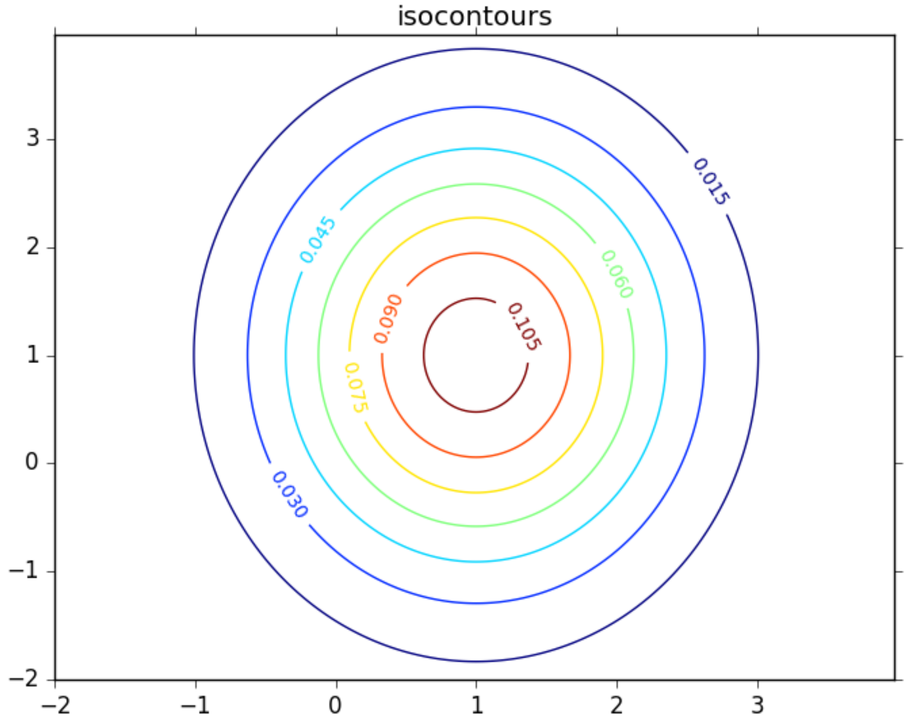
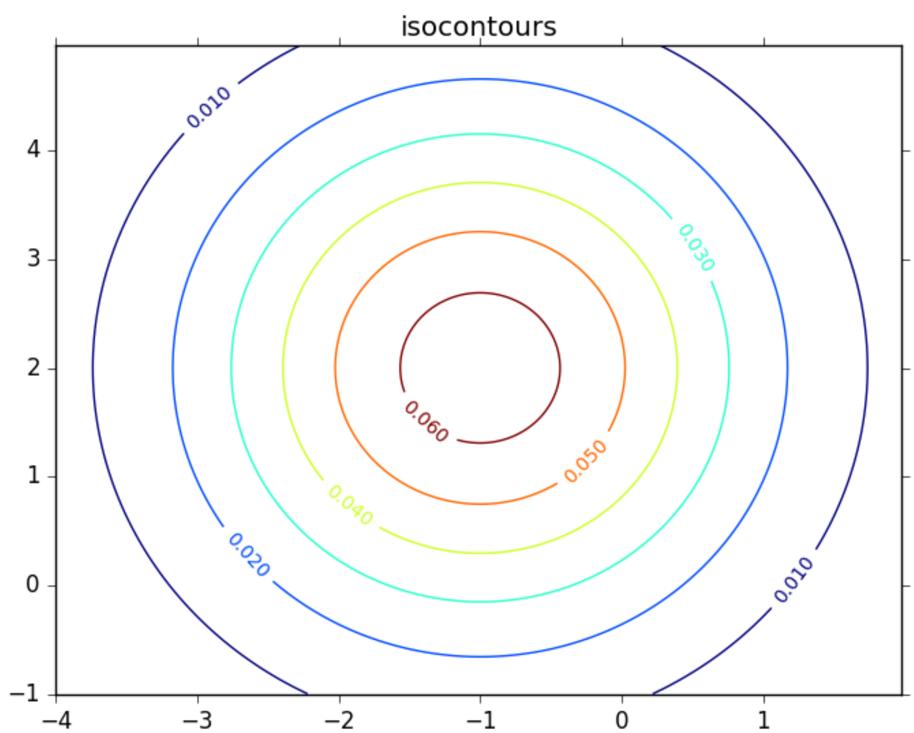
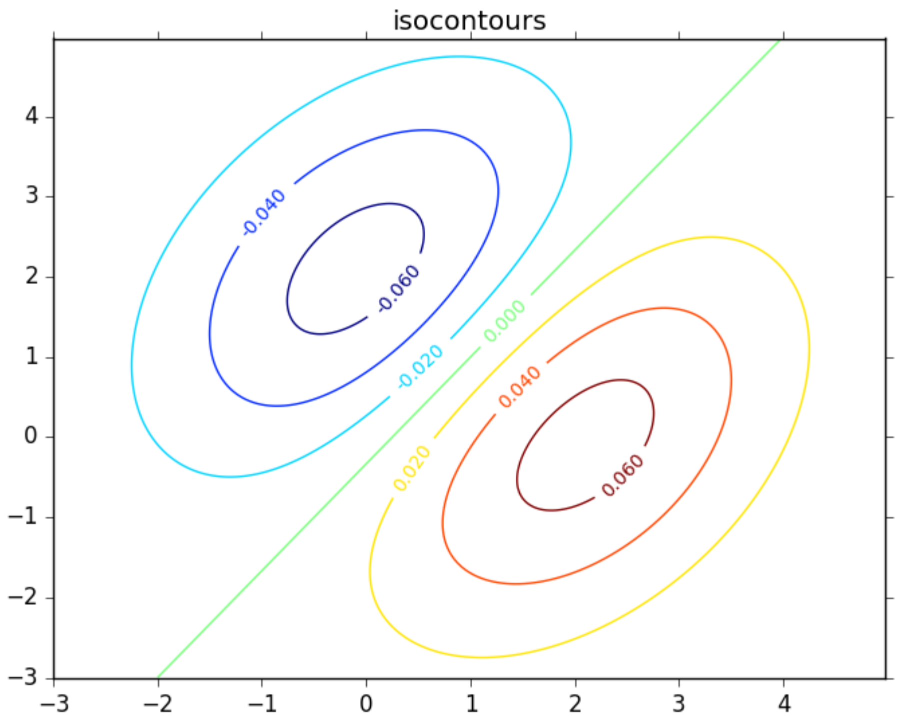
Q2 part(a):



