ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Μεταπτυχιακό Πρόγραμμα Συμπληρωματικής Ειδίκευσης
Εφαρμοσμένη Στατιστική

Μάθημα: Στατιστική Μάθηση
Statistical Learning **Εργασία 1η: Classification**

Καθηγήτρια: Παπαγεωργίου Ιουλία

Ημερομηνία Παράδοσης εργασίας: 2/4/2019

Κωνσταντίνα Τσάμη (p3621817)

ΠΕΡΙΕΧΟΜΕΝΑ

Περιγραφή του Προβλήματος

Τα δεδομένα που έχουμε στη διάθεσή μας αφορούν βιβλία τεσσάρων διαφορετικών συγγραφέων και προέρχονται από το βιβλίο 'Analysing Categorical Data' του Jeffrey S. Simonoff. Το αρχείο δεδομένων περιέχει 841 παρατηρήσεις για 71 μεταβλητές. Η μεταβλητή ενδιαφέροντος είναι το όνομα του συγγραφέα 'Author' (κατηγορική μεταβλητή με 4 επίπεδα), οι υπόλοιπες 69 μεταβλητές αφορούν το πλήθος κάποιων συνδετικών λέξεων που είναι γραμμένες σε κάθε κεφάλαιο του εκάστοτε βιβλίου όπως a, the, and, or, as, also κ.α. Η τελευταία μεταβλητή αφορά τον κωδικό του βιβλίου (BookID) .

	Number of Books	Number of Chapters/Scenes
Austen	5	317
London	6	296
Milton	2	55
Shakespeare	12	173
		841

Μας ενδιαφέρει να ερευνήσουμε αν μπορούμε να κατατάξουμε τα βιβλία στο όνομα κάθε συγγραφέα στηριζόμενοι στις μεταβλητές που μας δίνουν το πλήθος των λέξεων φτιάχνοντας έναν διαχωριστικό κανόνα. Για το σκοπό αυτό θα εξετάσουμε στη συνέχεια διάφορες μεθόδους classification και θα προσπαθήσουμε να βρούμε εκείνη με την καλύτερη προβλεπτική ικανότητα.

Μέθοδος K–Nearest Neighbors

Ξεκινάμε την ανάλυσή μας με μία μη παραμετρική μέθοδο. Η K-NN δε βασίζεται σε καμία υπόθεση για την κατανομή των παρατηρήσεων και καμία πληροφορία από τις επεξηγηματικές μεταβλητές, βασίζεται αποκλειστικά στην απόσταση των παρατηρήσεων.

Περνάμε λοιπόν τα δεδομένα μας στην R και μέσω εντολών κάνουμε μία αρχική προεργασία στα δεδομένα μας.

Διαβάζουμε το αρχείο.
as←read.table("authorship.txt",header=TRUE,sep=",")
Αφαιρούμε την μεταβλητή με τον κωδικό του βιβλίου (BookID) διότι δε θα χρειαστεί στην ανάλυσή μας.
as <-as[,-match('BookID',names(as))]

Αποκτάμε μία αρχική εικόνα των δεδομένων.
head(as)
str(as)
t(t(sapply(as,class)))
apply(is.na(as)>0,2,sum)
summary(as$Author)

Στα δεδομένα μας δεν υπάρχουν missing values.

Συνεχίζουμε χωρίζοντας το σετ των δεδομένων μας σε training dataset και test dataset. Για να γίνει αυτό θα επιλέξουμε τυχαία ένα δείγμα με 200 παρατηρήσεις από το αρχικό.
set.seed(99)
x<-sample(nrow(as),replace=FALSE,size=200)

test.as<-as[x,]
train.as<-as[-x,]

Παράγουμε το πίνακα συχνοτήτων της μεταβλητής Author για να βεβαιωθούμε ότι έχουμε παρατηρήσεις και από τις 4 κατηγορίες/συγγραφείς και στα δύο σετ δεδομένων.

table(test.as$Author)
table(train.as$Author)

Πίνακας 1- Πίνακας συχνοτήτων της μεταβλητής Author σε training και test. dataset

	Training	**Test**
Austen	249	68
London	213	83
Milton	41	14
Shakespeare	138	35
#	**641**	**200**

Παρατηρούμε στον Πίνακα 1 στην κατηγορία Milton ότι οι παρατηρήσεις μας είναι λίγες σε σχέση με τον αριθμό των μεταβλητών κάτι που θα μας απασχολήσει αργότερα.

Για να τρέξουμε τον αλγόριθμο της k - NN στην R θα τυποποιήσουμε πρώτα τις μεταβλητές επίδρασης ώστε να βεβαιωθούμε ότι είναι όλες στην ίδια κλίμακα και δεν θα μεροληπτεί ο κώδικάς μας λόγω της ευκλείδια απόστασης στην οποία στηρίζεται.

```
train.X<-scale(train.as[,-match("Author",names(as))])
test.X<-scale(test.as[,-match("Author",names(as))])
```

Επίσης θα χωρίσουμε την μεταβλητή Author σε train και test ώστε να χρησιμοποιήσουμε τις train παρατηρήσεις στην εκπαίδευση του κώδικα και τις test στην εκτίμηση του σφάλματος πρόβλεψης.

```
train.Y<-train.as[,match("Author",names(as))]
test.Y<-test.as[,match("Author",names(as))]
```

