Figure 1: TUC

ΑΝΑΠΤΥΞΗ ΧΗΜΕΙΟΜΕΤΡΙΚΟΥ ΑΛΓΟΡΙΘΜΟΥ ΠΡΟΒΛΕΨΗΣ ΙΔΙΟΤΗΤΩΝ ΠΕΤΡΕΛΑΙΟΕΙΔΩΝ ΑΠΟ ΔΕΔΟΜΕΝΑ ΥΠΕΡΥΘΡΗΣ ΦΑΣΜΑΤΟΣΚΟΠΙΑΣ

ΟΝΟΜΑ: ΑΝΤΩΝΙΟΣ - ΣΤΑΥΡΟΣ ΤΡΙΑΝΤΟΣ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

ΝΙΚΟΛΑΟΣ ΠΑΣΑΔΑΚΗΣ (ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ)

ΝΙΚΟΛΑΟΣ ΒΑΡΟΤΣΗΣ (ΚΑΘΗΓΗΤΗΣ) , ΒΑΣΙΛΕΙΟΣ ΓΑΓΑΝΗΣ (ΕΔΙΠ)

Το παρόν έγγραφο αποτελεί προϋπόθεση για την απόκτηση του διπλώματος Μηχανικού Ορυκτών Πόρων

Χανιά 07-Ιούνιος 2017

Contents

0.1 Χρήσιμοι συμβολισμοί . . . 3

1 Εισαγωγή **4**
1.1 Λεπτομέρειες για την εργασία . . . 4
1.2 Περίληψη . . . 4
1.3 Ευχαριστίες . . . 5
1.4 Χημειομετρία . . . 5

2 ΠΑΛΙΝΔΡΟΜΗΣΗ **7**
2.1 Θεωρία παλινδρόμησης . . . 7
2.2 Μαθηματική προσέγγιση Παλινδρόμησης . . . 7
2.3 Εφαρμογή στην χημειομετρία . . . 9
2.4 Χρήσεις . . . 9

3 ΜΕΘΟΔΟΙ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ **10**
3.1 MLR (Πολυμεταβλητή Γραμμική Παλινδόμηση) . . . 10
3.2 PLS (Μερικά Ελάχιστα Τετράγωνα) . . . 11
3.3 iPLS (Με διαστήματα) . . . 12
3.4 Μείωση διαστάσεων . . . 13
3.5 PCR (Παλινδρόμηση Κυρίων Συνιστωσών) . . . 14
3.6 SVD (Ανάλυση πίνακα σε ιδιάζουσες τιμές) . . . 16

4 ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ **18**
4.1 Παλινδρόμηση με επιλογή μεταβλητών . . . 18
4.2 Επιλογή λανθάνουσων μεταβλητών . . . 19
4.3 Παλινδρόμηση με προσαρμογή ποινής . . . 19

5 ΔΕΙΓΜΑΤΑ ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ ΜΕΤΡΗΣΗΣ **21**
5.1 Παρουσίαση Ιδιοτήτων δειγμάτων . . . 21
5.2 Ιδιότητες diesel . . . 23
5.3 Κετάνιο - Αριθμός Κετανίου . . . 23
5.4 Κινηματικό Ιξώδες . . . 24
5.5 Αριθμός Οκτανίου . . . 24
5.6 Φασματοσκοπία IR-NIR . . . 25

6 ΑΝΑΠΤΥΞΗ ΑΛΓΟΡΙΘΜΟΥ ΠΡΟΒΛΕΨΗΣ **27**
6.1 Εισαγωγη δεδομένων . . . 27
6.2 Διάγραμμα Οπτής Διαπερτότητας - Επιλογή Μεταβλητών . . . 28
6.3 Μετατροπή δεδομένων Οπτής Διαπερατότητας σε Απορρόφησης . . . 31
6.4 Επεξεργασία δεδομένων - Data pretreatment . . . 32
6.5 Μεθοδος PCR- Διάγραμμα PCA . . . 37
6.6 Cross Validation . . . 39
6.7 Διακύμανση εκφρασμένη με τον αριθμό των Κυρίων Συνιστωσών, για μεθοδολογίες PCA - PLS . . . 43
6.8 Επιλογή κατάληλλου αριθμού κύριων συνιστωσών για μείωση σφάλματος . . . 44
6.9 Διάγραμμα Βαρών PLS - PCA . . . 48
6.10 Μοντέλα Παλινδρόμησης με PLS . . . 49
6.11 Μοντέλα Παλινδρόμησης με PCR . . . 53

7 ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ **58**
7.1 Στοιχεία για την πρόβλεψη Ιδιοτήτων αγνώστου δείγματος . . . 60
7.2 Προσαρμογή περισσοτέρων Κυρίων Συνιστωσών στο Μοντέλο εκπαίδευσης . . . 61
7.3 Μελέτη σφαλμάτων - Αριθμού Κετανίου . . . 63
7.4 Ακρίβεια υπολογισμών - Σύγκριση με αποτελέσματα άλλων ερευνών . . . 67
7.5 Μοντέλα πρόβλεψης Ιξώδους . . . 67
7.6 Μοντέλα πρόβλεψης Αριθμού Οκτανίων . . . 75
7.7 Προτάσεις για βελτίωση μεθοδοδολογιών για καλύτερα αποτελέσματα . . . 79

0.1 Χρήσιμοι συμβολισμοί

- **ALL** : χρήση όλου του εύρους των κυματάριθμων.
- **1500-2500** : χρήση 1500-2500 του μέρους των κυματάριθμων.
- **2500-2500** : χρήση 2500-2500 του μέρους των κυματάριθμων.
- **St ή Stand** : Χρήση Κανονικοποιημένων Δεδομένων.
- **Sc** : Κλιμακωμένα Δεδομένα.
- **31** : Χρήση των 31 δειγμάτων.
- **29** : χρήση μόνο των 29 δειγμάτων (αφαίρεση των 2 προβληματικών τιμών).
- **PCs** : Κύριες Συνιστώσες.
- **CN** : Αριθμός Κετανίου.
- **Training Set** : Αρχικά δεδομένα για το μοντέλο πρόβλεψης.
- **rsquared** : Μέσο τετραγωνικό σφάλμα.
- **PCR** : Παλινδρόμηση Κυρίων Συνιστωσών.
- **PLS** : Μερικά Ελάχιστα Τετράγωνα.
- **BetaPCRn** : Διάνυσμα παλινδρόμησης για (n) Κύριες Συνιστώσες.
- **PCALoadings** : Διανύσματα Βαρών για προσαρμογή δεδομένων στον χώρο των Κυρίων Συνιστωσών.

Chapter 1

Εισαγωγή

1.1 Λεπτομέρειες για την εργασία

Η παρούσα εργασία εκπονήθηκε από τον φοιτητή Αντώνιο-Σταύρο Τριάντο, το χρονικό διάστημα του 10ου εξαμήνου φοίτησης στην σχολή των Μηχανικών Ορυκτών Πόρων του Πολυτεχνείο Κρήτης.

Η τριμελής επιτροπή αποτελείται απο τον υπεύθυνο και επιβλέποντα καθηγητή Πασαδάκη Νικόλαο και στην εξεταστική επιτροπή βρίσκονται ο Καθηγητής Νικόλαος Βαρότσης και ο Βασίλειος Γαγάνης (Εργαστηριακό Διδακτικό Προσωπικό).

Κατά την εκπόνηση της εργασίας, έγινε χρήση μεθόδων πολυμεταβλητής στατιστικής ανάλυσης δεδομένων (Multivariate Linear Regression), με το λογισμικό MATLAB R2014b. Επίσης, έγινε χρήση λογισμικού excel για την εισαγωγή των δεδομένων, αλλά και την παρουσίαση των αποτελεσμάτων. Η εργασία τυπώθηκε με χρήση του λογισμικού της LaTeX.

1.2 Περίληψη

Στην παρούσα εργασία δημιουργήθηκαν αλγόριθμοι για την πρόβλεψη χημικών ιδιοτήτων πετρελαϊκών κλασμάτων από δεδομένα απορρόφησης υπέρυθρης φασματοσκοπίας στο εγγύς-NIR και στο μέσο –IR. Συγκεκριμένα αναπτύχθηκαν μαθηματικά μοντέλα πρόβλεψης για το Κινηματικό Ιξώδες (15^oC) και τον Αριθμό Κετανίου για κλάσματα τύπου Ντίζελ, αλλά και τον Αριθμό Οκτανίων για βενζίνες.

Η χρήση ενός τέτοιου αλγορίθμου είναι ωφέλιμη καθώς μπορεί να αντικαταστήσει τις εργαστηριακές διαδικασίες μέτρησης αυτών των ιδιοτήτων, οι οποίες μπορεί να είναι χρονοβόρες, ή πολύ ακριβές ή και δύσκολες στο να πραγματοποιηθούν, σε αντίθεση με τη συλλογή δεδομένων υπέρυθρης φασματοσκοπίας, οπου είναι εύκολο να παραχθούν και να αξιολογιθούν.

Ο αλγόριθμος αναπτύχθηκε με τις μεθοδολογίες PLS (Μερικά ελάχιστα τετράγωνα) και PCR (Παλινδρόμηση Κυρίων Συνιστωσών) στο προγραμματιστικό περιβάλλον Matlab.

Για την πρόβλεψη του Αριθμού Κετανίου χρησιμοποιήθηκαν 31 δείγματα όπου με την μεθοδολογία

PLS, και έδωσαν ένα μέσο τετραγωνικό σφάλμα (RMSEP) = 0.3179.Το μοντέλο πρόβλεψης πρόβλεψης παρουσιάζει απόκλιση απο την πραγματική = ±0.2475%.

Από το ίδιο σετ φασματοσκοπικών δεδομένων για την πρόβλεψη του Κινηματικού Ιξώδους, έγινε χρήση 21 δειγμάτων όπου με την μεθοδολογία PLS παρουσιάζει μέσο τετραγωνικό σφάλμα (RMSEP) = 0.0319. Το μοντέλο πρόβλεψης παρουσιάζει απόκλιση απο την πραγματική = ±0.2425%.

Τέλος με ένα διαφορετικό σετ δεδομένων δημιουργήθηκε το μοντέλο πρόβλεψης Αριθμού Οκτανίων (RON) με χρήση 44 δειγμάτων βενζινών, Το οποίο με την μεθοδολογία PLS, παρουσιάζει ένα μέσο τετραγωνικό σφάλμα(RMSEP) = 0.0034. Το μοντέλο πρόβλεψης παρουσιάζει απόκλιση απο την πραγματική = ±0.2425%.

1.3 Ευχαριστίες

Ένα μεγάλο ευχαριστώ οφείλω να δώσω στον καθηγητή και υπεύθυνο της παρούσας εργασίας καθηγητή Πασαδάκη Νικόλαο, για την απεριόριστη εμπιστοσύνη και ενδιαφέρον, που έδειξε στο πρόσωπό μου, αλλά και τις γνώσεις που μου μετέδωσε μέσα από όλον αυτόν τον καιρό της συνεργασίας μας. Επίσης ευχαριστώ πολύ την εξεταστική επιτροπή της παρούσας εργασίας, τον καθηγητή Βαρότση Νικόλαο και τον Βασίλειο Γαγάνη (ΕΔΙΠ) για την υπέροχη συνεργασία τους.

Ένα ευχαριστώ θα ήταν λίγο για να περιγράψω την ευγνωμοσύνη που έχω στην οικογένεια μου, όλα τα χρόνια της εκπαιδευτικής μου πορείας. Φυσικά δε θα πρέπει να παραλείψω τους αγαπημένους φίλους μου, όπου υπήρξαν μια θαυμάσια επιρροή για εμένα, και στους οποίους εύχομαι τα καλύτερα στην ζωή τους.