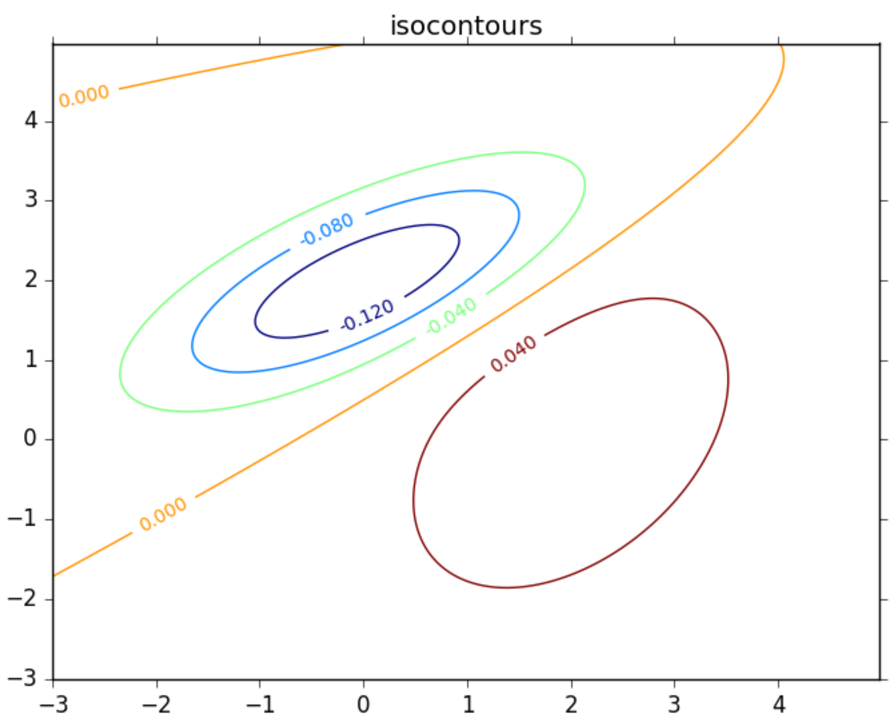
Q2 part(b):



