Question 1

(a)

This is implemented in the R script (see Rcode).

Plot showing the digit means is shown in Figure 1.

Plot showing pooled within-digit covariance matrix of the digits data is shown in Figure 2.

(b)

This is implemented in the R script (see Rcode).

Plot showing both the fraction of variance explained by each eigenvector and the cumulative amount of variance explained as a function of the number of eigenvectors retained is shown in Figure 3.

Report the minimal number (K) of eigenvectors needed to explain 99% of the overall variance in the data: K=346

(c)

This is implemented in the R script (see Rcode).

Plot showing the percent of variance explained as a function of

the digit labels 0-9 is shown in Figure 4.

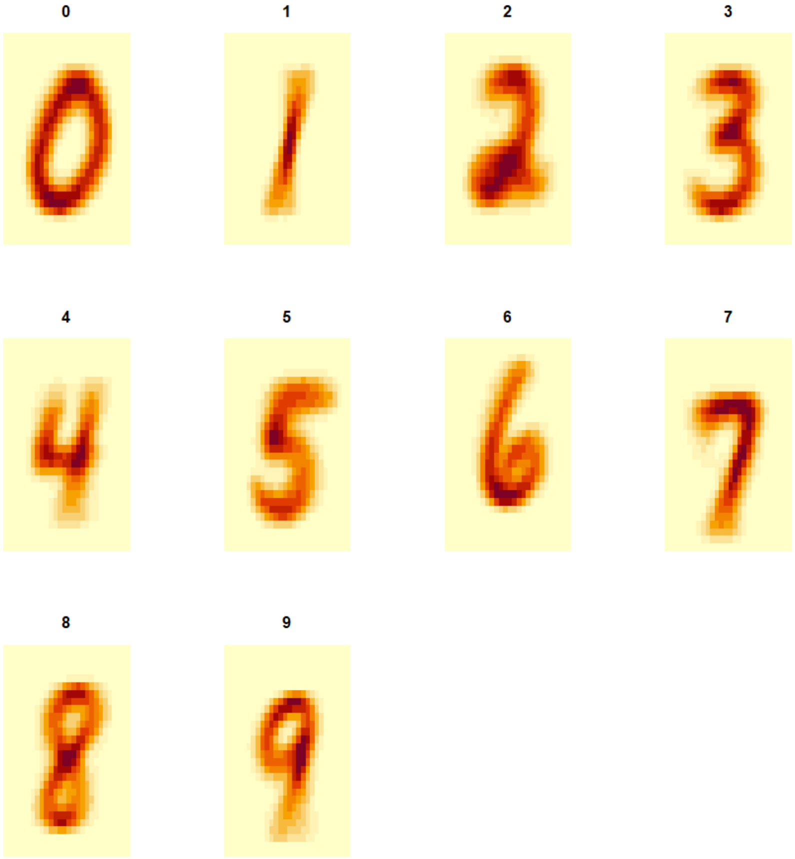

Figure 1: Plot showing the digit means.