1.4 Χημειομετρία

Η χημειομετρία είναι ο διεπιστημονικός κλάδος της επιστήμης, ο οποίος ασχολείται με την επεξεργασία δεδομένων, για την παραγωγή πληροφοριών ενός χημικού συστήματος. Ο κόσμος της χημειομετρίας εμφανίστηκε από την παρέμβαση των μαθηματικών στο περιβάλλον της χημείας.

Στην μαθηματική αυτή αντιμετώπιση των ερωτημάτων που παρατηρούνται, καλούνται συγκεκριμένοι κλάδοι των μαθηματικών, όπως της γραμμικής άλγεβρας, των εφαρμοσμένων μαθηματικών, της στατιστικής και της επιστήμης των υπολογιστών να συνδυαστούν και να συμπράξουν στην εξαγωγή πληροφορίας. Οι μαθηματικές αυτές τεχνικές μπορούν να αποτελέσουν κλειδί για πολύ σημαντικά συμπεράσματα, στο θέμα της παρούσας εργασίας, και γενικότερα και σε περιγραφικά μοντέλα και μοντέλα πρόβλεψης ιδιοτήτων.

Η δράση της χημειομετρίας σε εφαρμογές για την περιγραφή συστημάτων, αντλεί δεδομένα από τις ιδιότητες ενός χημικού συστήματος και προσπαθεί να μοντελοποιήσει το πρόβλημα για την εύκολη μελέτη, παρατήρηση και επεξεργασία του. Εντοπίζει συσχετίσεις μεταξύ των παραγόντων και παρουσιάζει τη δομή του συστήματος. Απαραίτητη θεωθείται η χρήση υπολογιστικών μεθόδων, διότι σε πολλές περιπτώσεις το σύνολο των δεδομένων που επεξεργάζονται (Data Analysis), μπορεί να περιέχει και εκατοντάδες χιλιάδες δεδομένα και συσχετίσεις μεταξύ αυτών.

Η χημειομετρία εχει σημαντικότατες εφαρμογές στην Αναλυτική Χημεία και στην Μεταβολισμική - Metabolomics (κλάδος βιοχημείας για μελέτες της χημείας κυττάρων, DNA-mRNA, κυτταρικά παράγωγα κ.α.).

Chapter 2

ΠΑΛΙΝΔΡΟΜΗΣΗ

2.1 Θεωρία παλινδρόμησης

Πρόκειται για μία στατιστική τεχνική για την εκτίμηση των σχέσεων μεταξύ των εξαρτημένων με τις μη-εξαρτημένες μεταβλητές ενός συνόλου. Χρησιμοποιείται για την επεξεργασία δεδομένων που προέρχονται από περισσότερες από μια μεταβλητές.

Ενδιαφέρον έχει να αναφέρθεί ότι, στην Ανάλυση Παλινδρομησης, (Regression Analysis), η εκτίμηση της εξαρτώμενης μεταβλητής παρουσιάζει μία μορφή στατιστικής κατανομής πιθανότητας (Statistical Probability Distribution), γύρω από τη συνάρτηση παλινδρόμησης (Regression function).

Υπάρχουν πολλές διαφορετικές μαθηματικές μέθοδοι για να γίνει ο υπολογισμός και η εκτίμηση των ιδιοτήτων. Για προβλήματα χημειομετρίας δεν υπάρχει μια μοναδική -χωρίς μεγάλα σφάλματα- υπολογιστική μέθοδος, και δεν είναι πάντα σωστή λύση μια απλή PLS Ανάλυση Παλινδρομησης. Γι αυτό, για κάθε υπολογισθείσα ιδιότητα θα πρέπει να χρησιμοποιούνται και τα κατάλληλα εργαλεία.

Η μεθοδολογία επεξεργασίας που πρέπει να ακολουθηθεί σε κάθε περίπτωση, εξαρτάται και από το είδος αλλά και από την μορφή των δεδομένων που έχουν παραχθεί. Γενικά, οι σημαντικότερες μεθοδολογίες που μπορούν να εκτελεστούν για την επιλογή των μεταβλητών είναι τρείς:

a) Παλινδρόμηση με επιλογή μεταβλητών (Variable Selection)
b) Παλινδρόμηση με προσαρμογή ποινής (Penalised Regression)
c) Επιλογή λανθάνουσων μεταβλητών (Latent Variable Selection)

2.2 Μαθηματική προσέγγιση Παλινδρόμησης

Χρησιμοποιώντας παλινδρομικές μεθόδους, γίνεται εφικτό να χρησιμοποιηθούν όλα τα δεδομένα που έχουν συλλεχθεί από μια πειραματική διαδικασία και να συσχετιστούν μεταξύ τους και να βρεθεί η σχέση που τα διέπει, για τον υπολογισμό της ιδιότητας που ενδιαφέρει. Η βασική σχέση είναι η παρακάτω:

$$Y_1(\lambda_1) = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n$$

δηλαδή, η ιδιότητα Y, υπολογίζεται ενώ είναι γνωστό το μέγεθος των μεταβλητών x, αλλά και των παραγόντων α, όπου είναι το βάρος (loading) των μεταβλητών x. Στην ουσία η τιμή του α δείχνει πόσο μεγάλη επιρροή έχει η μεταβολή της μεταβλητής x στη μεταβολή της ιδιότητας Y. Το λ, είναι ένα ιδιοδιάνυσμα (eigenvector) που πολλαπλασιάζεται με την ιδιότητα.

Στόχος, όπως αναφέρθηκε παραπάνω, είναι η κατασκευή μιας μαθηματικής φόρμουλας f, που θα περιγράφει τη σχέση εξαρτώμενων-μη εξαρτώμενων μεταβλητών. Αυτή η σχέση περιγράφεται ως εξής:

$$\Upsilon = f(X, \alpha) => E(\Upsilon|X) = f(X, \alpha)$$

όπου (α) είναι οι άγνωστοι παράμετροι, (X) οι ανεξάρτητες και (Υ) οι εξαρτώμενες μεταβλητές.

Η επιλογή επίλυσης προβλημάτων με Παλινδρομικές μεθόδους είναι καλή επιλογή, εφόσον υπάρχει μεγάλος αριθμός πειραμάτων και δεδομένων.

Η παραπάνω εξίσωση για n αριθμό πειραμάτων και m αριθμό μεταβλητών, μετατρέπεται στο σύνολο των παρακάτω εξισώσεων:

$$Y_1 = \alpha_{11} x_1 + \alpha_{12} x_2 + ... + \alpha_{1n} x_n$$
$$Y_1 = \alpha_{21} x_1 + \alpha_{22} x_2 + ... + \alpha_{2n} x_n$$
$$...$$
$$Y_1 = \alpha_{m1} x_1 + \alpha_{m2} x_2 + ... + \alpha_{mn} x_n$$

οι υπολογισμένες τιμές προκύπτουν όπως διακρίνεται από το γινόμενο των πινάκων:

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$$

Σε περιπτώσεις που ο αριθμός τών πειραμάτων – διαφορετικών εξισώσεων – που προκύπτουν, είναι μεγαλύτερος από τον αριθμό των ιδιοτήτων που επιθυμείται να προσεγγιστεί, τότε υπάρχουν τεχνικές που θα βοηθήσουν να λυθεί αυτό το πρόβλημα, και ρόλο σε αυτό θα παίξουν όλες οι εξισώσεις χωρίς την παράλειψη καμίας. Είναι σημαντικό να ληφθούν υπόψη όλες οι εξισώσεις χωρίς, την αυθαίρετη επιλογή μερικών μόνο, για την αποφυγή ακόμα και πολύ μεγάλων σφαλμάτων.

Όσο περισσότερες εξισώσεις υπάρχουν στο σύστημα, τόσο μικρότερο θα είναι και το μέσο τετραγωνικό σφάλμα μέτρησης.

2.3 Εφαρμογή στην χημειομετρία

Οι τεχνικές της Γραμμικής Παλινδρόμησης, δεν έμειναν «έξω» από τον κόσμο της χημείας. Πολλοί επιστήμονες-χημικοί ασχολήθηκαν με την επεξεργασία δεδομένων με πολυμεταβλητά κριτήρια[4] ήδη από το 1960 και μετά, χρησιμοποιώντας γλώσσες προγραμματισμού όπως: FORTRAN ή ALGOL.

Σημαντικό αντίκτυπο είχαν αυτές οι μέθοδοι υπολογισμών ακόμα και στην Κβαντική Χημεία, καθώς οι υπολογισμοί που έπρεπε να γίνουν με το χέρι «large hand calculations», πλέον γινόντουσαν γρήγορα και χωρίς ακούσια σφάλματα.

Ο όρος «χημειομετρία», πρωτοεμφανίστηκε στην επιστημονική κοινότητα το 1972, για να περιγράψει τις μαθηματικές εφαρμογές στον τομέα τις Χημείας. Έδωσε την ευκαιρία στους επιστήμονες να υπολογίζουν και να παράγουν δεδομένα με τεράστιες ταχύτητες και πολύ μεγάλη ακρίβεια. Οι χημειομετρικές αναλύσεις περιλαμβάνουν, εκτός από την πρόβλεψη ιδιοτήτων/χαρακτηριστικών μιας ουσίας, την ταξινόμηση και αναγνώρισή της, συνήθως με χρήση τεχνικών γραμμικής άλγεβρας αλλά και άλλων μαθηματικών και στατιστικών τεχνικών.

2.4 Χρήσεις

Οι τεχνικές χρησιμοποιήθηκαν από εταιρίες, με σκοπό να εξοικονομήσουν χρόνο και κόπο από τους υπολογισμούς, ώστε να χρησιμοποιηθεί σαν κέρδος. Ακόμα, η κατασκευή μαθηματικών μοντέλων «patterns», μπορεί να επιλύσει προβλήματα κόστους εφαρμογών και πειραματικών διαδικασιών.

Εκτός από τους ερευνητικούς, οι τεχνικές αυτές χρησιμοποιήθηκαν για εμπορικούς και καταναλωτικούς σκοπούς από εταιρείες, για τον έλεγχο της ποιότητας, τη βελτιστοποίηση των διαδικασιών κλπ. Χρησιμοποιήθηκε επίσης, για να προσδιοριστούν οι παράγοντες που επηρεάζουν περισσότερο μια ιδιότητα π.χ. (συστατικά που επηρεάζουν περισσότερο την τήξη σκοριών και τέφρας).[5]

Προσοχή πρέπει να δοθεί σε περιπτώσεις όπου υπάρχουν πολλοί παράγοντες που επηρεάζουν μια μεταβλητή, όπως το φαινόμενο της υπερθέρμανσης του πλανήτη, η καρκινογένεση ή η μόλυνση υδάτων. Τότε η μελέτη κάθε παράγοντα ξεχωριστά, μπορεί να επιφέρει λάνθασμένα αποτελέσματα, καθώς η μία μεταβλητή μπορεί να επηρεάζει την τιμή της άλλης, γραμμικά ή μη.

Πολλά προβλήματα της επιστημονικής κοινότητας δε θα μπορούσαν να επιλυθούν, χωρίς τη χρήση πολυμεταβλητής παλινδρόμησης «Multivariable Regression». Οι πολλές διαστάσεις που μπορεί να περιέχει ο υπολογισμός μιας ιδιότητας, μπορεί να είναι πρόβλημα για κάποιες τυπικές μαθηματικές διαδικασίες.

Chapter 3

ΜΕΘΟΔΟΙ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

3.1 MLR (Πολυμεταβλητή Γραμμική Παλινδόμηση)

Ένα σύνηθες πρόβλημα στη στατιστική, είναι να βρεθεί η σχέση που συνδέει ένα σύνολο ανεξάρτητων μεταβλητών X με ένα σύνολο εξαρτώμενων μεταβλητών Υ.