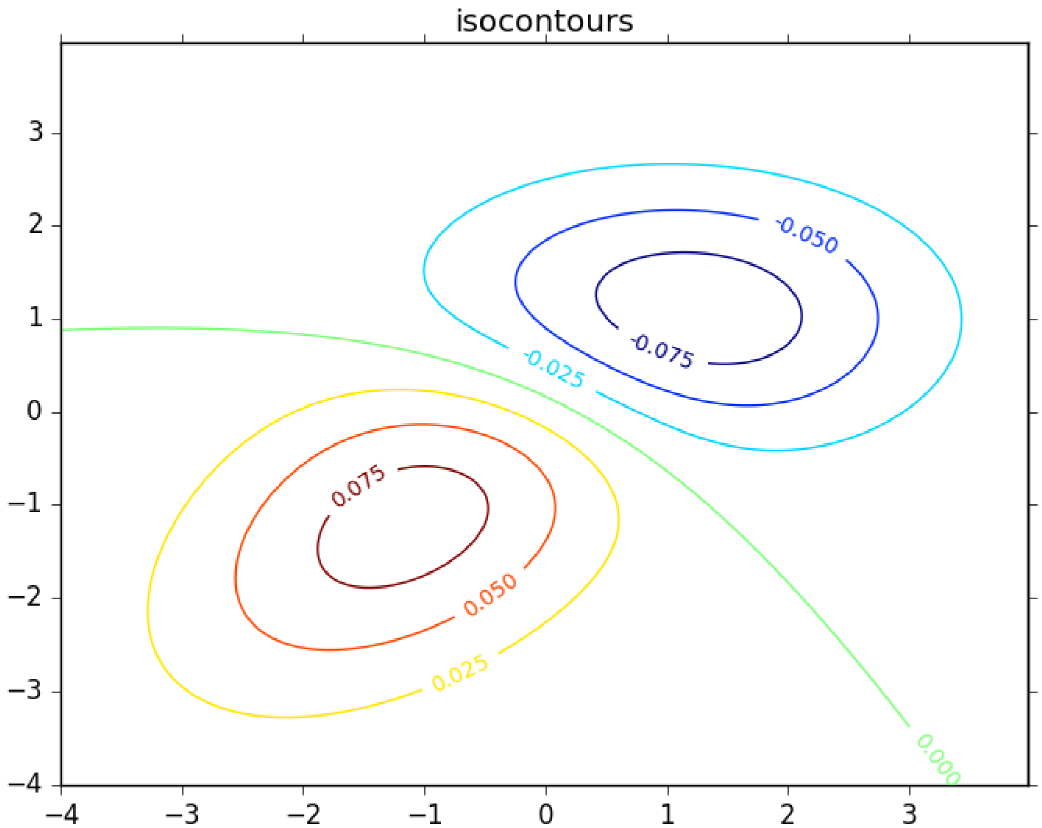
Q2 part(c):



