Homework 3: Gaussians

Sreesha Venkat

February 27, 2017

Collaborators: Sydney McMuldroch, Lynn Kong, Roshni Patel

Q1. Independence vs. Correlation

(a) X and Y are uncorrelated, but not independent.

To see that X and Y are not independent, we must prove that $P(X|Y) \neq P(X)$. We can see that this holds true for X and Y, because $P(X = 0|Y = 1) = 1$, but $P(X = 0) = 1/2$. Therefore X and Y are not independent.

To see that X and Y are uncorrelated, we must prove that their covariance is 0. Note that when X is a nonzero value, Y is zero, and vice versa. By graphing the potential values of (X, Y) in a plot, we also see that each of the four resulting points comes up with equal probability of $\frac{1}{4}$. From here we see that $E[X] = E[Y] = 0$, therefore $E[XY] = 0$. This tell us that the two terms are uncorrelated.

(b) X, Y, and Z are pairwise independent but not mutually independent.

To prove pairwise independence, we must show that for each pair of variables, $P(A, B) = P(A)P(B)$. We can see this for X and Y, as $P(X, Y) = P(X = 1)P(Y = 1) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$, and $P(X, Y) = P(X = 1)P(Y = 1) = P(B = 1, C = 0, D = 1) + P(B = 0, C = 1, D = 0) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$.

However, they are not mutually independent because we can derive X from $Y \oplus Z$.