Οι τρόποι σύνδεσης είναι πρακτικά άπειροι (εκθετική, λογαριθμική, γραμμική κ.α.), και εξαρτώνται από το είδος των δεδομένων. Για ένα μεγάλο σύνολο από ανεξάρτητες (βαθμωτές) μεταβλητές X, η διαδικασία καλείται Multiple Linear Regression. Αυτός ο όρος πρέπει να μην συγχέεται με τον όρο Πολυμεταβλητή Γραμμική Παλινδόμηση στον οποίο προβλέπονται οι πολλαπλές αλληλοσχετιζόμενες/εξαρτώμενες μεταβλητές (Multiple correlated dependent variables).

$$y = b_0 + \Sigma(b_i x_i) + E$$

Όταν ο αριθμός των μεταβλητών X είναι μικρότερος από 20, η MLR μπορεί να δώσει αξιόπιστα αποτελέσματα. Επίσης, η MLR χαρακτηρίζεται ασταθής, όταν στις μεταβλητές X υπάρχει αλληλεξάρτηση.

Η πολυπαραγοντική γραμμική παλινδρόμηση είναι μια πιο γενική μορφή της γραμμικής παλινδρόμησης, όμως, λαμβάνει υπόψη παραπάνω από μια ανεξάρτητες μεταβλητές. Το γενικό γραμμικό μοντέλο μπορεί να γραφτεί σαν :$Y = XB+U$. Το γενικό γραμμικό πολύ-παραγοντικό μοντέλο μπορεί να γραφτεί σαν:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + e_i$$

3.2 PLS (Μερικά Ελάχιστα Τετράγωνα)

Στη διαδιακσία της PLS, μελετάται η κατασκευή των διαγραμμάτων Score Values των μεταβλητών X , Υ και απεικονίζεται σε διάγραμμα [Score values (X) - Score Values (Y)], το οποίο κατασκευάζεται με τη βοήθεια των Διανυσμάτων Βαρών.

Παρατηρείται ότι καθώς αλλάζει η περιστροφή των Διανυσμάτων Βαρών, αλλάζει και η δομή των Scores στο Score-Score diagram. Η PLS περιστρέφει με τέτοιο τρόπο τα Διανύσματα Βαρών, έτσι ώστε η συνδιακύμανση των τιμών Scores Values X,Y να γίνει μέγιστη. Η συσχέτιση των μεταβλητών γίνεται ως εκτούτου μέγιστη.

Έτσι γίνεται εφικτό εάν από μια μεταβλητή X υπολογιστεί το Score X (μέσω του διαγράμματος και την ευθεία ελαχίστων τετραγώνων που θα αναπτυχτεί παρακάτω) να υπολογίζεται το Score Y και εν τέλει να γίνεται η πρόβλεψη της τιμής Y. Η διαδικασία της PLS είναι δομικά παρόμοια με την PCA, αλλά χρησιμοποιεί διαφορετικά κριτήρια.

Το γενικό γραμμικό μοντέλο λειτουργεί όταν η μεταβλητή Υ δεν είναι ένα βαθμωτό, αλλά ένα διάνυσμα. Ισχύει ότι:

$$E(y|x) = Bx.$$

Ένα διάνυσμα β αντικαθιστά τον πίνακα B. Πολυπαραγοντικά ανάλογα έχουν σχεδιαστεί για τις τεχνικές Ordinary Least Squares και Partially Least Squares.

Έστω ότι υπάρχει μια τυχαία μεταβλητή Υ, η οποία εξαρτάται από μία άγνωστη παράμετρο (η) και έστω ότι υπάρχει ένα σφάλμα μέτρησης αυτής της τιμής Υ που συμβολίζεται με (ε). Έστω ότι το (η) μπορεί να εκφραστεί στην παρακάτω μορφή:

$$\eta = \beta_0 + \beta_1\chi_1 +,,, + \beta_{p-1}\chi_{i,p-1} + \varepsilon_i, (i = 1, 2, ..., n)$$

όπου x_1, x_2, x_{p-1} είναι γνωστές μεταβλητές και φυσικά οι άγνωστοι παράγοντες που πρέπει να υπολογιστούν (β) Τότε το πρόβλημα αποκτά τη μορφή :

$$Y_1 = \beta_0 + \beta_1 x_{j1} + ... + \beta_{p-1}x_{i,p-1} + e_i$$

Matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & x_{12}...x_{u,p-1} \\ x_{10} & x_{11} & x_{12}...x_{u,p-1} \\ x_{10} & x_{11} & x_{12}...x_{u,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}$$

Δηλαδή πιο απλά:

$$\Upsilon = X\beta + \varepsilon$$

Η μέθοδος η οποία θα υπολογίσει το $(\hat{\beta})$ καλείται: Μέθοδος Ελαχίστων Τετραγώνων.

Αυτή η μέθοδος έχει ως κύριο στόχο να ελαχιστοποιήσει τα τετράγωνα, $\Sigma i(\varepsilon_i^2)$ Έτσι ώστε: αν $\theta = X\hat{\beta}$, ... $X'(\Upsilon - \theta) = 0$ ή $X'\theta = X'\Upsilon$

$$X'X\hat{\beta} = X'\Upsilon$$

$$\hat{\beta} = (X'X)^{-1}X'\Upsilon$$

Το $(\hat{\beta})$ καλείται (ordinary) Least Square του β

Το σφάλμα των μετρήσεων από την ευθεία ελαχίστων τετραγώνων προσδιορίζεται ως $\Delta = \delta_1^2 + \delta_2^2 + ... + \delta_n^2$ συνεπώς: $\Delta = (y_1 - \alpha - \beta * X_1)^2 + (y_2 - \alpha - \beta * X_2)^2 + ... + (y_n - \alpha - \beta * X_n)^2)$

Άρα: $\Delta = \sum_{i=1}^{n}(y_i - \alpha - \beta * x_i)^2$

Αυτή είναι η εξίσωση σφαλμάτων της γραφικής παράστασης (X,Y). Η μερική παράγωγος που μηδενίζει τη συνάρτηση, θα δώσει τις τιμές (β, α) που την ελαχιστοποιούν.

$\frac{\partial\Delta}{\partial\alpha} = 0$ ή $\frac{\partial * \sum_{i=1}^{n}(y_i - \alpha - \beta * X_i)^2}{\partial * \alpha} = 0$

$\frac{\partial\Delta}{\partial\beta} = 0$ ή $\frac{\partial * \sum_{i=1}^{n}(y_i - \alpha - \beta * X_i)^2}{\partial * \beta} = 0$

Άρα:

$\hat{\beta} = \frac{n * \sum x_i * y_i - \sum x_i * \sum y_i}{n * \sum (x_i)^2 - (\sum x_i)^2}$

$\hat{\alpha} = \overline{Y} - \hat{(}\beta) * \overline{X}$

Το $(\hat{\beta})$ σχηματίζει την ευθεία που συσχετίζει τις Score Values X και Y, για την πρόβλεψη τις τιμής Y και αποτελεί την ευθεία που διέρχεται καλύτερα από τα δεδομένα με μέγιστη συνδιακύμανση.

3.3 iPLS (Με διαστήματα)

Η iPLS, έχει την ικανότητα να επιλέγει μια ομάδα μεταβλητών η οποία θα δώσει μια ανώτερη πρόβλεψη σε σχέση με τη χρήση όλων των μεταβλητών.

Η iPLS ουσιαστικά υπολογίζει PLS με διάφορους τυχαίους συνδυασμούς την επιλογή των μεταβλητών και διαλέγει εκείνη στην οποία θα έχει το μικρότερο μέσο τετραγωνικό σφάλμα του cross-validation

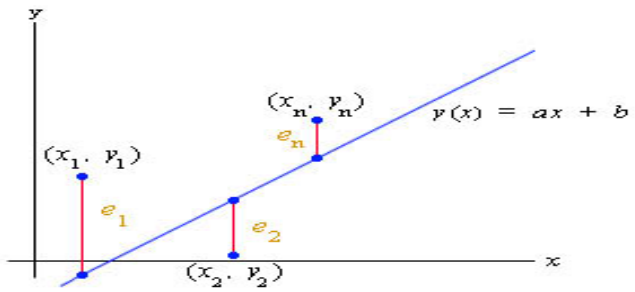


Figure 3.1: Partial Least Squares

που χρησιμοποιεί.

Η τεχνική αυτή, αν και επιλέγει μεταβλητές αυθαίρετα, μπορεί να βελτιώσει το σφάλμα της πρόβλεψης του μοντέλου. Επίσης υπάρχει η πιθανότητα ακούσιας παράλειψης χρήσιμων δεδομένων από ένα μοντέλο. Χρησιμοποιώντας ένα μικρότερο αριθμό μεταβλητών για να κάνει μια πρόβλεψη, σημαίνει ότι κάθε μεταβλητή έχει μια μεγαλύτερη επίδραση επί της τελικής πρόβλεψης.

Εάν οποιαδήποτε από αυτές τις μεταβλητές καταστραφεί, υπάρχουν άλλες μεταβλητές που μπορούν να χρησιμοποιήθουν στη θέση τους. Ομοίως, είναι πιο δύσκολο να ανιχνευθεί ένα σφάλμα.

Υπάρχει επίσης, δυσκολία στο να μελετηθούν και να υπολογιστούν τα σφάλματα που αποδίδει μια iPLS. Ως εκ τούτου, οι ανάγκες του τελικού μοντέλου θα πρέπει πάντα να εξετάζονται προσεκτικά, όταν η iPLS κάνει μια επιλογή μεταβλητής.

Αυτή η διαδικασία δεν αναπτύσεται στην παρούσα εργασία για την επεξεργασία των δεδομένων.

3.4 Μείωση διαστάσεων

Η μείωση των διαστάσεων είναι μια διαδικασία που γίνεται για να αποτυπωθούν, να μελετηθούν δεδομένα πολλών διαστάσεων, με ένα πιο βολικό και κατανοητό τρόπο.

Σκοπός είναι να αποτυπωθούν τα δεδομένα για να γίνει επεξεργασία τους, ώστε να βρεθούν οι σχέσεις τους, τα πιθανά πρότυπα (patterns) τα οποία θα οδηγήσουν σε μια στατιστική υπόθεση, χωρίς να χαθεί χρήσιμη πληροφορία από το σύστημα. Ακόμα γίνεται ορατή η δομή των δεδομένων σε γραφήματα (συνήθως 2 διαστάσεων) PCA Plot. Επιπρόσθετα, ομαδοποιούνται τα δεδομένα (clusterization), ομαλοποιούνται (smoothing), υπολογίζονται διαγράμματα πυκνότητας-πιθανότητας και ταξινομούνται.