Μέσω της συνάρτησης knn της R εκτελούμε τη διαδικασία.

```
library(class)
set.seed(1)
knn.pred←knn(train.X,test.X,train.Y,k=1)
```

Εξετάζουμε την διαχωριστική ικανότητα της μεθόδου k-NN για k=1

```
table(knn.pred,test.Y)
mean(test.Y==knn.pred)
mean(test.Y!=knn.pred)
```

Πίνακας 2 - Πίνακας συνάφειας μεταξύ των προβλεπόμενων τιμών της μεταβλητής Author, από τη διαδικασία k-NN για k=1 , και των πραγματικών τιμών της.

knn.pred	test.Y Austen	London	Milton	Shakespeare
Austen	**68**	4	0	0
London	0	**79**	0	0
Milton	0	0	**14**	0
Shakespeare	0	0	0	**35**

Οι διαγώνιες τιμές του Πίνακα 1 αναπαριστούν το πλήθος των τιμών που συμπίπτουν και στα δύο δείγματα, δηλαδή τις σωστά ταξινομημένες παρατηρήσεις που προέκυψαν από τον αλγόριθμό μας. Οι εκτός διαγωνίου παρατηρήσεις είναι μόλις τέσσερις. Δηλαδή 4 μόνο παρατηρήσεις απέτυχε να κατατάξει ο αλγόριθμος στη σωστή κατηγορία. Αυτό μας δίνει 4/200=2% misclassification error.

Θα δοκιμάσουμε να κάνουμε τη διαδικασία για διαφορετικές τιμές του k έως ότου να καταλήξουμε στο μικρότερο misclassification error που μπορούμε να έχουμε.

```
knn.pred<-knn(train.X,test.X,train.Y,k=2)
mean(test.Y!=knn.pred)

knn.pred<-knn(train.X,test.X,train.Y,k=3)
mean(test.Y!=knn.pred)

knn.pred<-knn(train.X,test.X,train.Y,k=4)
mean(test.Y!=knn.pred)


knn.pred<-knn(train.X,test.X,train.Y,k=5)
mean(test.Y!=knn.pred)
```

Στον Πίνακα 3 που ακολουθεί φαίνονται τα missclassification errors για κάθε k. Το ελάχιστο σφάλμα πρόβλεψης παρατηρείται για k=4 (0.005). Για k=5 το σφάλμα αυξάνεται (0.01), οπότε σταματάμε τις δοκιμές και επιλέγουμε την k-NN για **k=4** η οποία μας δίνει σφάλμα **0.5%.**

Πίνακας 3 – Τιμές του σφάλματος ταξινόμησης για διαφορετικές τιμές του k.

	k=1	k=2	k=3	k=4	k=5
Missclassification error	0.02	0.015	0.01	0.005	0.01

Γενικευμένο Γραμμικό Μοντέλο

Σε αυτή την ενότητα θα δοκιμάσουμε να προσαρμόσουμε ένα γενικευμένο γραμμικό μοντέλο πολυωνυμικής (multinomial) κατανομής (μεταβλητή απόκρισης κατηγορική μεταβλητή με 4 επίπεδα – γενίκευση διωνυμικού λογιστικού μοντέλου) και θα εξετάσουμε την απόδοσή του στη ταξινόμηση νέων παρατηρήσεων.

Προσαρμόζουμε λοιπόν το μοντέλο μας με όλες τις μεταβλητές επίδρασης στο training dataset.

```
library(nnet)
glm.as<-multinom(Author~.,data=train.as)
summary(glm.as)
plot(resid(glm.as),fitted(glm.as))
```

Εφαρμόζουμε το μοντέλο για την πρόβλεψη νέων τιμών.

```
glm.pred<-predict(glm.as,test.as,type="class")
```

Εξετάζουμε την ικανότητα πρόβλεψης.

```
table(glm.pred,test.as$Author)
mean(glm.pred!=test.as$Author)
```

Πίνακας 4 - Πίνακας συνάφειας μεταξύ των προβλεπόμενων τιμών της μεταβλητής Author, που προέκυψε από το πλήρες γενικευμένο μοντέλο , και των πραγματικών τιμών της.

glm.pred	Austen	London	Milton	Shakespeare
Austen	**68**	0	0	0
London	0	**83**	0	0
Milton	0	0	**13**	0
Shakespeare	0	0	1	**35**

Όπως παρατηρούμε στον Πίνακα 4 το πλήρες γραμμικό μοντέλο δεν ταξινόμησε σωστά μία μόνο παρατήρηση από το δείγμα, δίνοντας ένα πολύ ικανοποιητικό ποσοστό πρόβλεψης ίσο με (200-1)/200=99.5%. Το missclassification error είναι και εδώ πολύ μικρό και ίσο με **0.5%.**

Για να δούμε εάν αυτό το error μπορούμε τα το βελτιώσουμε θα δούμε αν μπορούμε να επιλέξουμε κάποιες μεταβλητές που ο συνδυασμός τους θα δίνει καλύτερη προσαρμογή και προβλεπτική ικανότητα. Ένας επιπλέον λόγος που το κάνουμε αυτό είναι ώστε να επιλέξουμε κάποιες μεταβλητές, αρκετά λιγότερες σε αριθμό από το σύνολο που διαθέτουμε, οι οποίες επιδρούν ισχυρότερα στην μεταβλητή ενδιαφέροντός μας, κάτι το οποίο θα μας βοηθήσει στην Discriminant Analysis που θα δούμε στις επόμενες ενότητες.