Q2. Isocontours of Normal Distributions

Figure 1: 2a

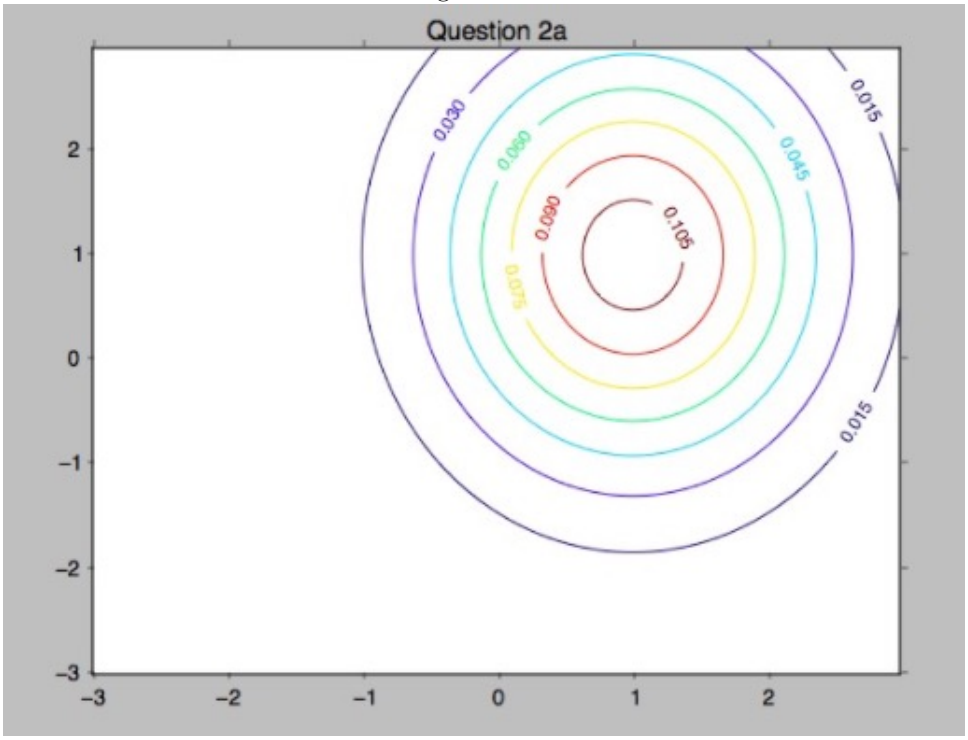