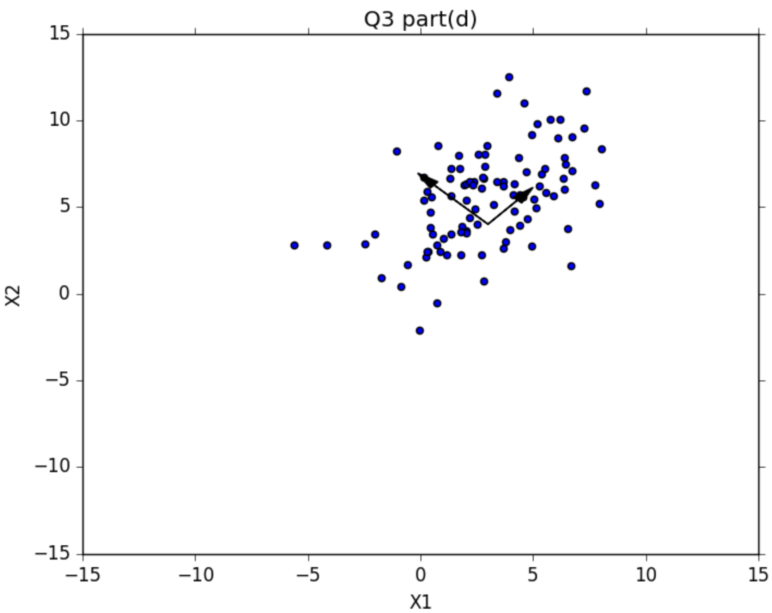
Q2 part(d):



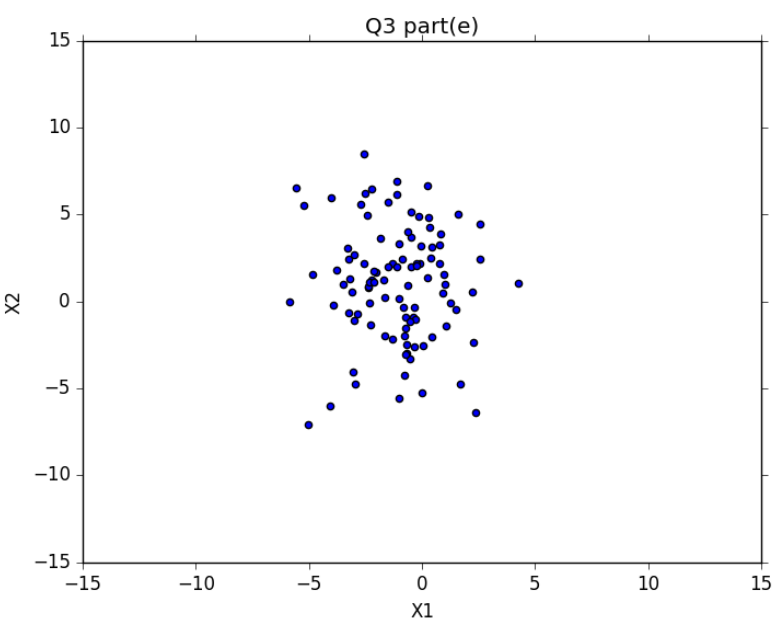
Q2 part(e):


isocontours
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
-0.050
-0.075
-0.025
0.075
0.050
0.025
0.000

```
Q3 part(a): [ 2.96241672  5.50936707]

Q3 part(b): [[ 7.04213055  3.57887393]
 [ 3.57887393  7.45579513]]

Q3 part(c) eigenvalues: [  3.66411721  10.83380847]

Q3 part(c) eigenvectors: [[-0.72721946  0.68640502]
 [ 0.68640502  0.72721946]]

Q3 part(d):
```