Λόγω του μεγάλου όγκου μεταβλητών η επιλογή είναι δύσκολη και θα καταφύγουμε στο να εφαρμόσουμε “βηματική” διαδικασία (stepwise procedure).

```
m.glm<-step(glm.as, direction = "both",data=train.as)
summary(m.glm)
formula(m.glm)
```

Το μοντέλο που προκύπτει είναι.
Author = and + her + if. + in. + it + may + must + my + of + should + such + the + to + will + e

Περιέχει 14 επεξηγηματικές μεταβλητές.

Η απόδοσή του όπως βλέπουμε στον Πίνακα 5 είναι χαμηλότερη από από αυτήν του μοντέλου που ενσωματώνει όλες τις μεταβλητές (**7.5% > 0.5%**).

Πίνακας 6 – Σφάλμα ταξινόμησης του μοντέλου με τις επιλεγμένες μεταβλητές

```
> mean(m.glm.pred!=test.as$Author)
[1] 0.075
```

Σε αυτή τη περίπτωση προτιμούμε το γενικευμένο γραμμικό μοντέλο στο σύνολο των μεταβλητών το οποίο δίνει συνολικό σφάλμα πρόβλεψης ίσο με 7.5%

Linear Discriminant Analysis

Η Linear Discriminant Analysis δίνει ελάχιστο missclassification error κάτω από συγκεκριμένες υποθέσεις. Οι υποθέσεις αυτές είναι:
Κανονική κατανομή των παρατηρήσεων σε κάθε γκρουπ .
Κοινός πίνακας συνδυακύμανσης σε όλα τα γκρούπ.

Αρχικά θα εφαρμόσουμε τον διαχωριστικό αυτό κανόνα. Στη συνέχεια θα εξετάσουμε την δυνατότητα που έχει στο να δίνει σωστές εκτιμήσεις σε νέα σετ δεδομένων και τέλος θα ελέγξουμε τις δύο υποθέσεις που προαναφέραμε.

Εφαρμόζουμε λοιπόν τη μέθοδο στα δεδομένα του training dataset.
Θα χρησιμοποιήσουμε όμως τη φόρμουλα με τις μεταβλητές που προέκυψαν από την "stepwise" procedure που είδαμε στην προηγούμενη ενότητα. Ο λόγος που το κάνουμε αυτό είναι ο μικρός αριθμός παρατηρήσεων που έχουμε στην κατηγορία Milton σε σχέση με τον αριθμό των μεταβλητών (41 παρατηρήσεις στο σύνολο και 70 μεταβλητές). Έχοντας τόσο μεγάλο αριθμό παρατηρήσεων και τόσες πολλές μεταβλητές η διακύμανση των παρατηρήσεων εντός της κατηγορίας (Σ_B) θα αυξανόταν δραματικά.

```
as.lda←lda(formula(m.glm), data = train.as)
print(as.lda)
```

Πίνακας 7 – Διαχωριστική ικανότητα των συναρτήσεων της LDA

```
Proportion of trace:
   LD1    LD2    LD3
0.5111 0.3263 0.1625
```

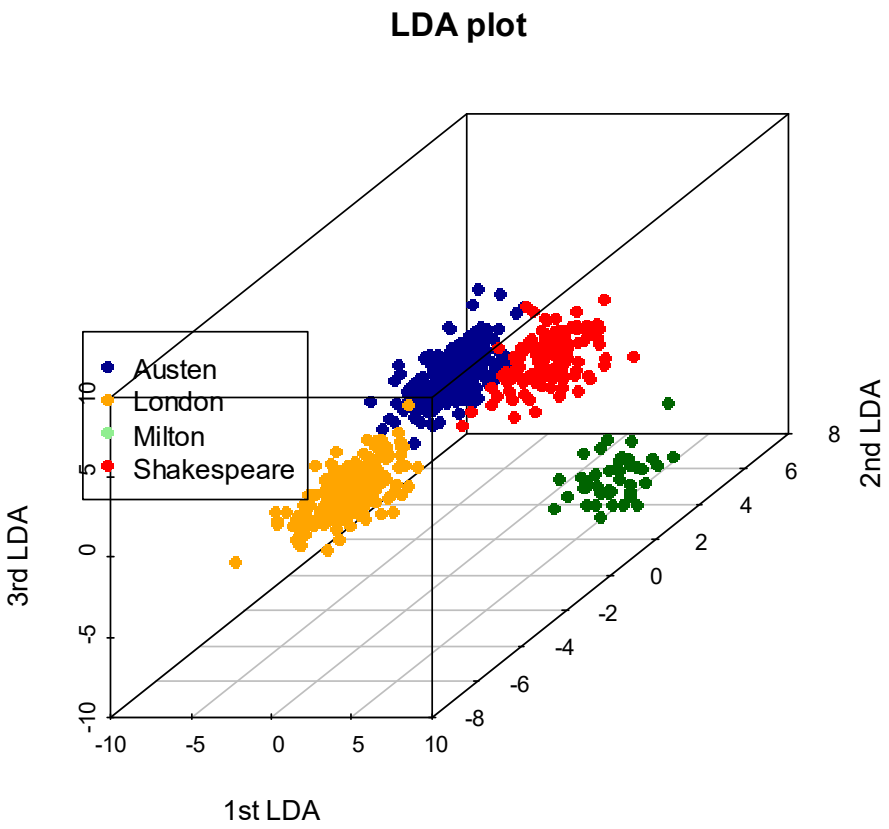
Το σημαντικό πλεονέκτημα της μεθόδου αυτής είναι η σημαντική μείωση των διαστάσεων του προβλήματος. Από 69 ανεξάρτητες μεταβλητές έχουμε μειώσει το πρόβλημα μας σε τρεις (C=4-1 κατηγορίες) ασυσχέτιστες μεταξύ τους συναρτήσεις, γραμμικό συνδυασμό των μεταβλητών. Τα αποτελέσματα που μας δίνονται από την LDA περιέχουν εκτός των άλλων τις περιθώριες κατανομές (prior probabilities) και τους συντελεστές των γραμμικών διαχωριστικών συναρτήσεων (LD1,LD2,LD3) όλων των μεταβλητών. Αυτό που μας ενδιαφέρει περισσότερο όμως είναι η διαχωριστική ικανότητα αυτών των συναρτήσεων. Σύμφωνα με τον Πίνακα 7 το ποσοστό διαχωριστικής ικανότητας της πρώτης συνάρτησης είναι 51.1%, της δεύτερης συνάρτησης 32.6%, της τρίτης συνάρτησης το υπόλοιπο 16.3%.