Figure 2: 2b

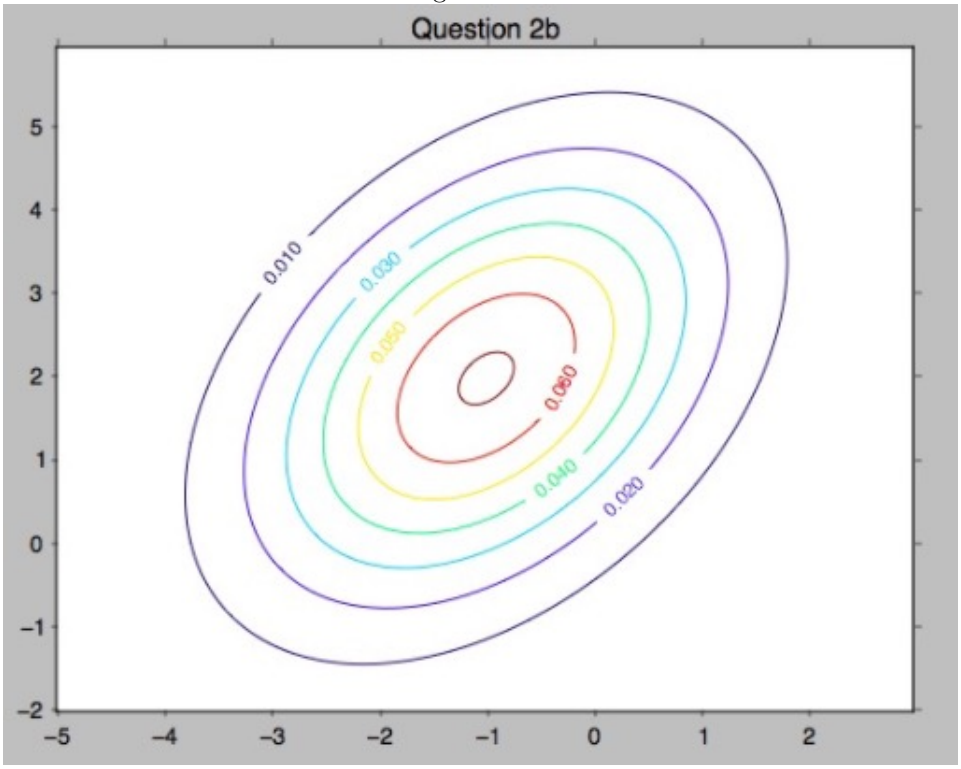


Figure 3: 2c

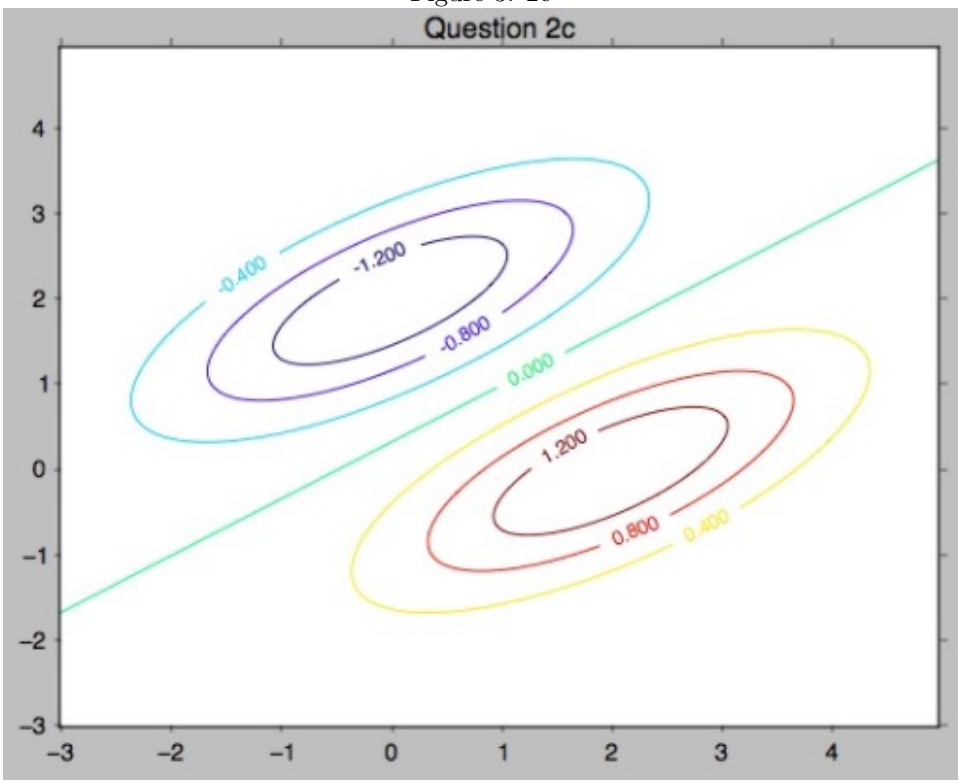