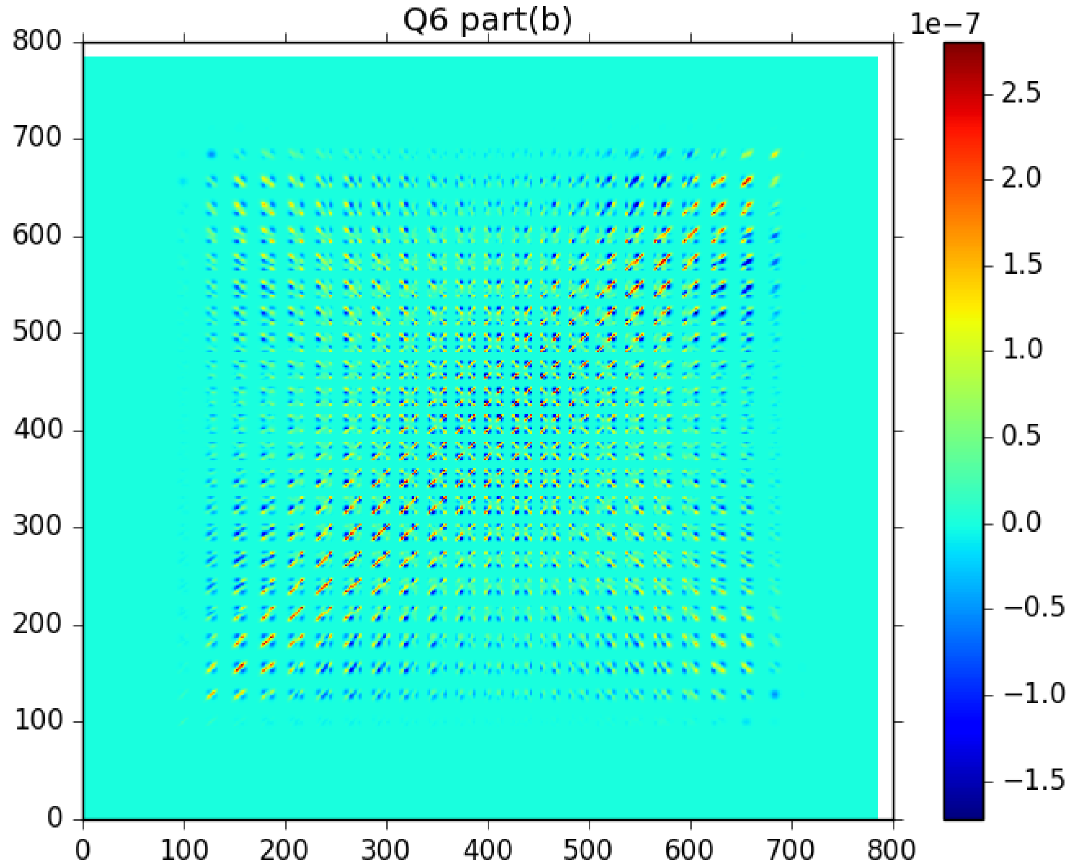


Q3 part(e):



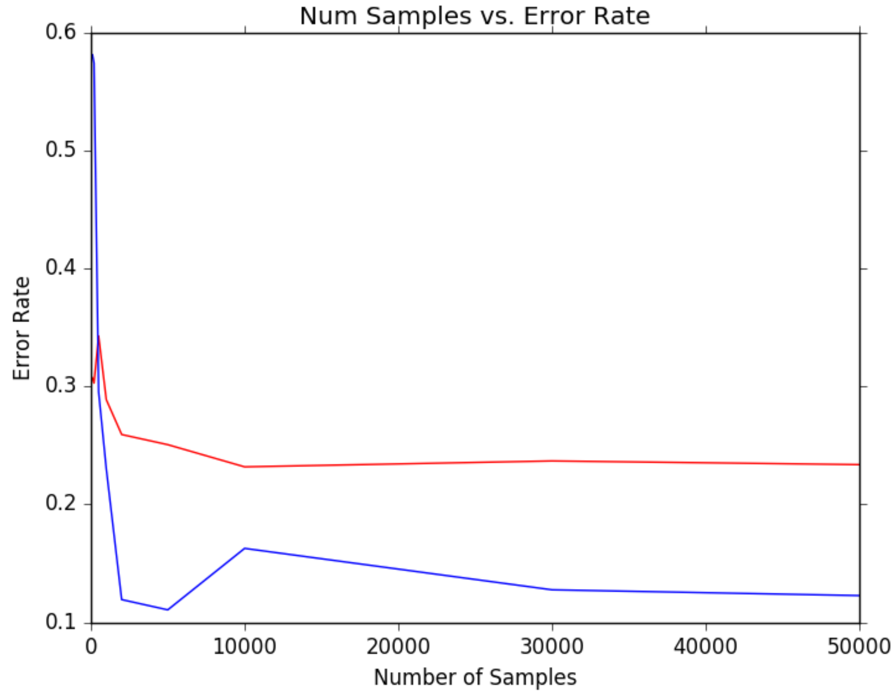
Q6 part(a): (It unnecessarily takes too many pages. Please look at q6_ab() function to see how I compute it.)