Την διαχωριστική ικανότητα των συναρτήσεων μπορούμε να την εξετάσουμε και γραφικά.

- Διάγραμμα σημείων – scatter plot 3D.

```
library(scatterplot3d)

cols <- c("darkblue", "orange", "darkgreen","red")

with(train.as, scatterplot3d(predict(as.lda)$x[,1],

                predict(as.lda)$x[,2],

                predict(as.lda)$x[,3],

                main="LDA plot",

                xlab = "1st LDA",

                ylab = "2nd LDA",

                zlab = "3rd LDA",

                pch = 16, color=cols[as.numeric(train.as$Author)]))

legend("left", legend = levels(train.as$Author),

col =  c("darkblue", "orange", "lightgreen","red"), pch = 16)
```

Διάγραμμα 1- Προβολές των σημείων/παρατηρήσεων στις ευθείες των γραμμικών συναρτήσεων LDA1,LDA2,LDA3

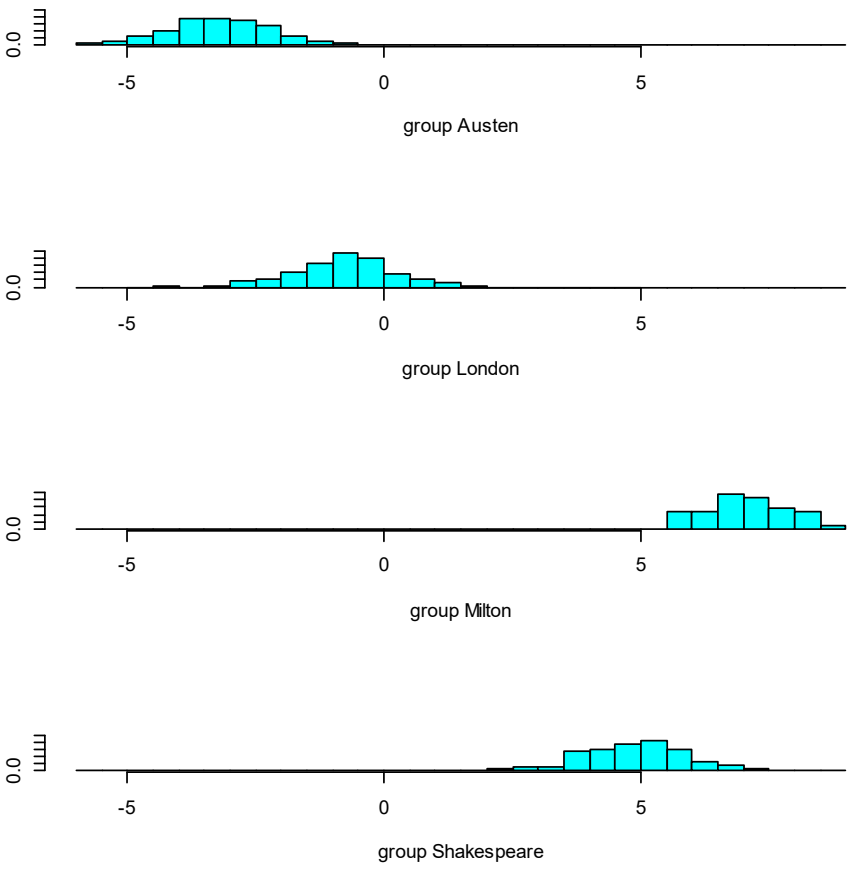


Από το Διάγραμμα 1 το οποίο αναπαριστά τις προβολές των σημείων στις τρεις ευθείες των διαχωριστικών συναρτήσεων βλέπουμε έναν αρκετά ικανοποιητικό διαχωρισμό. Οι παρατηρήσεις του κάθε group φαίνονται να είναι οι περισσότερες σε απόσταση με τις παρατηρήσεις που ανήκουν σε άλλα groups, ιδιαίτερα της κατηγορίας του συγγραφέα Milton που απεικονίζονται με πράσινο χρώμα και λίγες φαίνονται να εμπλέκονται μεταξύ δύο groups.