Figure 4: 2d

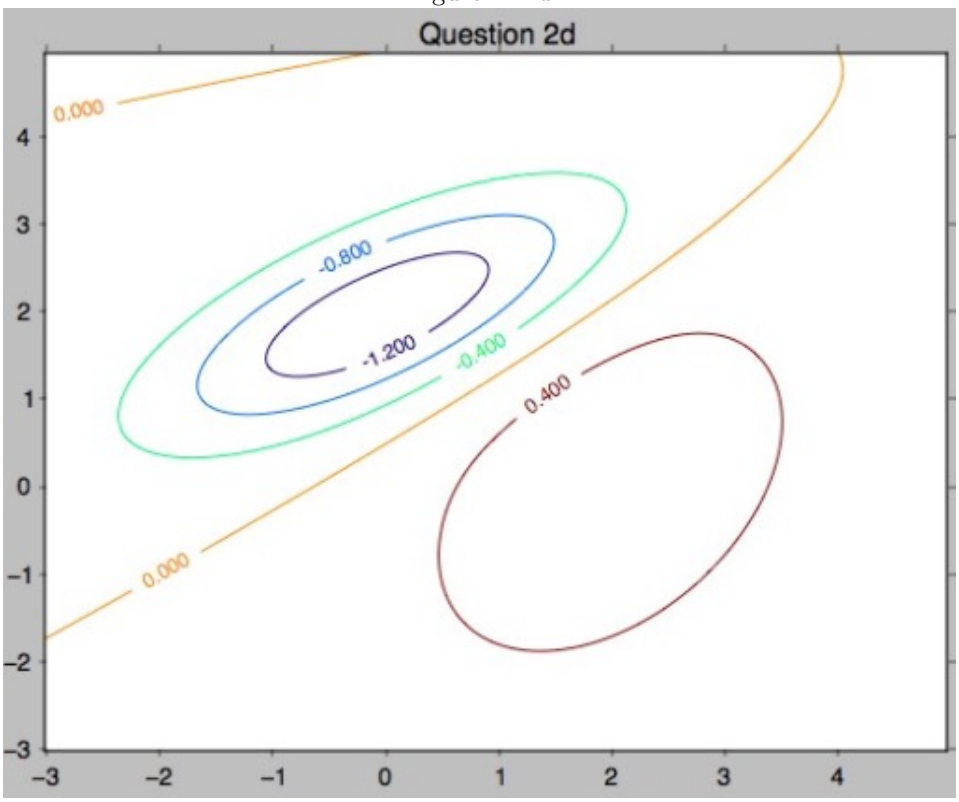
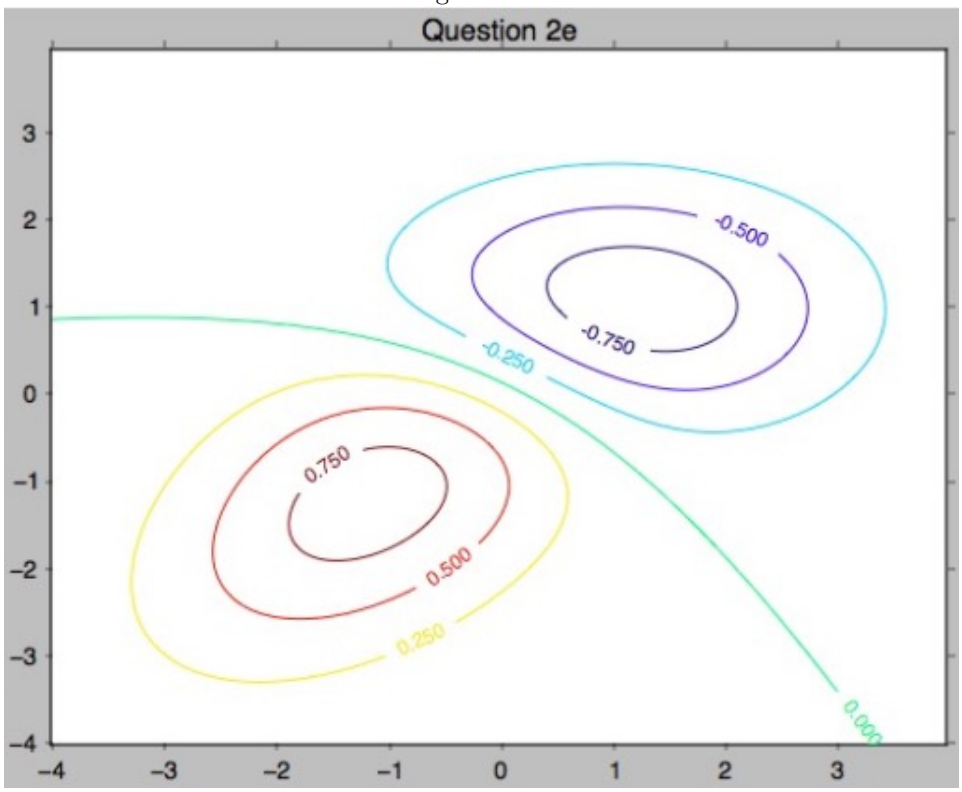