Q6 part(b): (visualization of covariance matrix for ‘0’ class)



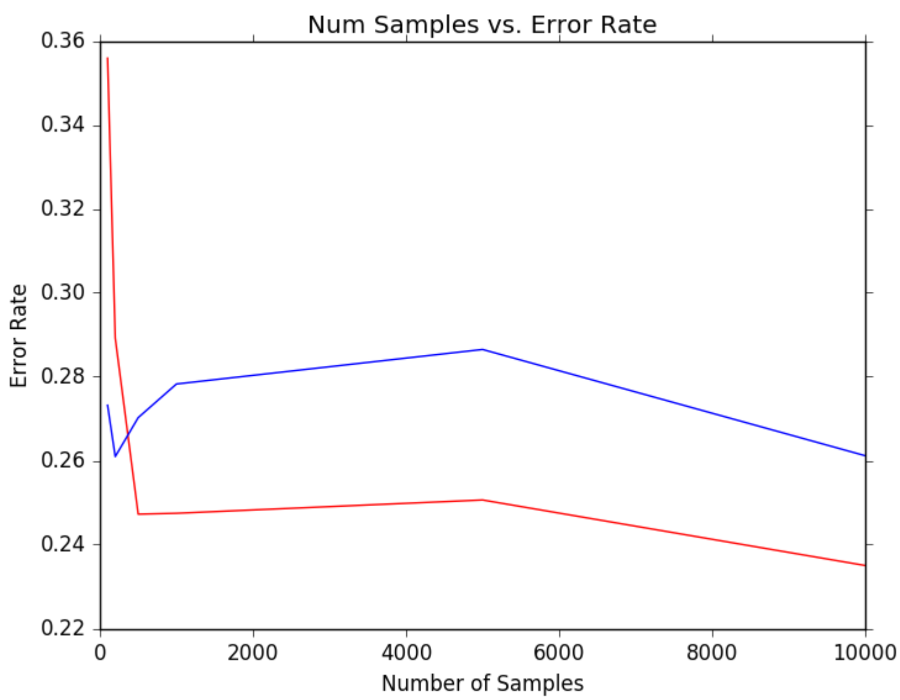
Diagonal terms are more likely to have somewhat higher covariance than off-diagonal terms. We can conclude that the pixels which are close to each other are more likely to relate to each other. The diagonal terms are sometimes zero because that specific features (pixel) are always zero.

Q6 part(c):
Red line is for LDA, and blue line is for QDA for both part c and d



Q6 part(d): (I know that part(d) does not require plotting, but I want to show why I choose LDA for spam data set.)

Kaggle ID: YONG-CHAN_SHIN
Digit score: 0.88000
Spam score: 0.75420



Yong-Chan Shin CS189 HW3

$\boxed{Q1}$ (a) (i) $X, Y \in \{-1, 0, 1\}$

(ii) $P[X=0, Y=0] = P[X\neq 0, Y\neq 0] = 0$

$P[X=0, Y\neq 0] = P[X\neq 0, Y=0] = \frac{1}{2}$

(iii) $P[Y=1 | X=0] = P[Y=-1 | X=0] = \frac{1}{2}$

$P[X=1 | Y=0] = P[X=-1 | Y=0] = \frac{1}{2}$

$\therefore P[X=0, Y=1] = P[X=0, Y=-1] = P[X=1, Y=0] = P[X=-1, Y=0] = \frac{1}{4}$

$\because E[X] = E[Y] = \frac{1}{2}\cdot 0 + \frac{1}{2}\left(\frac{1}{2} + \frac{-1}{2}\right) = 0$

$E[XY] = 0$ as either X or Y is always zero and if $X=0$, then $Y=0$ (vice versa).