Σε ένα σύστημα με πολλούς παράγοντες που επηρεάζουν τα -ως προς μελέτη- χαρακτηριστικά, θα ήταν λάθος να μειωθούν οι διαστάσεις του προβλήματος αν μειώνονταν η ποσότητα των διαστάσεων αυθαίρετα, καθώς αυτόματα θα απορρίπτονταν μεγάλο μέρος της πληροφορίας, βγάζοντας εσφαλμένα

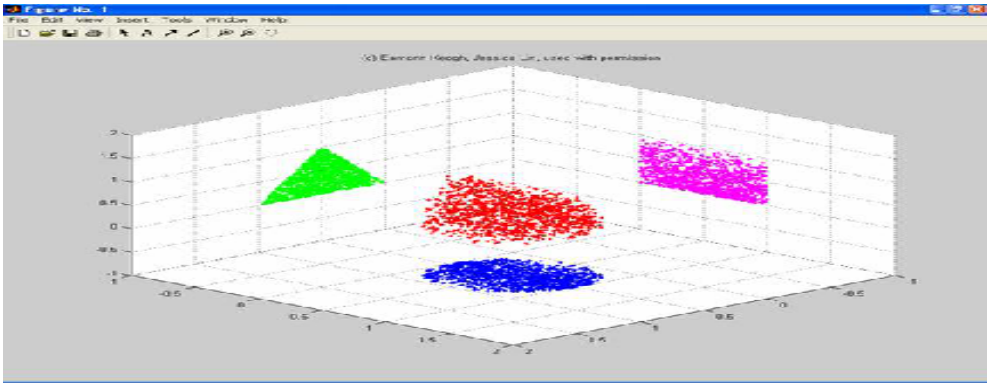


Figure 3.2: Dimensionality Reduction

αποτελέσματα. Εφικτό όμως θα ήταν να δημιουργηθούν καινούριες μεταβλητές όπου είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών, σε περίπτωση μικρού αριθμού διαστάσεων.

Τεχνικές μείωσης διαστάσεων υπάρχουν πολλές όπως: principal component analysis (PCA), singular value decomposition (SVD), nonnegative matrix factorization, factor analysis, και linear discriminant analysis.

Στην παρούσα εργασία το μειωμένο σύνολο διαστάσεων θα λάβει μέρος και στην μεθοδολογία PCA και στην PLS.

3.5 PCR (Παλινδρόμηση Κυρίων Συνιστωσών)

Πρόκειται για μια από τις πιο διαδεδομένες τεχνικές στα πολυπαραγοντικά στατιστικά συστήματα. Με την τεχνική αυτή γίνεται μια μελέτη της διακύμανσης και της συσχέτισης των μεταβλητών.

Είναι μια τεχνική η οποία βασίζεται στις PCA και MLR. Τυπικά εξετάζει τα δεδομένα που εξάγονται από ένα Standard Linear Regression Model και χρησιμοποιεί PCA για να προβλέψει τους άγνωστους συντελεστές της παλινδρόμησης στο μοντέλο αυτό.

Διαδικασία της PCA μπορεί να εκτελεστεί με αποδόμηση ιδιοτιμών της συνδιακύμανσης δεδομένων (data covariance) ή πίνακα συσχέτισης (correlation matrix), ή Αποδόμηση μοναδιαίας τιμής (SVD) του πίνακα δεδομένων, συνήθως μετά από Κανονικοποίηση (και Ομαλοποίηση ή χρήση Z-scores) του πίνακα δεδομένων για κάθε ιδιότητα.

Στην PCR, αντί η παλινδρόμηση να εξαρτάται από τις ιδιότητες ενός πίνακα δεδομένων, εξαρτάται από τις κύριες συνιστώσες Κυρίων Συνιστωσών και τις χρησιμοποιεί σαν Παλινδρομητές.

Όταν υπάρχουν δεδομένα με μεγάλο αριθμό διαστάσεων, που δεν μπορούν να απεικονιστούν σε ένα απλό διάγραμμα δύο ή τριών διαστάσεων για την ερμηνεία και εύκολη επεξεργασία των δεδομένων, θα πρέπει να χρησιμοποιείται η τεχνική της PCA (Principal Component Analysis) ή (factor Analysis).

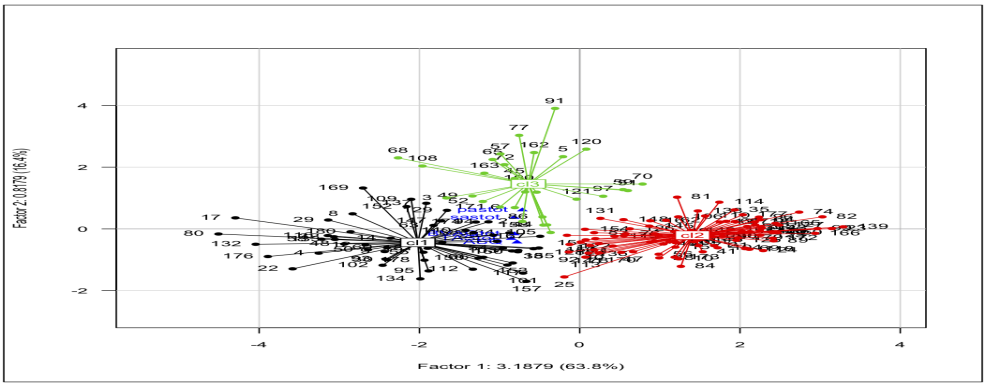


Figure 3.3: PCA Classification example

Με αυτήν την τεχνική, πολυδιάστατα δεδομένα απεικονίζονται σε δυσδιάστατα ή τρισδιάστατα διαγραμματικά μοντέλα. Στην ουσία μειώνονται οι διαστάσεις του προβλήματος και μένουν οι πιο σημαντικές διαστάσεις που επηρεάζουν τις ιδιότητες, και χωρίς να χαθεί πολύ μεγάλο μέρος της πληροφορίας,[8] δημιουργείται ένα σύνολο που προέρχεται από μεταβλητές που είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών.

Αν υπάρχει ένα διάνυσμα x με αρκετές p μεταβλητές, όπου υπάρχει συσχέτιση μεταξύ τους, τότε θα ήταν πάρα πολύ δύσκολο να μελετηθούν οι μεταβλητές και όλες οι $(1/2) * p * (p-1)$ συσχετίσεις ή συνδιακυμάνσεις τους (εκτός αν ο αριθμός μεταβλητών p είναι πολύ μικρός)[10]. Θα ήταν εφικτό, εκτός από το να μειωθούν οι διαστάσεις ενός συνόλου μεταβλητών, απλά να υπολογίζονταν η διακύμανση των δεδομένων κάθε στήλης ξεχωριστά και να μελετούνταν αυτές που έχουν την μεγαλύτερη τιμή. Αυτές είναι που θα επηρέαζαν περισσότερο τα αποτελέσματα και οι οποίες περιέχουν το μεγαλύτερο μέρος της πληροφορίας για μια PCA Analysis.

Στην τεχνική της PCA, υπάρχει η ικανότητα να διακριθούν τα δεδομένα σε ομάδες (clusters-Cluster Analysis) με κοινά χαρακτηριστικά, ως προς μια ιδιότητα. Βοηθάει στο να αναλυθούν τα δεδομένα στον πίνακα, να γίνει αντιληπτή η κατανομή τους, να βρεθεί ο βαθμός συγγραμμικότητας μεταξύ των ιδιοτήτων των δειγμάτων, να βρεθούν οι διαφορές τους και να βρεθούν οι σχέσεις που συσχετίζουν τις διάφορες στήλες και γραμμές του πίνακα. Αυτού του τύπου η ανάλυση καλείται και Classification and Discriminant Analysis.

PCA, χρησιμοποιείται όπου είναι δυνατόν για να διαχωρίσει ομάδες δεδομένων με κοινά χαρακτηριστικά με βάση τις διαθέσιμες μετρήσεις. Όπως και η PCA, έτσι και η discriminant function analysis, είναι βασισμένη στην ιδέα να βρεθεί ο καλύτερος γραμμικός συνδυασμός των αρχικών μεταβλητών.[6]

Μια ακόμα πληροφορία που λαμβάνεται από το διάγραμμα της PCA, προέρχεται από την εποπτεία των τιμών των Κυρίων Συνιστωσών πάνω στο διάγραμμα. Ρόλο παίζει η θέση τους και αν έχουν θετικά ή αρνητικά βάρη.

Η τεχνική της PCR χρησιμοποιεί τις κύριες συνιστώσες που προέρχονται από την ανάλυση PCA,

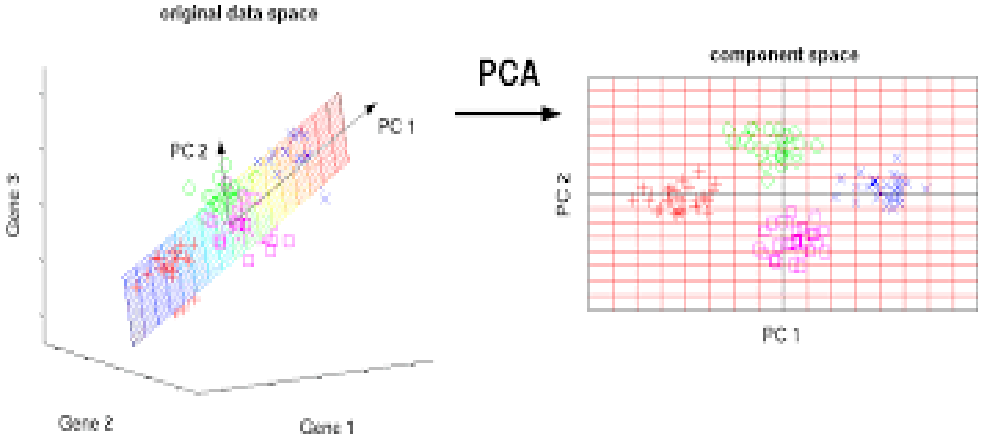


Figure 3.4: Principal Components

και εκεί εφαρμόζει MLR. Αυτό γίνεται εφικτό, μετατρέποντας τον αρχικό πίνακα σε ένα καινούριο «set» μεταβλητών Κυρίων Συνιστωσών, τα οποία είναι μη-συσχετισμένα μεταξύ τους και έχουν τέτοια παράταξη, ώστε τα πρώτα διατηρούν την μεγαλύτερη διακύμανση (variation) από όλες τις μεταβλητές [10].

Η PCA, έχει την δυνατότητα να υπολογίζει τη μέγιστη διακύμανση που έχουν τα δεδομένα ως προς μία διάσταση και να χρησιμοποιεί τα ιδιοδιανύσματα της κατεύθυνσης αυτής (γίνεται ουσιαστικά περιστροφή των διαστάσεων στον χώρο) όπου οι άξονες πρέπει να είναι κάθετοι μεταξύ τους, ώστε να υπολογίζει τις Κύριες Συνιστώσες PC(1) και PC(2).

Σε ένα τυπικό διάγραμμα PCA στον άξονα x, υπάρχουν οι τιμές του PC(1) και στον y, οι PC(2), τα οποία όπως ειπώθηκε διεκδικούν και το μεγαλύτερο μέρος της πληροφορίας.

Ισχύει ότι:

$$PC_{jk} = a_{j1}x_{k1} + a_{j2}x_{k2} + ... + a_{jn}x_{kn}$$

όπου PC_{jk}, είναι η τιμή για την κύρια συνιστώσα j του αντικειμένου k (δηλαδή το score value του αντικειμένου j στην συνιστώσα k), a_{j1} είναι το βάρος (Loading) της μεταβλητής 1 της συνιστώσας j, x_{k1}, είναι η μέτρηση της μεταβλητής 1 του αντικειμένου k και n είναι ο αριθμός των μεταβλητών.