Figure 5: 2e



Q3. Eigenvectors of the Gaussian Covariance Matrix

(a) $\mu = \begin{bmatrix} 3.13439813 \\ 5.67562221 \end{bmatrix}$

(b) $\Sigma = \begin{bmatrix} 9.18222344 & 5.07265441 \\ 5.07265441 & 6.99166833 \end{bmatrix}$

(c) $v_1 = \begin{bmatrix} 0.77815626 \\ -0.62807072 \end{bmatrix}$ with corresponding eigenvalue 13.27649844.

$v_2 = \begin{bmatrix} 0.62807072 \\ 0.77815626 \end{bmatrix}$ with corresponding eigenvalue 2.89739334.

Figure 6: 3d

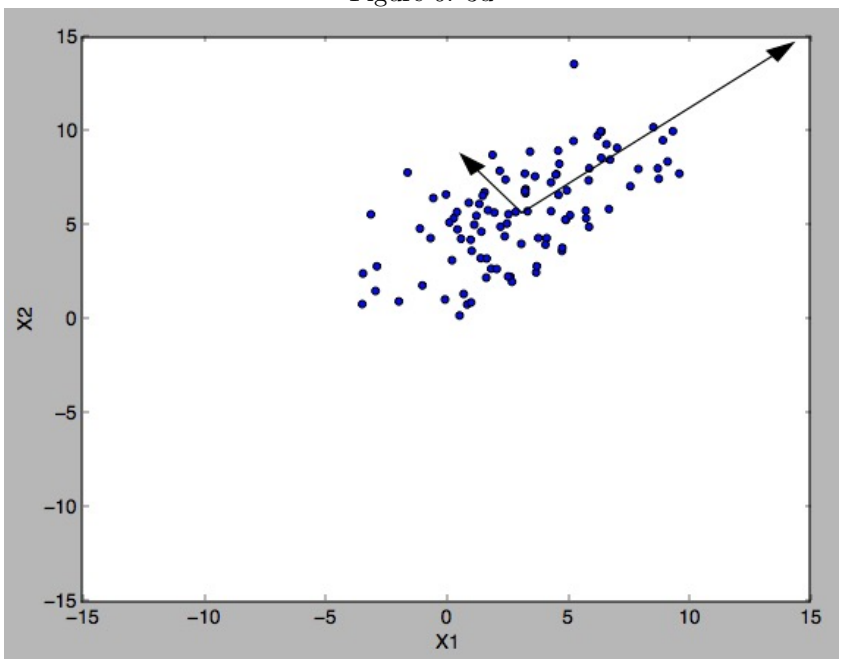
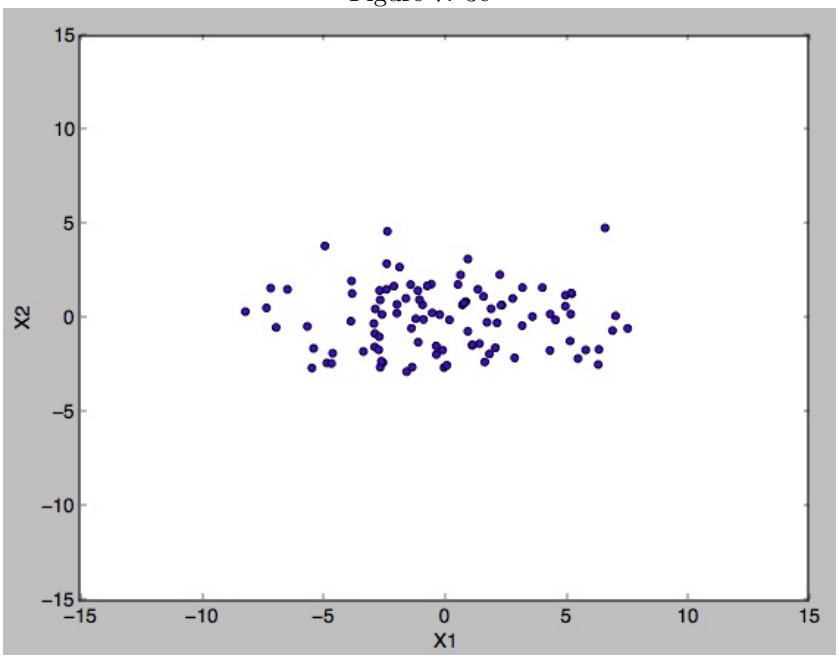


Figure 7: 3e



Q4. Maximum Likelihood Estimation

(a)

$$\mathcal{L} = \left(\frac{1}{(\sqrt{2\pi})^d\sqrt{|\Sigma|}}\right)^n \Pi_i \exp\Big(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big)$$

$$l = \frac{-nd}{2}(ln(2\pi)) - nln(\sigma) - \Big(\frac{1}{2}\Sigma_i(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big)$$

We can then expand out l to get:

$$l = \frac{-nd}{2}(ln(2\pi)) - nln(\sigma) - \frac{1}{2}\Sigma_i x_i^T\Sigma^{-1}x_i - \mu^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu$$

We can than take the transpose of $\mu^T\Sigma^{-1}x_i$, because it is a scalar and Σ^{-1} is symmetric, to get $x_i^T\Sigma^{-1}\mu$

$$l = \frac{-nd}{2}(ln(2\pi)) - nln(\sigma) - \frac{1}{2}\Sigma_i x_i^T\Sigma^{-1}x_i - 2x_i^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu$$

$$\frac{\partial l}{\partial \mu} = \Sigma_i x_i^T\Sigma^{-1} - \Sigma^{-1}\mu = 0$$