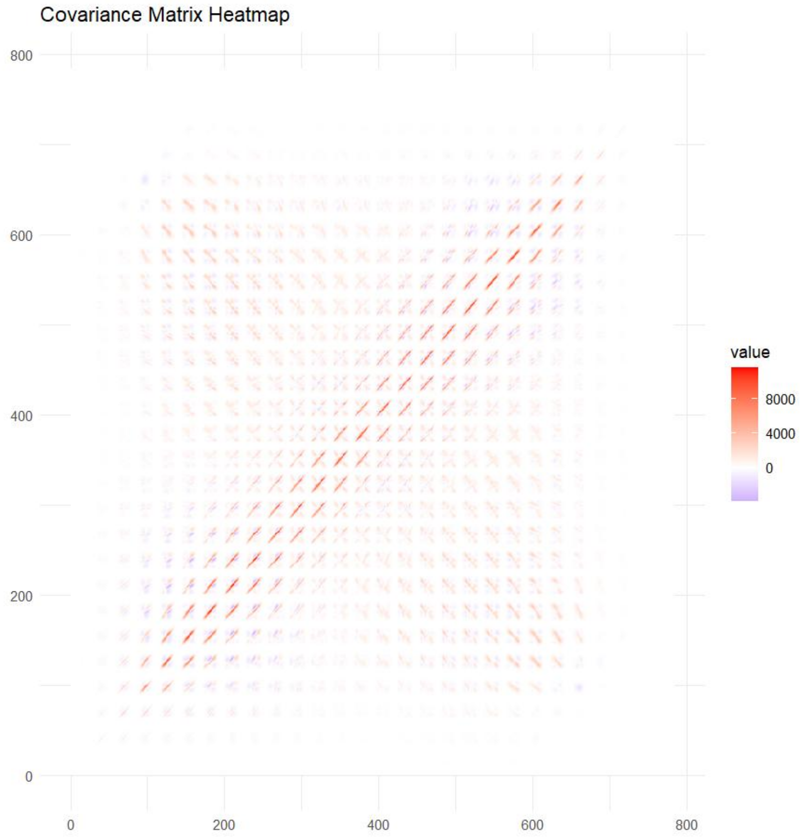

Figure 2: Plot showing pooled within-digit covariance matrix of the digits data.

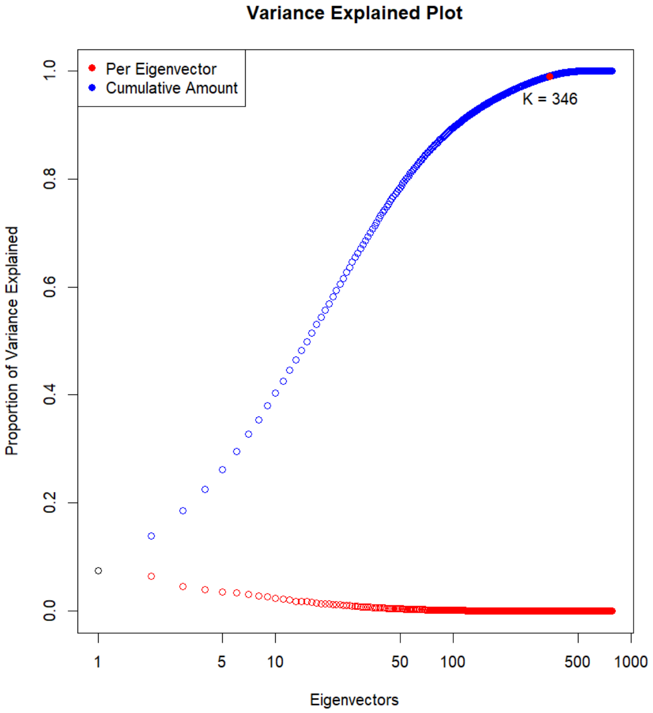

Figure 3: Plot showing both the fraction of variance explained by each eigenvector and the cumulative amount of variance explained as a function of the number of eigenvectors retained.