3.6 SVD (Ανάλυση πίνακα σε ιδιάζουσες τιμές)

Η τεχνική της PCA μπορεί να συσχετίζεται επίσης και με την τεχνική SVD (Αποδόμηση μοναδιαίας τιμής). Εκεί προβάλλονται σχετικά απλά οι κύριες συνιστώσες μιας συλλογής δεδομένων, για μια

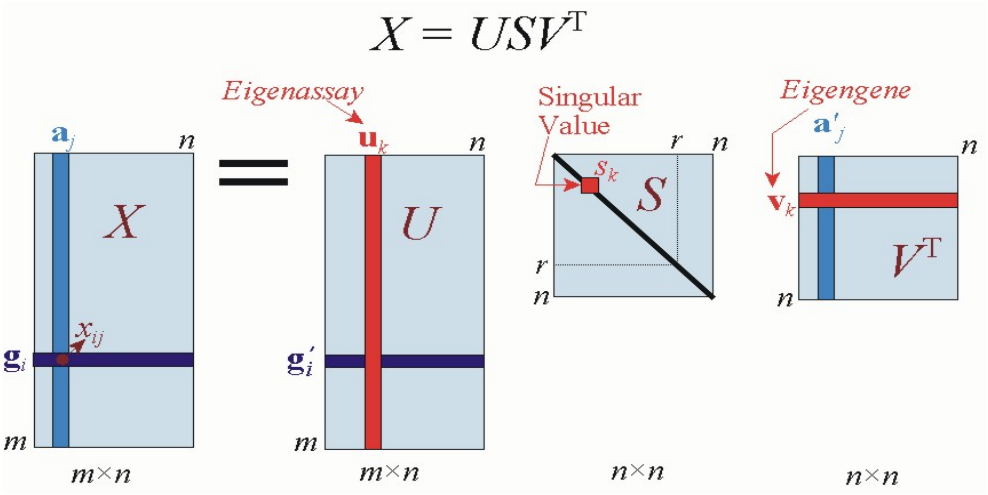


Figure 3.5: Singular Value Decomposition

κλασική PCA.

Η Ανάλυση κύριων συνιστωσών (PCA) μπορεί να γίνει από την αποσύνθεση της ιδιοτιμής (eigenvalue decomposition) ενός πίνακα συν-διασποράς ή συσχέτισης των δεδομένων ή την τεχνική SVD, με ομαλοποίηση ή χρήση Z-scores. Τα αποτελέσματα της PCA, δηλαδή τα PCs, χρησιμοποιούνται για την προβολή του διαγράμματος της PCA.

Η τεχνική SVD για τη διαδικασία της PCA έχει μερικά μειονεκτήματα, καθώς μπορεί να μη λάβει υπόψη της κάποιες ιδιότητες των μεταβλητών όπως: η ομαλότητα, οι ελάχιστες διαφορές και η ευρωστία τους[7]. Αναφορικά για την SVD: ένας πίνακας [A] με ανάλυση SVD θα δώσει τρείς πίνακες.

U (Left Singular Values Matrix)
S (Singular Values of matrix)- τετράγωνα του είναι τα ιδιοδιανύσματα του A)
V (Loading matrix).
*Matlab code

```
[U,S,V]=svd(A)

A=USV'

T=A*V
 % T is the Z-Score Matrix Used for PCA
```

Ο πίνακας score είναι μια ομάδα δεδομένων που προέρχεται από μεταβλητές που είναι γραμμικοί συνδυασμοί των αρχικών τιμών.

Chapter 4

ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

4.1 Παλινδρόμηση με επιλογή μεταβλητών

Στον τομέα της στατιστικής ανάλυσης, για την πρόβλεψη μεταβλητών, μπορεί να χρησιμοποιηθεί η μέθοδος παλινδρόμησης με επιλογή μεταβλητών ή Feature selection.

Η φιλοσοφία αυτής της μεθόδου συνήθως εφαρμόζεται πάνω σε ένα πολύ μεγάλο set δεδομένων, που σκοπό έχει την επεξεργασία των δεδομένων, ώστε να μειώσει την ποσότητα τους.

Στην παρούσα εργασία, η επιλογή των μεταβλητών γίνεται στο σύνολο των κυματάριθμων: από 1500-2500 cm^{-1}, 2500-3500 cm^{-1} και σε όλο το φάσμα.Η επιλογή έγινε με την εποπτεία των φασματογραφημάτων.

Η τεχνική αυτή κρατάει μόνο τις τιμές των μεταβλητών που έχουν ικανοποιητική σημασία για το αποτέλεσμα και ως απώτερο σκοπό έχει την μείωση της περιπλοκότητας του συστήματος και την αποτροπή της υπερπροσαρμογής των δεδομένων (overfitting). Έτσι τα αποτελέσματα μπορούν να αξιολογούνται καλύτερα από τους ερευνητές και να μειώνεται ο χρόνος επεξεργασίας των δεδομένων, χωρίς όμως να χάνεται σημαντική πληροφορία από το σύστημα. Η μείωση των μεταβλητών μπορεί επίσης να προβλέψει το πρόβλημα της πολυσυγγραμικότητας (Multicollinearity) των μεταβλητών. Οταν υπάρχει αυτο το πρόβλημα είναι δύσκολο κάποιες φορές να ερμηνευθεί, ποιά μεταβλητή επιρρεάζει περισσότερο το αποτέλεσμα.

Με τη μέθοδο Forward Stepwise Regression (FSR), αρχικά γίνεται η επιλογή μεταβλητών που θα μείνουν για την επεξεργασία, ενώ διαγράφονται οι άλλες που δεν είναι σημαντικές. Η επιλογή ενός υποσυνόλου από όλο το σύνολο των δεδομένων, χρειάζεται εξειδικευμένους υπολογιστικούς αλγόριθμους.

Οι αλγόριθμοι αυτοί διαχωρίζονται σε Wrappers, Filters και Embedded. Οι Wrappers είναι κομμάτια του αλγόριθμου που επιλέγουν υποσύνολα και τα αξιολογούν, εφαρμόζοντας ένα μαθηματικό μοντέλο στο υποσύνολο αυτό. Είναι σχετικά χρονοβόροι και υπάρχει κίνδυνος overfitting στο μοντέλο. Τα filters και Embedded είναι παρόμοιας φύσης.[15]

Η τεχνική FSR ξεκινά αρχικά χωρίς προβλεπόμενους παράγοντες (predictors) και εν συνεχεία

συμπεριλαμβάνει νέους, όπου κρίνει σωστό. Η διαδικασία μπορεί ακόμα και σε ένα επόμενο στάδιο να διαγράψει μεταβλητές.

Για τους υπολογισμούς αυτής της μεθοδολογίας, θα πρέπει να αναφερθεί ότι για την πρόβλεψη της σημαντικότητας των predictors χρησιμοποιούνται οι p-values του partial F statistical test και η τεχνική MLR, για να συσχετίσει τους predictors με τις μεταβλητές.

4.2 Επιλογή λανθάνουσων μεταβλητών

Αυτή η μεθοδολογία εξαρτάται από τη συσχέτιση που έχουν οι Predictors και οι μεταβλητές απόκρισης. Συγκεκριμένα, θεωρεί ότι η παρατηρούμενη μεταβλητότητα είναι αποτέλεσμα μόνο μερικών μεταβλητών που έχουν την κύρια σημασία στο αποτέλεσμα. Αυτές οι μεταβλητές ονομάζονται Λανθάνουσες Μεταβλητές, και μπορούν να υπολογιστούν με μεθόδους γραμμικών συνδυασμών.

Για Latent variable methods γίνεται χρήση δύο κύριων σχέσεων.

$$X = TP^T + E$$

$$y = Tc^T + f$$

Το P είναι ένας ρ x a πίνακας βαρών, το T είναι ένας n x a ένας πίνακας scores,c είναι ένα διάνυσμα 1 x a που σχετίζεται με τα score των λανθανουσών μεταβλητών και της ανταπόκρισης. E και f είναι τυχαία σφάλματα, με διαστάσεις n x p και n x 1, αντίστοιχα.

Σε αυτή την κατηγορία μεθόδων, θα γίνει χρήση Principal Component Regression (PCR) Principal Component Regression with a Forward Stepwise procedure (PCR-FS), Partial Least Squares (PLS), και interval PLS (iPLS).

Αυτή η μέθος δεν εκτελείται στην συνέχεια της εργασίας.

4.3 Παλινδρόμηση με προσαρμογή ποινής

Η μέθοδος αυτή εκτιμά ταυτόχρονα τη μεταβλητή και τον συντελεστή κατά την επεξεργασία των δεδομένων. Σκοπός είναι να προστεθεί μια ποινή στις τιμές των συντελεστών που παράγονται από μια διαδικασία Ελαχίστων Τετραγώνων, αποσκοπώντας σε μια εκδοχή της μεθόδου MLR, αλλά παρουσιάζοντας μικρότερες διακυμάνσεις και μικρότερα σφάλματα, με το αποτελέσματα της καλύτερης πρόβλεψης.

Υπάρχουν τέσσερις κύριες ομάδες μεθόδων γι αυτή την κατηγορία : Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Nets (EN) και Support Vector Regression (SVR).

Στατιστικά προβλήματα τέτοιου τύπου, έχουν ως βάση την θεωρία της μέγιστης πιθανοφάνειας.

Για την αντιμετώπισή τους μεγιστοποιείται η σχέση

$$M(\theta) = l(\theta|x) - \lambda P(\theta)$$

[17]

P: είναι μια συνάρτηση που προσθέτει ποινή σε αυτές τις τιμές που είναι λιγότερο ρεαλιστικές.
λ: ονομάζεται regularization parameter.

Στο Regression, συνήθως επιθυμείται η ελαχιστοποίηση μιας συνάρτησης απώλειας (συνήθως το τετράγωνο απώλειας σφάλματος), αντί για τη μεγιστοποίηση της πιθανοφάνειας. Ισοδύναμα υπολογίζεται το ϑ σαν:

$$M(\theta) = L(\theta|x) - \lambda P(\theta)$$

Εδώ, το L είναι μια συνάρτηση απώλειας (συνήθως είναι ανάλογη με - log πιθανότητας, όπως το υπολειπόμενο άθροισμα τετραγώνων)
Το (P) είναι επομένως μια λειτουργία η οποία αλλάζει τους συντελεστές που βρίσκονται πιο μακριά από το μηδέν.

Τα Penalties αυτά προκύπτουν για δύο από τις προαναφερθείσες μεθόδους

$$Ridge : P(\beta) = \sum_{j=1}^{p} \beta_j^2$$

$$Lasso : P(\beta) = \sum_{j=1}^{p} |\beta_j|$$

[16]

Η τεχνική αυτή δεν αναπτύσεται στην παρούσα εργασία.

Chapter 5

ΔΕΙΓΜΑΤΑ ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ ΜΕΤΡΗΣΗΣ

5.1 Παρουσίαση Ιδιοτήτων δειγμάτων

Εν συνεχεία παρουσιάζεται το σύνολο των δεδομένων της εργασίας.

Για τα μοντέλα πρόβλεψης ιδιοτήτων Αριθμού Κετανίου και κινηματικού ιξώδους με την φασματοσκοπία IR δώθηκαν: δεδομένα φυσικών ιδιοτήτων για τα περισσότερα δείγματα : Ιξώδες (κινηματικό σε 15 βαθμούς Κελσίου), αλλά και ο Αριθμός Κετανίου. Τέλος, κάθε δείγμα αποτελείται από τα φασματοσκοπικά του δεδομένα διαπερατότητας.

Τα φασματοσκοπικά δεδομένα που δόθηκαν για επεξεργασία, αρχικα ήταν σε 37 δείγματα όπου μετά από αξιολόγηση, κρατήθηκαν και θεωρήθηκαν σωστά μόνο τα 31. Το κάθε δείγμα έχει 1764 δεδομένα φασματοσκοπίας από μήκος κύματος 599,8617 μέχρι 4000,3643, με βήμα -1.9277. Αυτά σχηματίζουν τον πίνακα 54674 δεδομένων Οπτής Διαπερατότητας (31x1764) ο οποίος επεξεργάζεται στην παρούσα εργασία.