$\therefore E[XY] = E[X]E[Y]$

$\therefore$ X, Y are uncorrelated.

$P[X=0] = \frac{1}{2}$, and $P[X=0 | Y=1] = 1$.

$\therefore P(X|Y) \neq P(X)$

Thus, X and Y are not independent.

(b) For Y, B is random in $\{0,1\}$, so $P(X|Y) = P(X)$

For Z, C is random in $\{0,1\}$, so $P(Y|Z) = P(Y)$

For X, D is random in $\{0,1\}$, so $P(Z|X) = P(Z)$

Thus, X, Y, and Z are pairwise independent.

If Y, Z are given, X is determined automatically.

$Y \oplus Z = (C \oplus D) \oplus (B \oplus D) = B \oplus C \oplus 0 = B \oplus C = X$

Thus, X, Y and Z are not mutually independent.

Q4 (a).

Since Σ is diagonal matrix, all the covariance values between any pair of X_i and X_j is zero $(i \neq j)$ so $X_1, \dots X_n$ are mutually independent.

$$\therefore P(X) = \underbrace{P(X_1) \cdot P(X_2) \cdots P(X_n)}_{\text{univariate Gaussians}}$$

$$= \prod_{i=1}^{n} P(X_i)$$

$$= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(X_i-\mu)^T \Sigma^{-1} (X_i-\mu)\right)$$

$$= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \sum_{j=1}^{d} \sigma_j^{-2} (X_{ij}-\mu_j)^2\right)$$

$$= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^d \, |\sigma_1 \cdots \sigma_d|} \exp\left(\frac{-1}{2} \sum_{j=1}^{d} \sigma_j^{-2} (X_{ij}-\mu_j)^2\right)$$

$$\ln P(X) = \ln P(X_1) + \ln P(X_2) + \dots + \ln P(X_n)$$

$$= -n \cdot \ln\left((\sqrt{2\pi})^d \cdot \sigma_1 \cdots \sigma_d\right) + \frac{-1}{2} \sum_{k=1}^{n} \sum_{j=1}^{d} \sigma_j^{-2} (X_{kj}-\mu_j)^2$$

$$\frac{\partial \ln P(X)}{\partial \sigma_i} = -n \cdot \frac{1}{\sigma_i} + \frac{-1}{2} \sum_{k=1}^{n} (X_{ki}-\mu_i)^2 \cdot (-2) \cdot \sigma_i^{-3} = 0$$

$$-n \sigma_i^2 + \frac{-1}{2} \sum_{k=1}^{n} (X_{ki}-\mu_i)^2 \cdot (-2) = 0$$

$$\therefore \hat{\sigma}_i^2 = \frac{\sum_{k=1}^{n} (X_{ki}-\hat{\mu}_i)^2}{n}$$

$$\nabla_\mu \ln P(X) = \begin{bmatrix} \frac{\partial}{\partial \mu_1} \ln P(X) \\ \vdots \\ \frac{\partial}{\partial \mu_d} \ln P(X) \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \cdot 2 \cdot (-1) \cdot \sum_{k=1}^{n} \sigma_1^{-2} (X_{k1}-\mu_1) \\ \vdots \\ -\frac{1}{2} \cdot 2 \cdot (-1) \cdot \sum_{k=1}^{n} \sigma_d^{-2} (X_{kd}-\mu_d) \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} \sigma_1^{-2} (X_{k1}-\mu_1) \\ \vdots \\ \sum_{k=1}^{n} \sigma_d^{-2} (X_{kd}-\mu_d) \end{bmatrix} = 0$$

$$\therefore \hat{\mu}_i = \frac{1}{n} \sum_{k=1}^{n} X_{ki}$$

$$\therefore \hat{\mu} = \frac{1}{n} \begin{bmatrix} \sum_{k=1}^{n} X_{k1} \\ \sum_{k=1}^{n} X_{k2} \\ \vdots \\ \sum_{k=1}^{n} X_{kd} \end{bmatrix} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