$$n\Sigma^{-1}\mu = \Sigma_i x_i^T\Sigma^{-1}$$

$$\mu = \frac{1}{n}\Sigma_i\Sigma x_i^T\Sigma^{-1}$$

$$\mu = \frac{1}{n}\Sigma_i x_i$$

For σ_j:

$$\frac{\partial l}{\partial \sigma_j} = -\frac{n}{\sigma_j} + \frac{\Sigma_i(x_{ij}-\mu_j)^2}{\sigma_j^3} = 0$$

$$\frac{n}{\sigma_j} = \frac{\Sigma_i(x_{ij}-\mu_j)^2}{\sigma_j^3}$$

$$\sigma_j^2 = \frac{\Sigma_i(x_{ij}-\mu_j)^2}{n}$$

$$\sigma = \sqrt{\frac{\Sigma_i(x_{ij}-\mu_j)^2}{n}}$$

(b)

$$\mathcal{L} = \left(\frac{1}{(\sqrt{2\pi})^n\sqrt{|A|^n}}\right)\exp -\frac{1}{2}\Sigma_i((x_i-A\mu)^T\Sigma^{-1}(x_i-A\mu)\Big)$$

$$l = \frac{-n}{2}ln(2\pi) - \frac{n}{2}ln(\Sigma) - \frac{1}{2}\Sigma_i((x_i-A\mu)^T\Sigma^{-1}(x_i-A\mu)\Big)$$

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2}\Sigma_i(2A^T\Sigma^{-1}A\mu - 2A^T\Sigma^{-1}X_i)\Big)$$

$$\frac{\partial l}{\partial \mu} = A^T\Sigma^{-1}\Sigma_i(X_i - A\mu) = 0$$

$$\mu = A^{-1}\frac{\Sigma_i X_i}{n}$$

Q5. Covariance Matrices and Decompositions

(a) $\hat{E}$ is not invertible if there exists a hyperplane in d - 1 dimensions such that the sample points X_i can lie on that plane. In linear algebra terms, this essentially equates to the matrix $\hat{E}$ not having full rank.

(b) Our starting matrix $\hat{E}$ has dimension dxd. One way to fix it so that it becomes invertible is to loop through the matrix's corresponding column and row pairs, removing each pair one at a time. After removing a row-column pair, we check to see that our matrix has $d-1$ rank and is of dimension $(d-1)*(d-1)$. If this holds true then we are done, because our matrix has full rank and is now invertible. If not, replace the row-column pair that was removed, and move on to the next pair. Repeat this until you find the invertible matrix.

(c) If $\mu = 0$, $f(x)$ becomes:

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

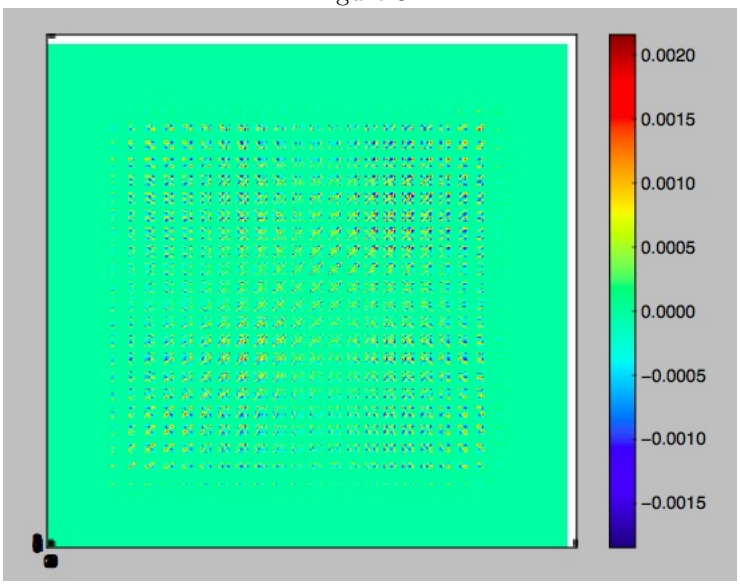
The value of $f(x)$ relies on the value it is being raised to in the latter part of the equation. If e is raised to a large negative number, the value of $f(x)$ becomes smaller, whereas raising it to a small negative number causes the value of $f(x)$ to grow larger. Therefore, the vector x of length 1 that minimizes $f(x)$ is the one that picks up the largest value of Σ^{-1}, while the x that maximizes $f(x)$ is the one that picks up the smallest value of Σ^{-1}.