Για τα μοντέλα πρόβλεψης ιδιοτήτων Αριθμού Οκτανίου με την φασματοσκοπία NIR δώθηκαν: 44 δείγματα με τον αντίστοιχο Αριθμό Οκτανίου μετρημένο κατα RON (Research Octane Number). Τα φασματοσκοπικά δεδομένα ξεκινούν απο 894.73 μέχρι 1791.96 cm^{-1} με βήμα 3.5, συνολικά 256 σημεία κυμματάριθμων. Δώθηκαν επίσης τα δεδομένα Απορρόφησης.

Όλα τα δεδομένα αποτυπώθηκαν σε αρχεία txt, excel και Matlab, σε μορφή Table και Vectors για την ομαλή επεξεργασία τους. Ολοκληρωμένα δεδομένα ιδιοτήτων υπάρχουν μόνο από τα παρακάτω δείγματα του πίνακα Α. Τα δεδομένα αυτά θα χρησιμοποιηθούν αναλόγως.Παρακάτω παρατίθεται ο πίνακας των ιδιοτήτων των δειγμάτων που θα χρησιμοποιηθούν για τα μοντέλα πρόβλεψης, δηλαδή

τον Αριθμό Κετανίου , το Ιξώδες(15 C) και τον Αριθμό Οκτανίου για NIR και IR φασματοσκοπία.

α/α	Samples IR	CN	Viscosity(15)	Samples NIR	Octane Number
1	HC 20-225	38.57	-	REF21	99.3
2	HC 225-250	44.7	-	REF25	100
3	HC 250-275	49.55	3.808	REF28	99.3
4	HC 275-300	55.87	5.981	REF37	99.9
5	HC 300-325	62.1	9.168	REF38	99.3
6	HC-RES	72.52	-	REF48	101.7
7	HDS 25-225	43.24	-	REF60	102.4
8	HDS 200-225	47.62	1.691	REF61	99.6
9	HDS 225-250	51.29	2.435	REF62	100.6
10	HDS 250-275	54.26	3.692	REF76	99.5
11	HDS 275-300	58.63	5.712	REF84	100
12	HDS 300-325	61.95	8.734	REF85	100
13	HDS RES	63.26	-	REF86	101
14	HT 225-250	43.51	-	REF87	99.3
15	HT 250-275	49.99	3.840	REF88	92.2
16	HT 275-300	55.51	5.793	REF89	99.4
17	HT 300-325	60.12	9.088	REF90	99.5
18	MOH HDS 25-200	-	1.046	REF91	99
19	MOH HDS 200-225	41.24	1.813	REF92	99.14
20	MOH HDS 225-250	45.53	3.763	ISO97	86.3
21	MOH HDS 225-250	45.53	3.763	ISO97	86.3
22	MOH HDS 250-275	50.01	-	ISO98	86.3
23	MOH HDS 275-300	55.23	5.597	GAS1	97
24	MOH HDS 300-325	58.61	8.669	GAS6	96.6
25	MOH HDS RES	59.55	-	GAS10	99.9
26	MOH MHC 25-200	37.59	-	GAS14	97
27	MOH MHC 200-225	43.21	1.939	GAS36	95.1
28	MOH MHC 225-250	46.29	2.855	GAS52	95
29	MOH MHC 250-275	52.37	3.756	GAS56	95
30	MOH MHC 275-300	58.11	6.601	GAS67	97.1
31	MOH MHC 300-325	62.59	9.494	GAS73	97
32	MOH MHC 325-350	69.33	-	GAS80	95.2
33	MOH MHC RES	71.23	-	FCC82	94
34	-	-	-	FCC100	93
35	-	-	-	FCC102	92
36	-	-	-	FCC103	91.8
37	-	-	-	DIM11	95.8
38	-	-	-	DIM45	95.5
39	-	-	-	DIM55	95.4
40	-	-	-	DIM70	95.8
41	-	-	-	DIM75	95.6
42	-	-	-	DIM77	95.5
43	-	-	-	DIM94	95.3
44	-	-	-	DIM95	95.5
44	-	-	-	ALK9	95

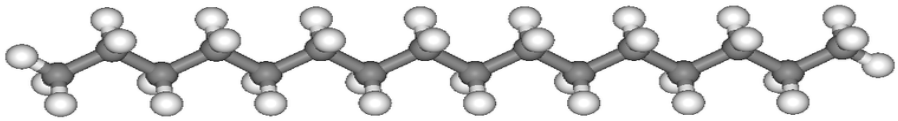


Figure 5.1: images/Cetane

5.2 Ιδιότητες diesel

Το πετρέλαιο ντίζελ είναι ένα μίγμα υδρογονανθράκων, με σημεία βρασμού στην περιοχή από 150 έως 380ο C, οι οποίοι παράγονται από το πετρέλαιο κυρίως με την μέθοδο της διαφορικής διύλισης από το αργό πετρέλαιο.

Το αργό πετρέλαιο αποτελείται από υδρογονάνθρακες τριών κύρίων κατηγοριών: παραφινικά, ναφθενικά (ή κυκλοπαραφινικών), και αρωματικών υδρογονανθράκων.

Το αργό πετρέλαιο μετατρέπεται σε καύσιμα μεταφοράς – όπως της βενζίνης, καύσιμα αεριωθουμένων, και ντίζελ καυσίμου αλλά και άλλα προϊόντα πετρελαίου, όπως το υγροποιημένο αέριο πετρελαίου (LPG), τα καύσιμα θέρμανσης, λιπαντικά έλαια, κερί, και άσφαλτο. Υψηλής πυκνότητας προϊόντα του αργού πετρελαίου (έλαια), περιέχουν περισσότερα από τα ελαφρύτερα προϊόντα που χρειάζονται για την παραγωγή καυσίμων μεταφορών και γενικά έχουν χαμηλότερη περιεκτικότητα σε θείο.

Το κλάσμα της Diesel, αποτελείται από περίπου 75% κορεσμένους υδρογονάνθρακες (παραφίνες κυρίως συμπεριλαμβανομένων n, iso, και κυκλοπαραφίνων), και περίπου 25% αρωματικούς υδρογονάνθρακες (συμπεριλαμβανομένων των ναφθαλίνων και αλκυλοβενζολίων). Η μέση χημική φόρμουλα του κοινού ντίζελ καυσίμου είναι C12H23, που κυμαίνεται περίπου από C10H20 να C15H28.[11].

Συνήθως παγώνει σε θερμοκρασίες γύρω στους -8,1oC, ενώ το ιξώδες είναι αντιστρόφως ανάλογο της θερμοκρασίας. Μετατρέπεται σε gel περίπου στους -19oC. Συμβατικά καύσιμα Diesel εξατμίζονται σε θερμοκρασίες μεταξύ 149oC και 371oC [12]

5.3 Κετάνιο - Αριθμός Κετανίου

Ο Σκοπός της εργασίας είναι ο προσδιορισμός του αριθμού κετανίου που χαρακτηρίζει ένα πετρέλαιο Diesel, με βάση τα φασματοσκοπικά δεδομένα που υπάρχουν από ένα άγνωστο δείγμα.
Ο Αριθμός Κετανίου του Μοντέλου εκπαίδευσης, υπολογίστηκε με βάση το International Standard 5165, με την τεχνική επιτροπή ISO/TC 28, Petroleum products and lubricants. [18]

Το κετάνιο ή αλλιώς δεκαεξάνιο $C_{16}H_{34}$ (Hexadecane), είναι ένας υδρογονάνθρακας που ανήκει στην κατηγορία των αλκανίων, δηλαδή πρόκειται για μια ευθεία ανθρακική αλυσίδα με απλούς δεσμούς μεταξύ των ανθράκων. Αποτελείται από 10359 συντακτικά ισομερή.
Χρησιμοποιείται κυρίως για την έννοια του αριθμού κετανίου, ένα μέγεθος που αφορά την ποιότητα

για την αυτό-ανάφλεξη /εκρηκτικότητα των πετρελαίων Diesel. Ο προσδιορισμός του είναι πολύ σημαντικός γιατί από αυτόν καθορίζεται ο τρόπος που θα λειτουργήσει μια μηχανή πετρελαίου. Η ανάφλεξη του κετανίου επηρεάζεται πολύ από την πίεση στην οποία βρίσκεται.

Το καθαρό κετάνιο, έχει αριθμό κετανίου το 100, και ουσιαστικά δηλώνει την περιεκτικότητα % του κετανίου των μειγμάτων. Χρησιμοποιείται ως βάση αναφοράς για άλλα μείγματα. Δε θα πρέπει να συσχετίζεται με τον δείκτη κετανίου (Cetane Index), διότι αυτός είναι και συνάρτηση της πυκνότητας του καυσίμου και του τρόπου διύλισης.

Η μεθοδολογία του Linear Regression Analysis δίνει πάρα πολύ καλά αποτελέσματα για τέτοιου είδους προβλέψεις. Γενικά η σχέση που θεωρείται για το Regression είναι:

$$CN = K + ax_1 + bx_2 + cx_3 + ...+$$

όπου $K, a, b, c..$, είναι οι σταθερές που προκύπτουν από το αποτέλεσμα του Regression και $x_1, x_2, x_3...$ είναι οι μη εξαρτώμενες μεταβλητές (εδώ απορρόφηση σε συγκεκριμένο μήκος κύμματος). [13]

5.4 Κινηματικό Ιξώδες

Πρόκειται για ένα μέτρο το οποίο περιγράφει την αντίσταση που παρουσιάζει ένα υγρό κατά την άσκηση σε αυτό μιας διατμητικής τάσης. Το κινηματικό ιξώδες μπορεί να εκφραστεί και ως η αντίσταση στην ροή που προβάλλει.

Η αντίσταση αυτή οφείλεται στο είδος των διαμοριακών σχέσεων (εφαρμογή δυνάμεων συνοχής) και τις συγκρούσεις που υπάρχουν στα μόρια της υπό μελέτης ουσίας, αλλά και από το γεγονός ότι σε μια ροή, οι ταχύτητες των μορίων είναι διαφορετικές, και εξαρτούνται από την θέση στην οποία βρίσκονται όπως χαρακτηριστικά συμβαίνει κατά την διέλευση ενός υγρού σε έναν σωλήνα, όπου το υγρό κοντά στα τοιχώματα κινείται πολύ πιο αργά απ ότι υγρό στο κεντρικό τμήμα.

Η διατμητική τάση που εφαρμόζεται στο ρευστό μπορεί να εκφραστεί ως $\tau = \mu * dc/dy = \mu * \gamma$ ενώ το κινηματικό ιξώδες $\nu = \mu/\rho$

Το ιξώδες μετριέται με διάφορες μεθόδους κυρίως με την χρήση του ιξωδομέτρου, και ανάλογα με τον τύπο της μεθόδου και το είδος του ρευστού που τίθεται προς ανάλυση. Οι μέθοδοι για την πρόβλεψη του αριθμού ιξώδους για πετρελαιοειδή είναι διάφορες ανάλογα τον τύπο της ουσίας, ASTM 975 και ASTM D445 - ISO 3104 .

5.5 Αριθμός Οκτανίου

Πρόκειται για ένα μέτρο που χρησιμοποιείται για τον έλεγχο ποιότητας της βενζίνης. Είναι απαραίτητος δείκτης για την χρήση του καυσίμου αυτού στις μηχανές εσωτερικής καύσης και άλλες βενζινομηχανές.

Δείχνει τον βαθμό της αντικρουστικής ικανότητας του καυσίμου αυτού.

Μεγαλύτερη τιμή του αριθμού Οκτανίων σημαίνει ότι και το καύσιμο μπορεί να αντέξει μεγαλύτερες συνθήκες συμπίεσης μέσα στην μηχανή χωρίς να αναφλεχτεί, και ως εκ τούτου η απόδοση του κινητήρα αυξάνει.