(Q4)

(b) We already got that

$$\ln P(X) = -n\ln\left((\sqrt{2\pi})^d \sigma_1 \cdots \sigma_d\right) - \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{d} \sigma_j^{-2}(X_{kj} - \mu_j)^2$$

Now, instead of μ, we'll use $A\mu$.

Then, $\ln P(X) = -n\ln\left((\sqrt{2\pi})^d \sigma_1 \cdots \sigma_d\right) - \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{d} \sigma_j^{-2}\left(X_{kj} - (A\mu)_j\right)^2$

$$\nabla_\mu \ln P(X) = \begin{bmatrix} \sum_{k=1}^{n} \sigma_1^{-2}(X_{k1} - (A\mu)_1) \\ \vdots \\ \sum_{k=1}^{n} \sigma_d^{-2}(X_{kd} - (A\mu)_d) \end{bmatrix} = 0.$$

$$(A\mu)_j = \frac{1}{n}\sum_{k=1}^{n} X_{kj}$$

$$\therefore A\hat{\mu} = \frac{1}{n}\begin{bmatrix} \sum_{k=1}^{n} X_{k1} \\ \vdots \\ \sum_{k=1}^{n} X_{kd} \end{bmatrix}$$

$$\therefore \hat{\mu} = \frac{1}{n} A^{-1}\begin{bmatrix} \sum_{k=1}^{n} X_{k1} \\ \vdots \\ \sum_{k=1}^{n} X_{kd} \end{bmatrix} = \frac{1}{n} A^{-1} \sum_{k=1}^{n} X_k$$

Q5 (a) $\hat{\Sigma}$ is not invertible when all the sample points lie on a hyperplane which has, by its definition, has dimension strictly less than d.

(b) Instead of using the raw covariance matrix, add $\alpha \cdot I$ where α is an hyperparameter and I is $d \times d$ identity matrix. α Should be small enough to not change the result too much, so value around minimum of the non-zero eigenvalue of the covariance matrix will be proper enough.

(c) From HW2 Q4 (a) and (b), we get a corollary

$$\begin{cases} \lambda_{max}(\Sigma^{-1}) = \max_{\|x\|_2=1} x^T \Sigma^{-1} x \\ \lambda_{min}(\Sigma^{-1}) = \min_{\|x\|_2=1} x^T \Sigma^{-1} x \end{cases} \quad \text{where } \Sigma^{-1} \geq 0 \text{ and symmetric}$$

and we know that value that maximizes/minimizes $\ln f(x)$ maximizes/minimizes $f(x)$. Also, $\mu = 0$, so we want x such that maximizes/minimizes

$$\ln f(x) = -\ln\sqrt{(2\pi)^d |\Sigma|} - \frac{1}{2} x^T \Sigma^{-1} x$$

Since $-\ln\sqrt{(2\pi)^d|\Sigma|}$ is constant, vector that maximizes $x^T\Sigma^{-1}x$ will minimize $\ln f(x)$, and vector that minimizes $x^T\Sigma^{-1}x$ will maximize $\ln f(x)$ due to the negative sign in front of $\frac{1}{2}x^T\Sigma^{-1}x$. From the corollary above, if we choose eigenvector of $\lambda_{min}(\Sigma^{-1})$, it will maximize $f(x)$, and if we chose eigen vector of $\lambda_{max}(\Sigma^{-1})$, it will minimize $f(x)$.