Q6. Gaussian Classifiers for Digits and Spam

(a) See code in appendix; Kaggle Username: SreeshaVenkat

(b) The diagonal terms have a higher covariance in comparison to the off-diagonal terms. From this we can conclude that the covariance between each sample point and itself, as well as each sample point and adjacent points, is higher than the covariance between a sample point and a point farther away.

Figure 8:

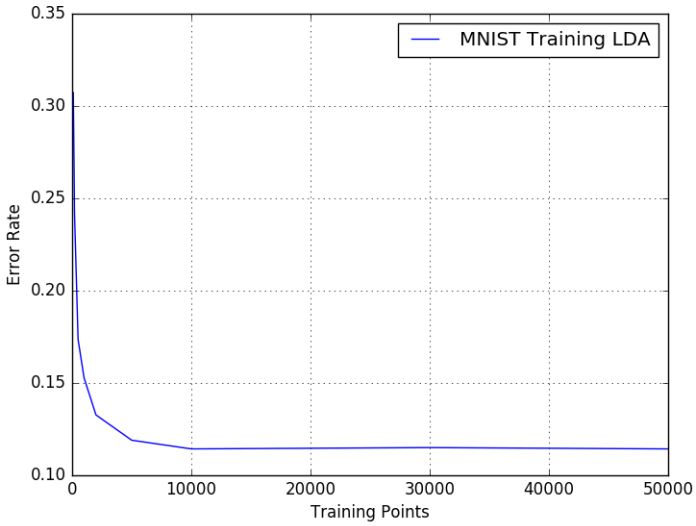


(c)

(i)

Training Points	Error Rate
100	.3073
200	.2423
500	.1736
1000	.1528
2000	.1327
5000	.119
10000	.1143
30000	.115
50000	.1143

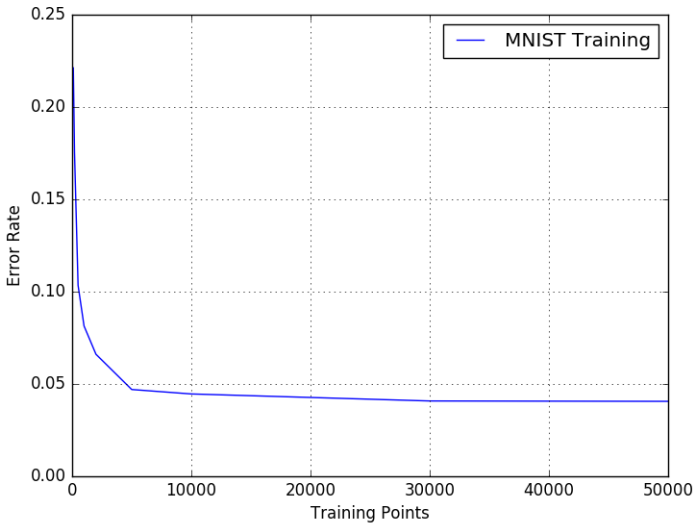
Figure 9:



(ii)

Training Points	Error Rate
100	.2212
200	.1755
500	.1034
1000	.0814
2000	.0661
5000	.047
10000	.0446
30000	.0408
50000	.0406

Figure 10:



(iii) QDA performed better, because it looks at the estimated covariance matrix for each class,

whereas LDA looks at the mean of covariance matrices across all classes, which ultimately ends up being less accurate.

(iv) Optimum Prediction Rate: 0.94980

(d) Optimum Prediction Rate: 0.77260