Στην κλίμακα του Οκτανίου το μηδέν αποτελεί το κανονικό επτάνιο και το εκατό το ισοοκτάνιο (2,2,4-τριμεθυλο-πεντάνιο). Η κλίμακα ουσιαστικά είναι η ποσότητα % κατ όγκο ισοοκτανίου σε πρότυπο μείγμα με n-επτάνιο έτσι ώστε το υπό μελέτη δείγμα να παρουσιάζει την ίδια αντικροτική συμπεριφορά με το πρότυπο αυτό.

Ο προσδιορισμός του αριθμού οκτανίου μιας βενζίνης γίνεται με τη βοήθεια ενός πρότυπου κινητήρα, στον οποίο αρχικά μπαίνει η εξεταζόμενη βενζίνη και μετριέται η συμπίεση στην οποία ακούγεται το κτύπημα. Έπειτα, μπαίνει κανονικό επτάνιο στο οποίο προστίθεται ισοοκτάνιο ωσότου ακουστεί το κτύπημα στην ίδια πίεση με εκείνη της εξεταζόμενης βενζίνης. Το επί της εκατό ποσοστό του ισοοκτανίου που υπάρχει στο μείγμα δίνει τον αριθμό οκτανίου της βενζίνης

Τα καύσιμα με μικρότερο αριθμό Οκτανίου αλλά με μεγαλύτερο αριθμό Κετανίου, είναι ιδανικά για μηχανές Diesel καθώς σε αυτές τις μηχανές η καύσιμο εισέρχεται ξεχωριστά στον θάλαμο καύσης (που έχει προθερμανθεί λόγο της συμπίεσης), ενώ έχει προηγηθεί η συμπίεση του αέρα στην μηχανή.

Αντιθέτως στις βενζινομηχανές η συμπίεση του καυσίμου γίνεται ταυτόχρονα με την συμπίεση του αέρα στο τελικό στάδιο της συμπίεσης. Έτσι επιτυγχάνεται μεγάλη απόδοση ενέργειας, και γι αυτό για μια καλή βενζινομηχανή χρειάζεται μεγάλο αριθμό Οκτανίου.

5.6 Φασματοσκοπία IR-NIR

Η υπέρυθρη ακτινοβολία ανακαλύφθηκε το 1800 από τον Sir William Herschel με πειράματα θερμικής ακτινοβολίας. Μετά από μια σειρά πειραμάτων ο Herschel μέτρησε την απορρόφηση της ακτινοβολίας από διάφορες ουσίες.

Έναν αιώνα αργότερα, αφού κατανοήθηκε καλύτερα η φύση της ακτινοβολίας, αναπτύχθηκε η οργανολογία και η θεωρία της Φασματοσκοπίας Υπερύθρου, και οι επιστήμονες ήταν σε θέση να βγάλουν συμπεράσματα και να πάρουν πληροφορίες σχετικά με την μοριακή δομή των οργανικών ενώσεων. Από το 1930 και μετά, οι χημικοί κατανόησαν πλήρως τις ιδιότητες της φασματοσκοπίας και χρησιμοποιήθηκε για οργανική και ποσοτική ανάλυση.[14].

Το φάσμα της ακτινοβολίας που μετρείται στην υπέρυθρη περιοχή είναι μεταξύ $4000 - 400 cm^{-1}$. Οι μοριακές δονήσεις είναι πολλών μορφών: stretching, bending,wagging vibrations,rocking, twisting, scissoring or deformation και πολλές άλλες.

Τα άτομα μιας χημικής ένωσης δονούνται όταν απορροφούν κβάντα υπέρυθρης ηλεκτρομαγνητικής ακτινοβολίας (φωτόνια), που έχουν ένα συγκεκριμένο ποσό ενέργειας. Τα ηλεκτρόνια μεταβαίνουν σε μια υψηλότερη ενεργειακή στάθμη και κατά την αποφόρτιση τους (πρόσπτωση) επανεκπέμπουν φωτόνια διαφορετικής ενεργειακής κατάστασης με τα προσπιπτώμενα. Η επανεκπεμπόμενη ενέργεια των φωτονίων/ηλεκτρομαγνητικής ακτινοβολίας από την εκάστοτε χημική ουσία, είναι συνάρτηση των

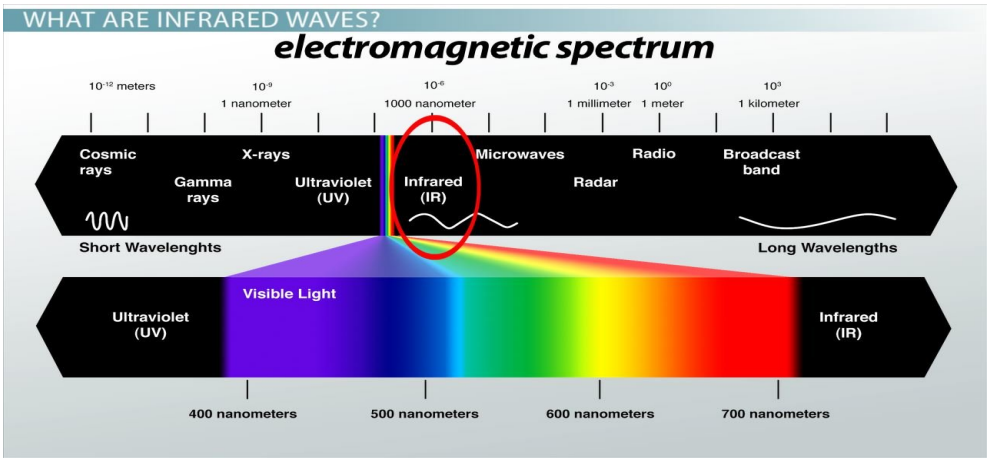


Figure 5.2: Infra Red Spectrum

χημικών δεσμών της ουσίας.

Η ενέργεια που αποβάλλει μια χημική ουσία κατά την μετάπτωση των ηλεκτρονίων στις ενεργειακές στοιβάδες, προσδιορίζεται από την σχέση $E = (\nu + \frac{1}{2})hv$, ενώ η συχνότητα των δονήσεων περιγράφεται $v = \frac{1}{2\pi}\sqrt{\frac{K}{\mu}}$ και είναι χαρακτηριστικό για το φάσμα της ουσίας. Κάθε διαφορετική ουσία παρουσιάζει διαφορετικό είδος φασματοσκοπήματος, ακριβώς λόγω των διαφορετικών δεσμών που τη χαρακτηρίζουν.

Η μελέτη του φασματοσκοπήματος καθορίζει και τα είδη των χημικών ουσιών που υπάρχουν σε ένα μίγμα. Κατά την πρόσπτωση συγκεκριμένου μήκους κύματος ακτινοβολίας, πολύ συγκεκριμένες ουσίες στο μίγμα θα απορροφήσουν την ακτινοβολία αυτή. Η απορρόφηση αυτή καταγράφεται και παρουσιάζεται σε ένα διάγραμμα (wavelength – Absorption) για ένα τυπικό φασματογράφημα όπως αυτό την παρούσα εργασία.

Η φασματοσκοπία στο εγγύς (Near-IR) ηλεκτρομαγνητικό φάσμα, είναι μεταξύ 700nm -2500nm. Βασίζεται κι αυτή με την σειρά της στην θεωρία των μοριακών δονήσεων, και είναι ίδιας φιλοσοφίας με την τεχνική της IR φασματοσκοπίας.

Η NIR φασματοσκοπία έχει ένα αρκετά μικρό εύρος έκτασης σε μήκη κύματος. Πλεονέκτημα της είναι ότι έχει καλύτερη ικανότητα διείσδυσης σε σχέση με την IR φασματοσκοπία. Χαρακτηριστικά δεν είναι πολύ ευαίσθητη τεχνική, αλλά μπορεί να δώσει καλά αποτελέσματα για δεδομένα μιας πρόχειρης δειγματοληψίας με μικρή ή καθόλου προ επεξεργασία δείγματος.

Η μεθοδολογία της NIR πολλές φορές είναι δύσκολο να δώσει ξεκάθαρα αποτελέσματα για την χημειοσύνθεση μιας άγνωστης ουσίας, αλλά με τις μαθηματικές τεχνικές που εφαρμόζονται στην παρούσα εργασία, γίνεται εφικτό να ληφθούν οι επιθυμητές χημικές πληροφορίες.

Chapter 6

ΑΝΑΠΤΥΞΗ ΑΛΓΟΡΙΘΜΟΥ ΠΡΟΒΛΕΨΗΣ

Ο παρκάτω αλγόριθμος αναπτύχθηκε για την πρόβλεψη της ιδιότητας του Αριθμού Κετανίου. Εν συνεχεία χρησιμοποιήθηκε για την πρόβλεψη και των άλλων ιδιοτήτων του Κινηματικού Ιξώδους και του Αριθμού Οκτανίων.

Σκοπός κατά την επεξεργασία των δεδομένων, είναι να βρεθεί με σχετική ακρίβεια το διάνυσμα (X)-Διάνυσμα παλινδρόμησης, το οποίο θα πολλαπλασιαστεί με το διάνυσμα των τιμών των δεδομένων Απορρόφησης και θα προβλέπει την εκάστοτε Ιδιότητα. Για την προσαρμογή των δεδομένων του διανύσματος των τιμών Απορρόφησης στον πολυχώρο των Κυρίων Συνιστωσών, θα πρέπει υπολογιστεί και το Διάνυσμα Βαρών το οποίο θα πρέπει πρώτα να πολλαπλασιάστεί με το διάνυσμα Απορρόφησης του αγνώστου δείγματος.

6.1 Εισαγωγη δεδομένων

Ο παρακάτω αλγόριθμος εισάγει τα δεδομένα από ένα αρχείο excel. Το αρχείο αυτό είναι μορφής xlsx. Τα δεδομένα του αριθμού κετανίου δίνονται στην πρώτη γραμμή από την δεύτερη στήλη και μετά. Ο κυματάριθμος στην πρώτη στήλη από την δεύτερη γραμμή και τα δεδομένα Οπτής Διαπερατότητας αποτελούν τον υπόλοιπο χώρο του αρχείου. Επίσης εισάγονται και τα ονόματα των δειγμάτων.

```
%% [SCRIPT 1] PLS & PCR code according to the Matlab example of octane
    number prediction
% Data tips: first column must be wavelength vector
% starting from second box. First row must be the dependent variable of
    the
% problem starting from the sendond box.
% Import IR data
clc
close all
clear all
```

```
9
10    [filename, filepath] = uigetfile({'*.xlsx'; '*.xls'}, 'select data
          file');
11    [Data,Labels] = xlsread(strcat(filepath,filename));
12    CN=Data(1,2:end)';
13    wavenumber = Data(2:end,1)';
14    Transmittance = Data(2:end,2:end)';
15
16 clearvars -except Transmittance wavenumber CN  Data Labels
17 format
```

6.2 Διάγραμμα Οπτής Διαπερτότητας - Επιλογή Μεταβλητών

Είναι σημαντικό να γίνει μια παρουσίαση των δεδομένων του φασματοσκοπήματος, σε ένα διάγραμμα ώστε να γίνει μια οπτή αξιολόγηση. Γι αυτό είναι απαραίτητο να γραφεί ο αλγόριθμος που κατασκευάζει το διάγραμμα αυτό.

Επιπλέον, ο αλγόριθμος δίνει τη δυνατότητα στο χρήστη να επιλέξει ποιες μεταβλητές θέλει να χρησιμοποιήσει. Η επιλογή αυτή, όπως προαναφέρθηκε, ονομάζεται Variable Selection και είναι πολύ σημαντική για την αξιοπιστία του αποτελέσματος, δηλαδή την ακρίβεια της πρόβλεψης του αριθμού κετανίου.