- Ιστογράμματα για κάθε συνάρτηση ξεχωριστά.

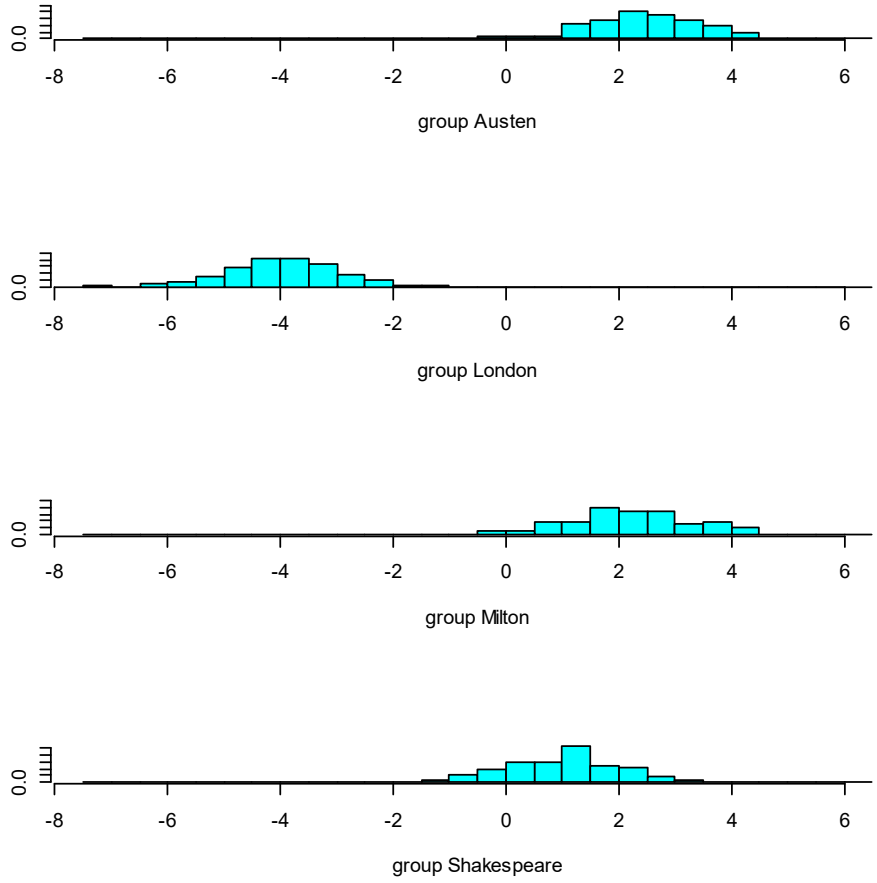
```
ldahist(data = predict(as.lda)$x[,1], g=train.as$Author , main="LD1")

ldahist(data = predict(as.lda)$x[,2], g=train.as$Author , main="LD2")

ldahist(data = predict(as.lda)$x[,3], g=train.as$Author , main="LD3")
```

Διάγραμμα 2 - Ιστόγραμμα συχνοτήτων των τιμών της LD1 σε κάθε group



Οι παρατηρήσεις στο Διάγραμμα 2 φαίνονται επαρκώς διαχωρισμένες με μία μικρή επικάλυψη στα groups Milton και Shakespeare όπως επίσης και στα groups Austen και London. Οι διάμεσοι ωστόσο των παρατηρήσεων φαίνεται να διαφέρουν σημαντικά μεταξύ τους.

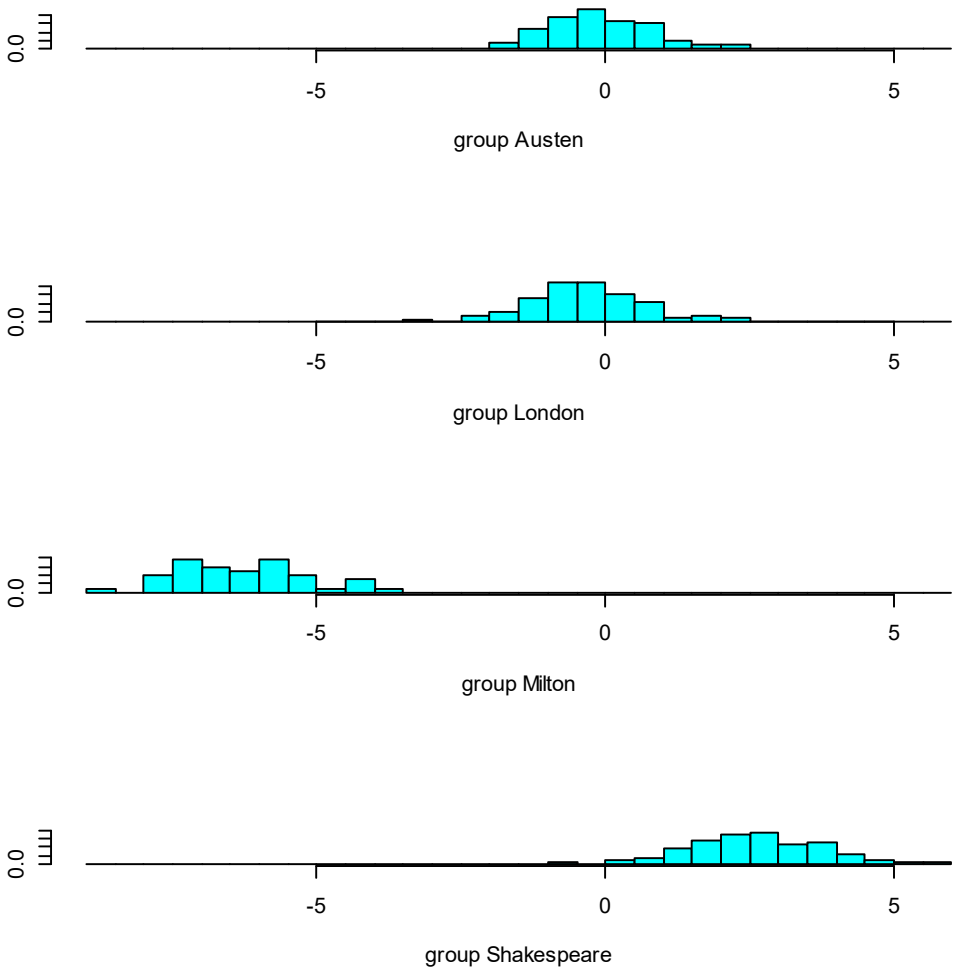
Διάγραμμα 3 - Ιστόγραμμα συχνοτήτων των τιμών της LD2 σε κάθε group