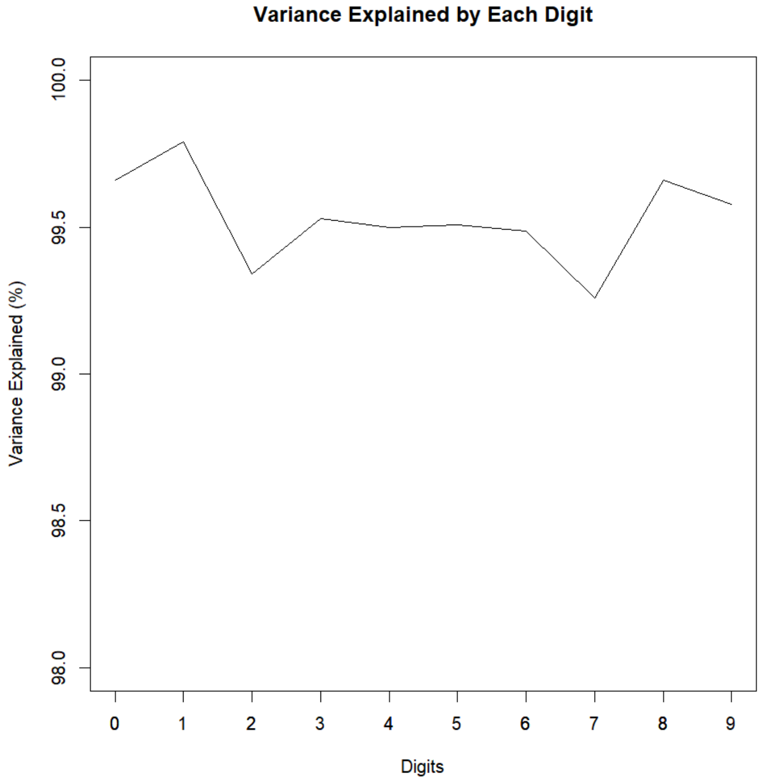

Figure 4: Plot showing the percent of variance explained as a function of the digit labels 0-9.

Question 2

(a)

This is implemented in the R script (see Rcode).

Plot showing 784 x 15 array of associated cluster means for d=2 is shown in Figure 5.

(b)

This is implemented in the R script (see Rcode).

An array of boxplots, similar to digits2.pdf, with the rss-values by digit is shown in Figure 3.

(c)

This is implemented in the R script (see Rcode).

Re-substitution misclassification percentage rate by digit added to the array of boxplots from part (b) with the rss-values by digit is shown in Figure 6.

