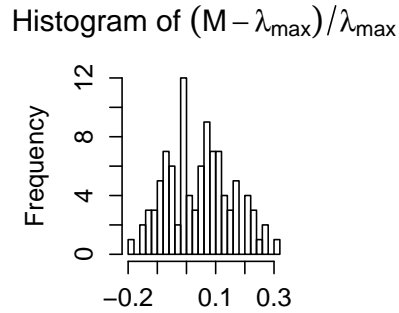
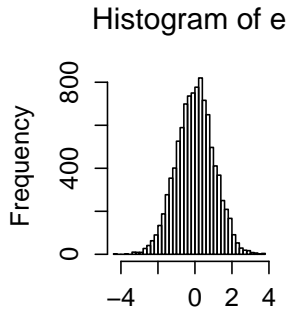
```

The following plots are histograms of \mathbf{e} , $(\mathbf{M} - \lambda_{\max})/\lambda_{\max}$ and $(\mathbf{m} - \lambda_{\min})/\lambda_{\min}$ for $d = 10, 100$ and $n = 150, 1000$.

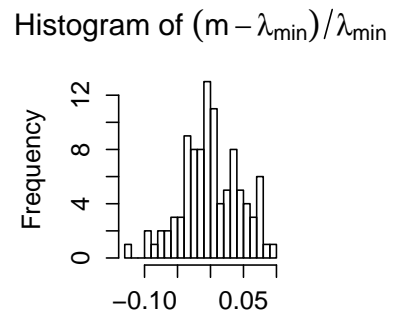
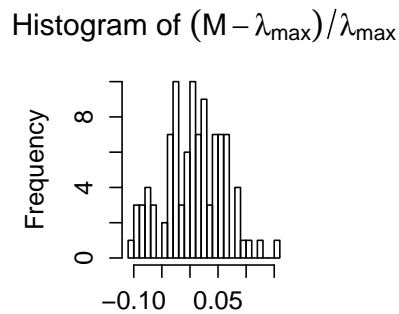
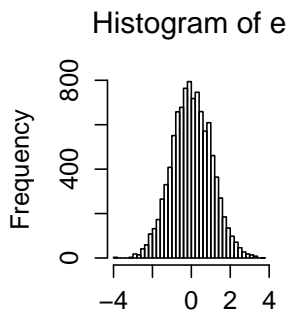
It seems that the standardized error of the covariance estimate approximately follows the standard normal distribution in all cases. For the relative errors of $\hat{\lambda}_{\max}$ and $\hat{\lambda}_{\min}$, their histograms don't clearly show that they follow the normal distribution although they get closer to some extent to bell-shaped distributions as n increases. Therefore, the estimates of the covariance seem to converge faster and are more stable than those of the eigenvalues.

When $d > n$, we don't have enough data points to compute reasonable estimates for all entries of Σ . Estimates might be bad in that $\hat{\lambda}_{\min}$ has negative values.

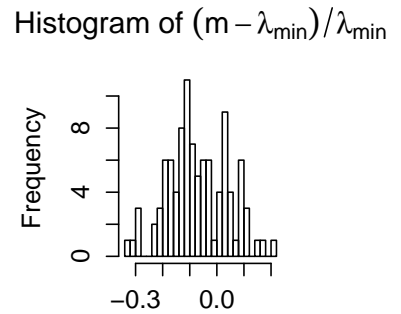
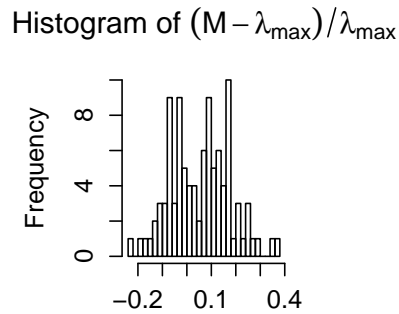
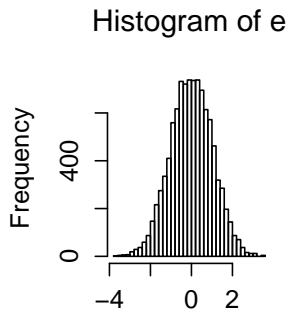
d=10, n=150



d=10, n=1000



d=100, n=150



d=100, n=1000