Στο Διάγραμμα 3 οι τιμές της δεύτερης διαχωριστικής συνάρτησης (LD2) φαίνονται να έχουν μια μεγαλύτερη επικάλυψη στα groups Austen, Milton και London.

Στο Διάγραμμα 4 παρακάτω, παρατηρείται μεγάλη επικάλυψη στα groups Austen και London.

Διάγραμμα 4- Ιστόγραμμα συχνοτήτων των τιμών της LD3 σε κάθε group



Σειρά έχει τώρα να εξετάσουμε το misclassification error της LDA.

Κάνουμε εκτιμήσεις μέσω του test dataset.

```
lda.pred<-predict(as.lda,test.as)

table(lda.pred$class,test.as$Author)

mean(lda.pred$class!=test.as$Author)
```

Πίνακας 8 – Πίνακας συνάφειας μεταξύ των προβλεπόμενων τιμών της μεταβλητής Author, και των πραγματικών τιμών της για τη μέθοδο LDA.

	Austen	London	Milton	Shakespeare
Austen	66	2	0	3
London	0	81	0	2
Milton	1	0	14	1
Shakespeare	1	0	0	29

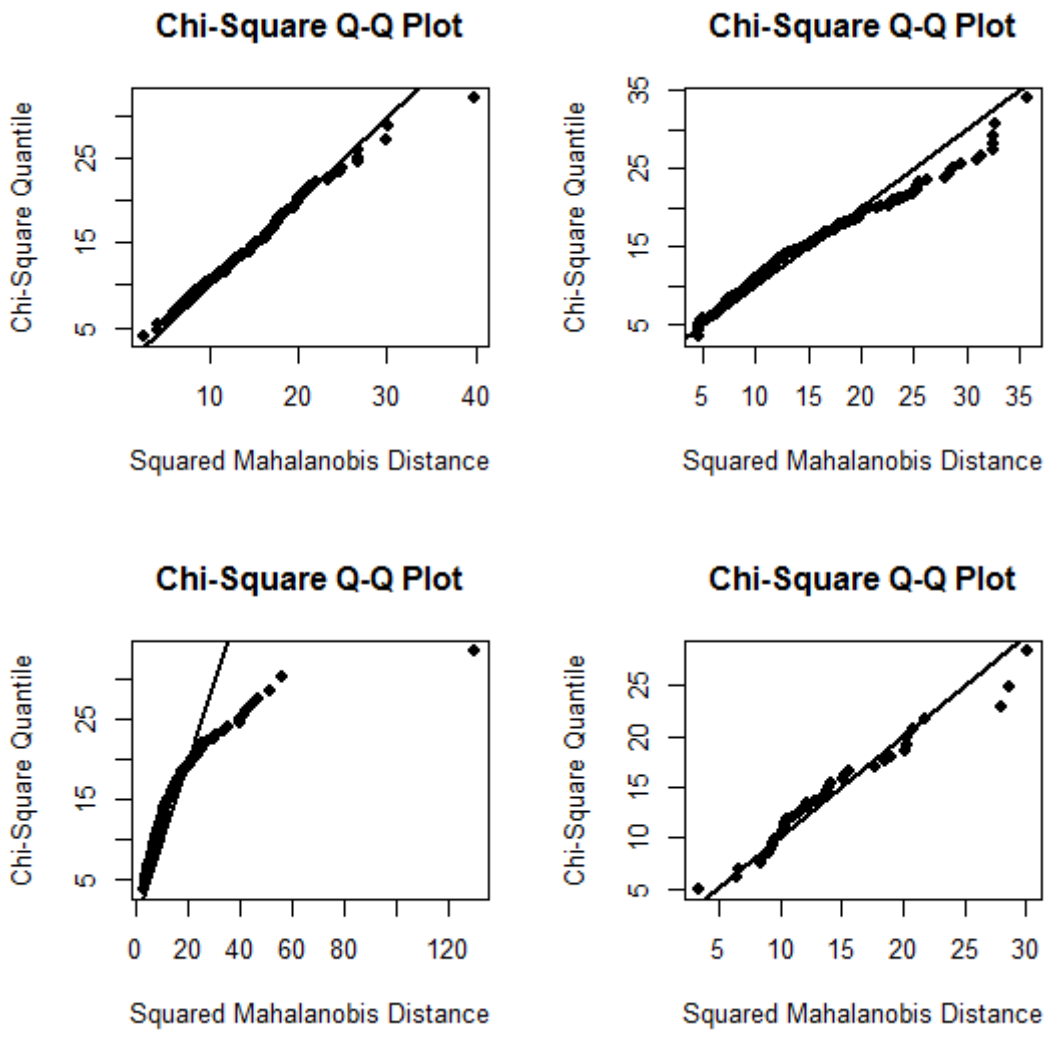
Σύμφωνα με τον Πίνακα 8, 10 παρατηρήσεις συμολικά δεν έχουν ταξινομηθεί σωστά δίνοντας έστι στην LDA μέθοδο σφάλμα πρόβλεψης **0.5%**.