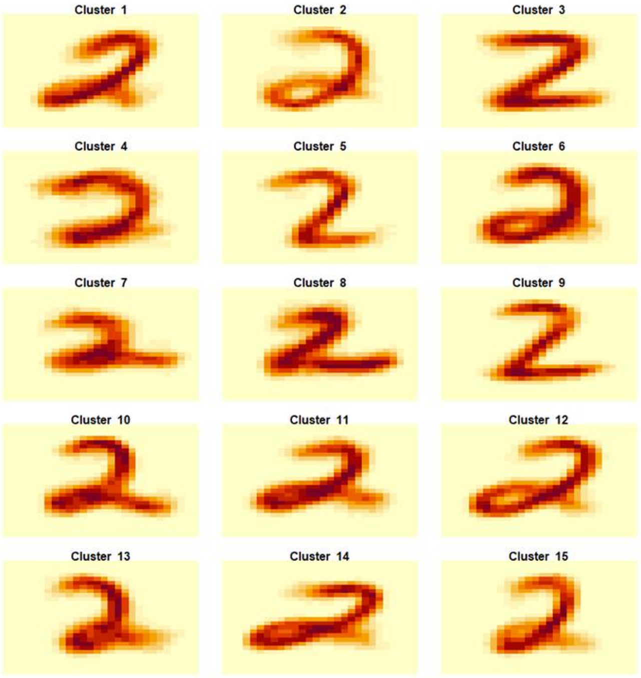

Figure 5: Plot showing 784 x 15 array of associated cluster means for d=2

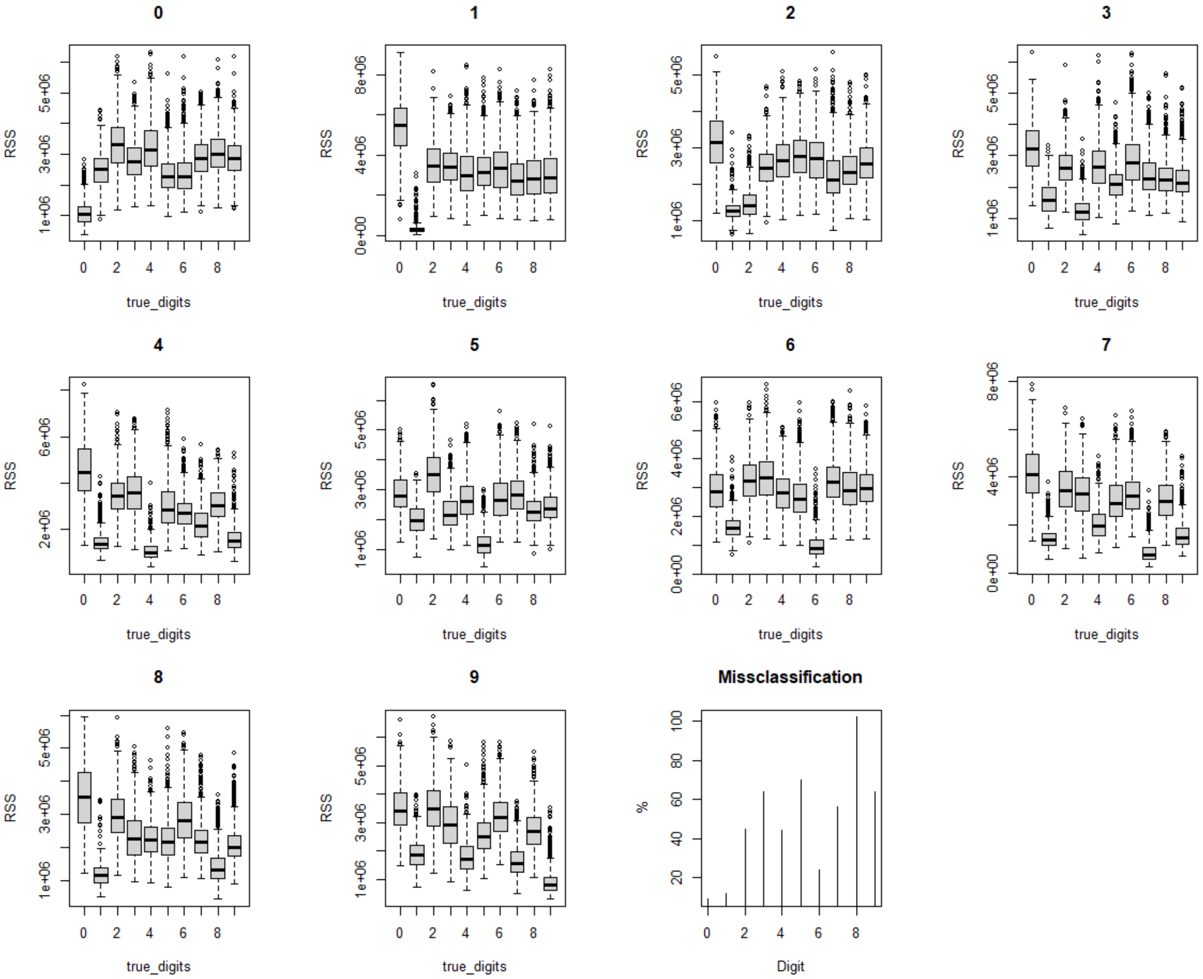

Figure 6: An array of boxplots, with the rss-values by digit & Re-substitution misclassification percentage rate by digit.

Question 3

(a)

This is implemented in the R script (see Rcode).

Report the number of cases removed from each class:

BYDra	EA	EW	Mira	RR	RSCVN	SR
217	45	75	2211	1	134	5134

(b)

(i)

This is implemented in the R script (see Rcode).

4x3 layout showing loading vectors associated with the best 4 discriminant variables, a plot showing the set of 10 F-tests values for assessment of the separation achieved by each of the discriminant variables & boxplots of the 10 discriminant variables obtained is shown in Figure 7.

(ii)

This is implemented in the R script (see Rcode).

3x3 pairwise plots of linear discriminant variables with the mean of each of the 11 object types identified with top 20 outliers removed is shown in Figure 8.

(c)

(i)

This is implemented in the R script (see Rcode).

For lda() and qda(), report well-formatted tables of re-substitution and cross-validated misclassification rates by class.

Class	**Misclassification Rates**			
	Re-substitution		**Cross-Validation**	
	LDA	**QDA**	**LDA**	**QDA**
BYDra	0.441	0.270	0.441	0.270

CEP	0.849	0.302	0.850	0.309
CEPII	0.983	0.564	0.989	0.592
DSCT	0.754	0.051	0.755	0.052
EA	0.231	0.180	0.231	0.180
EW	0.064	0.195	0.064	0.195
Mira	0.090	0.005	0.091	0.005
RR	0.179	0.040	0.180	0.040
RRC	0.887	0.083	0.887	0.084
RSCVN	0.877	0.738	0.877	0.739
SR	0.269	0.079	0.269	0.079

(ii)

Report the overall re-substitution and cross-validated misclassification rates.

Misclassification Rates			
Re-substitution		**Cross-Validation**	
LDA	**QDA**	**LDA**	**QDA**
0.2677953	0.2286397	0.2678716	0.2288065

Comment on the misclassification characteristics obtained - by type and overall:

By class:

- Mira, DSCT, and RRC show extremely low misclassification rates under QDA, indicating they are well-separated from other classes.
- CEPII, CEP, and RSCVN perform poorly under LDA but improve significantly under QDA, again highlighting the advantage of quadratic decision boundaries.
- Some classes (e.g., SR and EA) perform similarly across LDA and QDA, suggesting their boundaries may be relatively linear.
- RSCVN still shows high misclassification even under QDA (≈74%), suggesting that this class overlaps heavily with others or has higher within-class variance.

Overall, QDA consistently outperforms LDA across both re-substitution and cross-validation, as seen in the lower overall misclassification rates (22.86% vs 26.78% for cross-validation). This suggests that the class boundaries are likely non-linear, making QDA more suitable.

In summary, QDA captures the underlying class structures better for this dataset, particularly for those with more complex distributions.

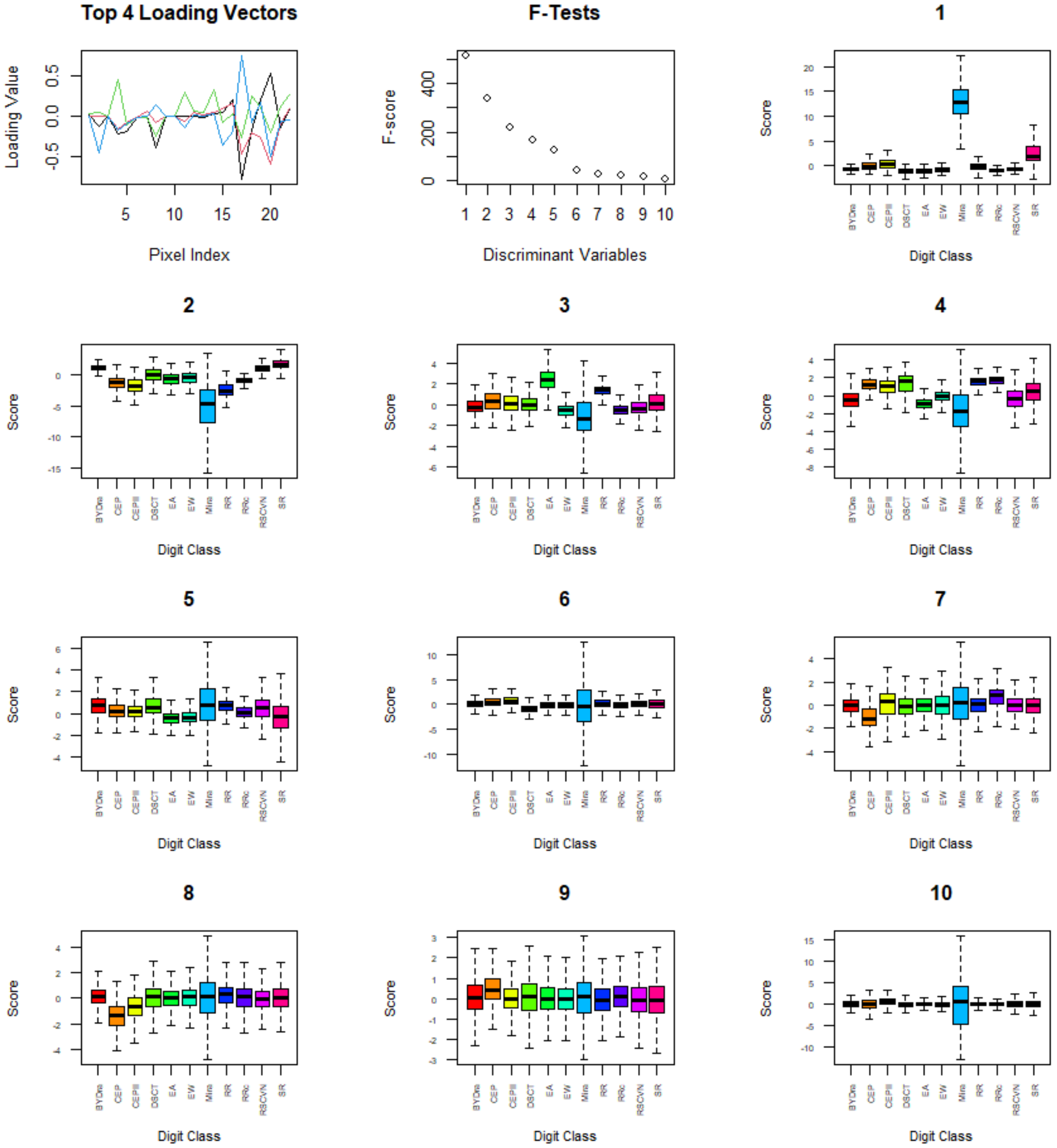

Figure 7: 4x3 layout showing loading vectors associated with the best 4 discriminant variables, a plot showing the set of 10 F-tests values for assessment of the separation achieved by each of the discriminant variables & boxplots of the 10 discriminant variables obtained

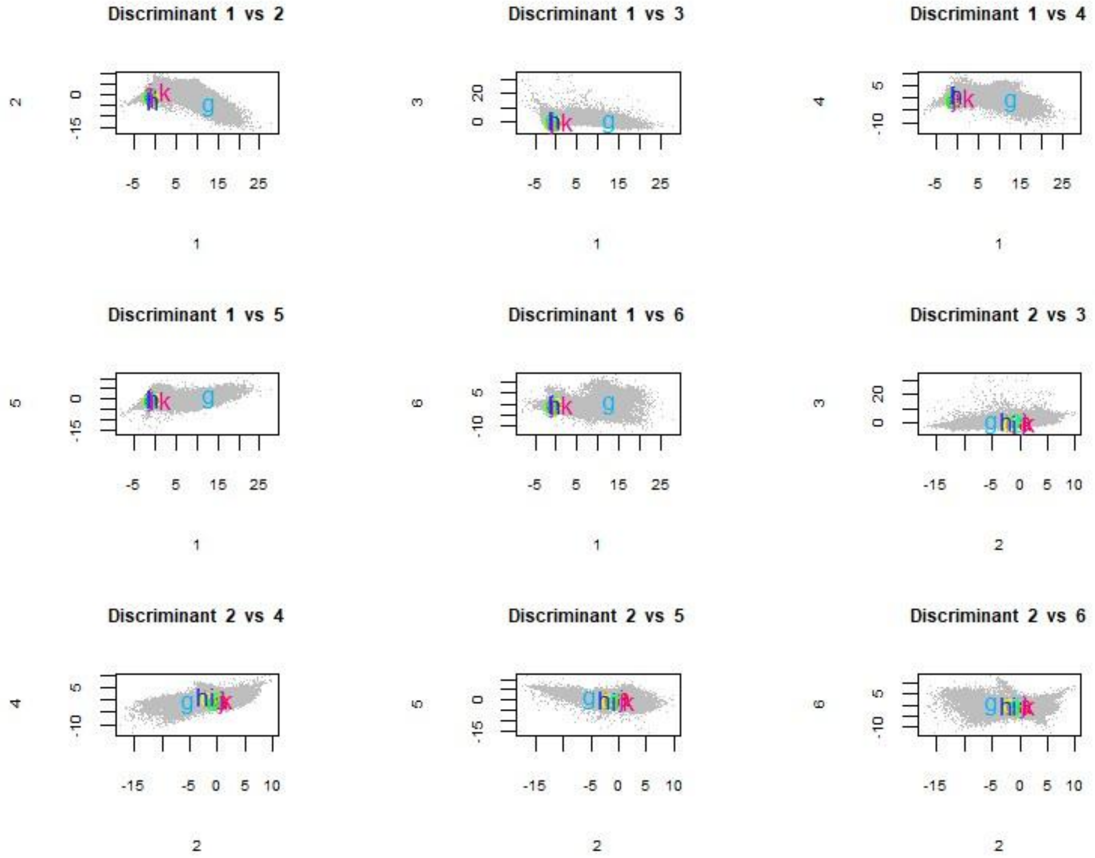

Figure 8: 3x3 pairwise plots of linear discriminant variables with the mean of each of the 11 object types identified with top 20 outliers removed.