Εν συνεχεία, θα δούμε όπως είπαμε αν καλύπτονται οι υποθέσεις της μεθόδου. Αυτό θα το κάνουμε χρησιμοποιώντας κάποιους ελέγχους για πολυμεταβλητά δεδομένα. Ξεκινώντας από τον έλεγχο κανονικότητας “Mardia’s Multivariate Normality Test” και τα διαγράμματα των “Chi--squared q-q plot” (Διάγραμμα 5).

```
#Mardia's multivariate normality test kai chi-sq qq plot

library(MVN)

par(mfrow=c(2,2))

x1<-train.as[train.as$Author=='Shakespeare',]

x2<-x1[,match(c('and','her','if.','in.', 'it', 'may',
'must','my','of','should','such','the','to','will'),names(as))]

mvn(x2,mvnTest="mardia",multivariatePlot="qq",showOutliers=TRUE)


x1<-train.as[train.as$Author=='Austen',]

x2<-x1[,match(c('and','her','if.','in.', 'it', 'may',
'must','my','of','should','such','the','to','will'),names(as))]

mvn(x2,mvnTest="mardia",multivariatePlot="qq",showOutliers=TRUE)
```

```
x1<-train.as[train.as$Author=='London',]

x2<-x1[,match(c('and','her','if.','in.', 'it', 'may',
'must','my','of','should','such','the','to','will'),names(as))]

mvn(x2,mvnTest="mardia",multivariatePlot="qq",showOutliers=TRUE)


x1<-train.as[train.as$Author=='Milton',]

x2<-x1[,match(c('and','her','if.','in.', 'it', 'may',
'must','my','of','should','such','the','to','will'),names(as))]

mvn(x2,mvnTest="mardia",multivariatePlot="qq",showOutliers=TRUE)
```

Διάγραμμα 5- Chi-squared q-q plots των παρατηρήσεων κάθε κατηγορίας



Αν και προσεγγιστικά θα μπορούσαμε να δεχτούμε ότι υπάρχει κανονικότητα λόγω ενός μεγάλου σχετικά μεγέθους δείγματος σε κάθε κατηγορία, στο Διάγραμμα 5 αυτή η συνθήκη δε φαίνεται να ικανοποιείται με βεβαιότητα καθώς αρκετά σημεία της δειγματικής πολυμεταβλητής κατανομής εμφανίζουν απόκλιση από τα ποσοστιαία σημεία της κανονικής κατανομής κυρίως στο group “London” (διάγραμμα κάτω αριστερά) και στο group “Austen” (διάγραμμα πάνω δεξιά).

Σειρά έχει ο έλεγχος ίσότητας των πινάκων συνδυακύμανσης των groups. Θα χρησιμοποιήσουμε τον έλεγχο “ Box-M” για τον ελεγχο ομοιογένειας των πινάκων συνδυακύμανσης.

```
library(heplots)

x2<-train.as[,match(c('and','her','if.','in.', 'it', 'may',
'must','my','of','should','such','the','to','will'),names(as))]

boxM(x2,train.as$Author)
```

Πίνακας 9 – ‘Ελεγχος ισότητας διακυμάνσεων για πολυμεταβλητά δεδομένα “Box-M”

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  x2
Chi-Sq (approx.) = 1481.4, df = 315, p-value < 2.2e-16
```

Σύμφωνα με τον έλεγχο η μηδενική υπόθεση της ισότητας των πινάκων διακυμάνσεων απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας 5% (Πίνακας 9, p-value<2.2e-16). Οι πίνακες συνδυακυμάνσεων διαφέρουν σημαντικά μεταξύ τους με πιθανότητα μεγαλυτερη από 95%.

Σε αυτό το σημείο θα ήταν χρήσιμο να αναφέρουμε ότι οι υποθέσεις αυτές που ελέγξαμε προηγουμένως δεν αποτελούν αναγκαία συνθήκη της LDA αλλά απλά αποτελούν ιδανική συνθήκη κάτω από την οποία η LDA έχει την καλύτερη δυνατότητα σωστής πρόβλεψης.

Quadratic Discriminant Analysis

Η Quadratic Discriminant Analysis εφαρμόζεται στην περίπτωση που δεν έχουμε κοινό πίνακα συνδυακυμάνσεων σε όλα τα groups. Επειδή η συγκεκριμένη μέθοδος υπολογίζει τον πίνακα συνδυακύμανσης κάθε group, δε θα μπορούσε να λειτουργήσει αν δεν είχαμε επιλέξει τις 14 μεταβλητές.

```
as.qda<-qda(formula(m.glm), data = train.as)
```

Για να δούμε τη διαχωριστική ικανότητα της QDA θα εφαρμόσουμε MANOVA (anova για πολυμεταβλητά δεδομένα) και 'Wilks' έλεγχο.

```
author.type=as.double(train.as$Author)
author.var<-as.matrix(train.as[,match(c('and','her','if.','in.', 'it',
                      'may', 'must','my','of','should','such','the','to','will'),names(as))])

manova(author.var~ author.type)
summary(manova(author.var~ author.type), test='Wilks')
```

Πίνακας 10 – 'Ελεγχος "Wilks" για την ισότητα των μέσων κάθε group

```
              Df  Wilks  approx F num Df   den Df  Pr(>F)
author.type    1 0.26206   125.91     14    626 < 2.2e-16 ***
Residuals    639
 ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

'Οπως βλέπουμε στον Πίνακα 10 υπάρχουν στατιστικά σημαντικές διαφορές ανάμεσα στους διανυσματικούς μέσους των groups (Wilks p-value $<2.2e^{-16}$,0.05) σε επίπεδο σημαντικότητας 5%. Επίσης η τιμή του Λ είναι ίση με **0.26,** δηλαδή το ποσοστό μεταβλητότητας της εξαρτημένης μεταβλητής που δεν εξηγείται από την μεταβλητότητα τον groups των επεξηγηματικών μεταβλητών. Θα λέγαμε λοιπόν ότι έχουμε μία καλή διαχωριστική ικανότητα.

Στο σημείο αυτό θα εξετάσουμε το σφάλμα ταξινόμησης της QDA στο σετ δεδομένων test.

```
qda.pred<-predict(as.qda,test.as)
table(qda.pred$class,test.as$Author)
mean(qda.pred$class!=test.as$Author)
```

Πίνακας 11 – Πίνακας συνάφειας μεταξύ των προβλεπόμενων τιμών της μεταβλητής Author, και των πραγματικών τιμών της για τη μέθοδο QDA.

	Austen	London	Milton	Shakespeare
Austen	67	2	1	5
London	0	80	1	1
Milton	1	0	12	0
Shakespeare	0	1	0	29

Δώδεκα παρατηρήσεις δεν έχουν ταξινομηθεί σωστά όπως μπορούμε να παρατηρήσουμε στον Πίνακα 11, το συνολικό missclassification error της μεθόδου είναι **6%.**

Συμπεράσματα

Ανακεφαλαιώνοντας, στον Πίνακα 12 συνοψίζονται όλες οι μέθοδοι classification που έχουμε δοκιμάσει και τα σφάλματα ταξινόμησης που δίνουν. Τα μεγαλύτερα σφάλματα πρόβλεψης έχουν οι μέθοδοι LDA και QDA και αυτό συμβαίνει κατά πάσα πιθανότητα διότι όπως είδαμε δεν βρίσκονται στις ιδανικές συνθήκες τους. Οι μέθοδοι k-NN και GLM (Γενικευμένο Γραμμικό Μοντέλο) έχουν την ισχυρότερη ικανότητα πρόβλεψης έχοντας και οι δύο missclassification error ίσο με **0.5%.**

Πίνακας 12 – Συνολικό Σφάλμα ταξινόμησης των μεθόδων

Μέθοδος	**k-NN**	**GLM**	**DA**	**QDA**
Missclassification error	0.005	0.005	0.05	0.06

Ανάμεσα στις δύο μεθόδους θα λέγαμε ότι υπερτερεί η μέθοδος k-NN καθώς δε χρειάζεται καμία υπόθεση, δε βασίζεται σε καμία πληροφορία από τις ανεξάρτητες μεταβλητές και είναι πιο εύκολη στο σχεδιασμό και τη χρήση της. Το γενικευμένο γραμμικό μοντέλο (πολλαπλό μοντέλο λογιστικής παλινδρόμησης) από την άλλη απαιτεί ελέγχους όπως ελέγχο ανεξαρτησίας καταλοίπων και καλης προσαρμογής (goodness of fit) που είναι δύσκολο να γίνουν και απαιτείται η δημιουργία επιμέρους απλών λογιστικών μοντέλων. Επίσης όπως είδαμε στην κατηγορία “Milton” ο αριθμός των παρατηρήσεων είναι αρκετά μικρός σε σχέση με τον αριθμό παραμέτρων οπότε υπάρχει ο κίνδυνος υπερμετροποίησης (overfitting) δίνοντας ασταθείς εκτιμήσεις.

Σε ασφαλέστερα συμπεράσματα θα οδηγούμασταν αν είχαμε στη διάθεσή μας μεγαλύτερο μέγεθος δείγματος, εξετάζοντας και την πρόβλεψη νέων τιμών με μεγαλύτερο σετ δεδομένων (test dataset). Μία ακόμη προσέγγιση θα ήταν να εφαρμόσουμε κάποια Bootstraping μέθοδο ή Cross-Validation έλεγχωντας το σφάλμα σε μεγαλύτερο αριθμό δειγμάτων. Όπως και να έχει, η καλύτερη επιλογή επίλογή που έχουμε με τα δεδομένα που διαθέτουμε φαίνεται να είναι η k-NN.