Ο αλγόριθμος παρατίθεται παρακάτω

```
1  %% [SCRIPT 2] Variable Selection of Transmittance Data
2  % At this script user is able to select a specific range of
       Transmittance
3  % data.
4  % plot the whole range of Transmittance
5
6      plot(wavenumber,Transmittance')
7  (' Wavenumber Data Range starts from');
8      Data(2,1)
9  ('To')
10     Data(end,1)
11
12 grid minor
13 title(' Transmittance ')
14 xlabel('wavenumber')
15 ylabel('Transmittance')
16
17 % User may insert the desirable range of data.
18 %If the initial data is ok , this step can be negleted.
```

```
19
20    From=input('Insert Starting wavenumber');
21    To=input('Insert Final wavenumber');
22  figure()
23  clc
24
25  % Plot the new-range of Transmittance Data
26
27        wavenumber2=Data(Data(1:end,1)>=From & Data(1:end,1)<=To);
28        [o p]=size(wavenumber2);
29        Transmittance2=Data(find(Data(1:end,1)>=From & Data(1:end,1)<=To)
              ,:);
30        Transmittance2(:,1)=[];
31
32  plot(wavenumber2,Transmittance2);
33  grid minor
34  title(' Transmittance ')
35  xlabel('wavenumber')
36  ylabel('Transmittance')
37
38        Transmittance=Transmittance2';
39        wavenumber=wavenumber2;
40
41  clearvars -except Absorbance Transmittance wavenumber CN  Data Labels
```

Από το διάγραμμα αυτό γίνεται εμφανές ότι οι μετρήσεις στα πρώτα Wavenumbers δεν αποτελούν αντικειμενικές μετρήσεις Οπτής Διαπερατότητας, καθώς οι τιμές ξεπερνούν ακόμα και το 100% σε κάποια δείγματα, ενώ παρατηρείται πολύ μεγάλος θόρυβος.

Αυτό μπορεί να είναι αποτέλεσμα κακής βαθμονόμησης (Calibration) του οργάνου μέτρησης. Συνήθως τέτοια δεδομένα υπάρχουν σε ουσίες που παρουσιάζουν φωταύγεια σε κάποια μήκη κύματος.

Το διάγραμμα κατασκευάστηκε με την ιδιότητα να τοποθετεί τα δείγματα με αυξανόμενο αριθμό κετανίου CN.

Παρακάτω παρουσιάζονται οι εικόνες από το φάσμα και τα τμήματα του φάσματος που θα μελετηθούν.

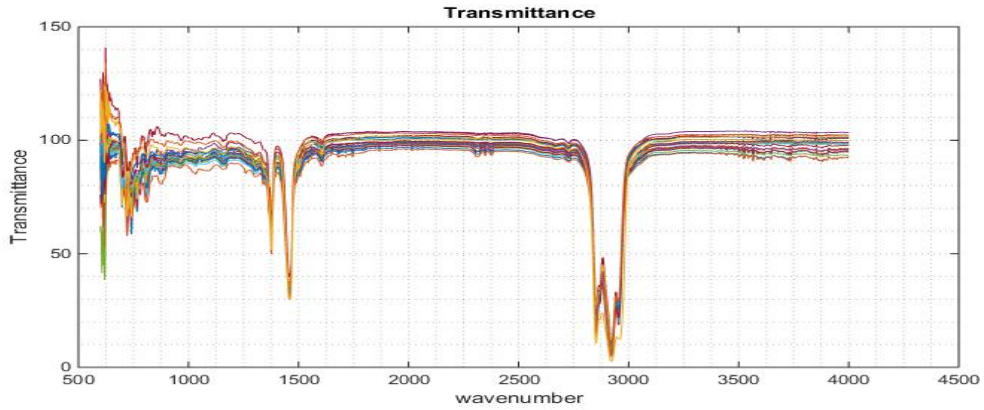


Figure 6.1: All- Spectra Variables

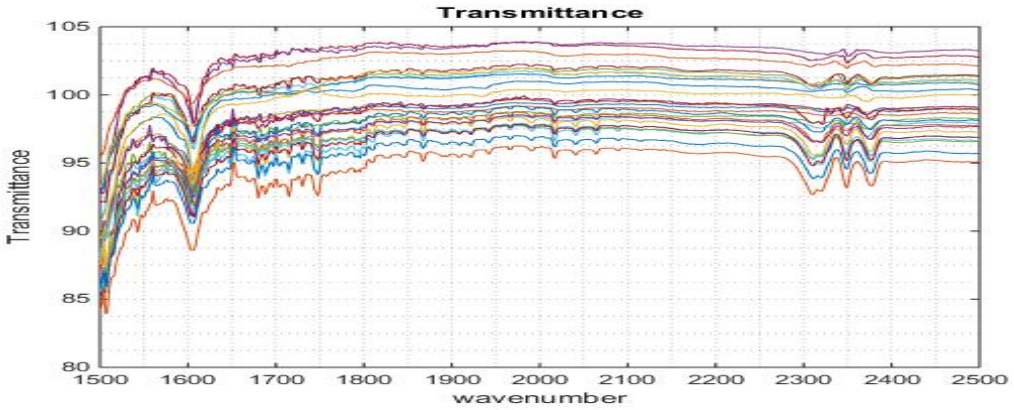


Figure 6.2: 1500-2500 - Spectra Variables

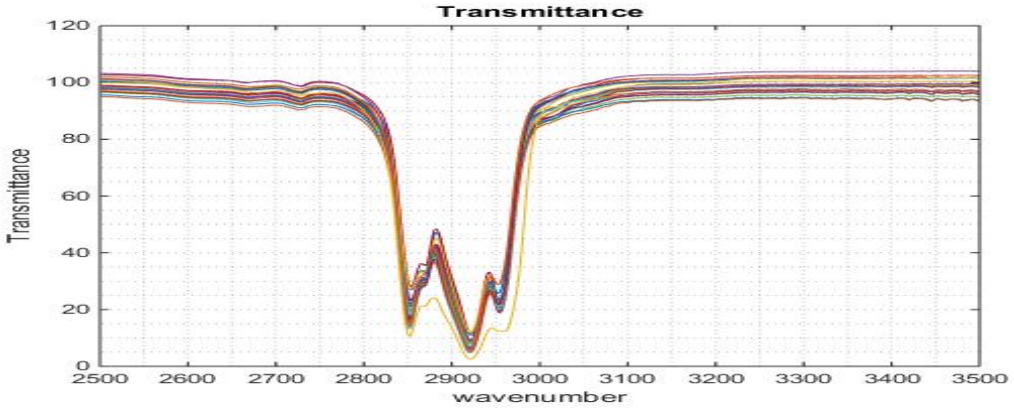


Figure 6.3: 2500-3500 - Spectra Variables

6.3 Μετατροπή δεδομένων Οπτής Διαπερατότητας σε Απορρόφησης

Για την επεξεργασία των δεδομένων μετατρέπονται τα δεδομένα Οπτής Διαπερατότητας, σε δεδομένα Απορρόφησης.

Η Οπτή Διαπερατότητα είναι ουσιαστικά το ποσοστό της ηλεκτρομαγνητικής ακτινοβολίας που διαπερνά ένα μέσο (εδώ δείγμα Diesel) και εξέρχεται από αυτό. Η υπόλοιπη ακτινοβολία είτε απορροφάται είτε ανακλάται. Είναι σημαντικό τα δεδομένα να μετατραπούν σε απορρόφησης για την καλύτερη επεξεργασία και τη μελέτη τους.

Για τη μετατροπή των δεδομένων απορρόφησης σε δεδομένα Οπτής Διαπερτότητας, εφαρμόζεται η παρακάτω σχέση που προκύπτει από τον νόμο του Beer.

$$A = log(100/\%T)$$

Ο αλγόριθμος της Matlab για την μετατροπή, αλλά και την παρουσίαση των αποτελεσμάτων σε ένα διάγραμμα, παρουσιάζεται εδώ.

```
%% [SCRIPT 3] Transmittance converted to Absorbnace
clc
close all

% Transmittance=(Transmittance-mean(mean((Transmittance))/std(
    Transmittance)));
% This code will convert Transmittance to Absorbance according to Beers
    Law

for i=1:size(Transmittance,2)
    for j=1:size(Transmittance,1)
        X1(j,i) = log(100/(Transmittance(j,i)));
    end
end

    [dummy,h2]=sort(CN);
    Absorbance=X1;
    oldorder=get(gcf,'defaultAxesColorOrder');
    set(gcf,'DefaultAxesColorOrder',jet(size(Absorbance,1)));

% This is the plot of the Absorbance

plot3(wavenumber,repmat(CN(h2),1,size(Absorbance,2))',Absorbance(h2,:)
    ')
```

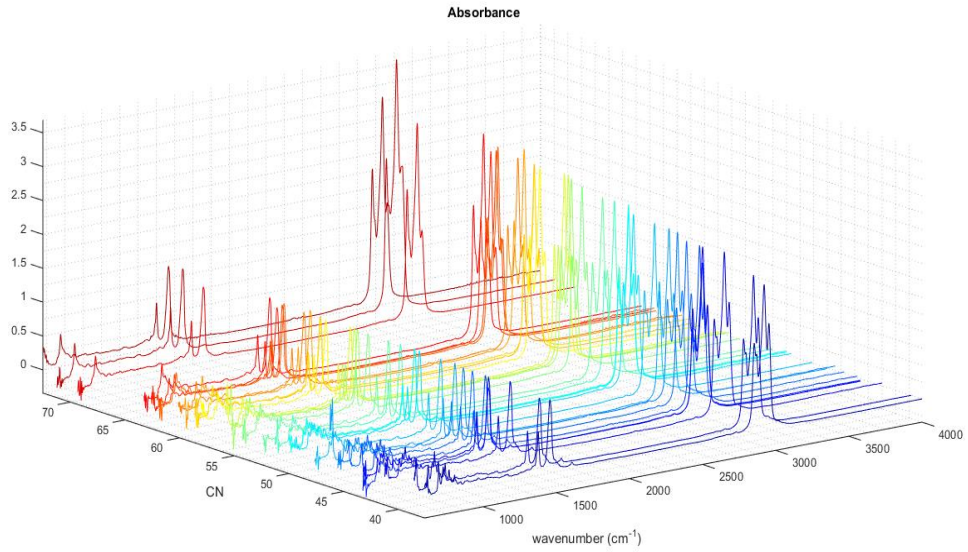


Figure 6.4: Absorbance

```
22 title('{\bf Absorbance}');xlabel('wavenumber (cm^{-1})'); ylabel('CN');
       axis('tight');
23 grid minor
24
25 clearvars -except Absorbance wavenumber CN  Data Labels
```

Το διάγραμμα (Absorbance) παρουσιάζει σφάλματα σχετικά με τις τιμές της απορρόφησης, καθώς ορισμένες τιμές ξεπερνούν κατά πολύ το 100%. Αυτό οφείλεται πιθανώς σε λάθος κάτω ή άνω ορίου ανίχνευσης απορρόφησης, ή λάθη κατά την μέτρηση. Εδώ, για την απαλοιφή των σφαλμάτων αυτών θα δράσει το επόμενο πρόγραμμα το οποίο σκοπό έχει την κανονικοποίηση των δεδομένων (Standardization).

6.4 Επεξεργασία δεδομένων - Data pretreatment

Η Κανονικοποίηση των δεδομένων είναι ένα πολύ χρήσιμο εργαλείο στη στατιστική ανάλυση των δεδομένων.

Σε ένα πολυδιάστατο σύνολο δεδομένων, η μεθοδολογία της PCA -όπως έχει προαναφερθεί- έχει ως σκοπό να μεταφέρει την πληροφορία/διακύμανση στις κύριες συνιστώσες οι οποίες έχουν αρχικά και τη μεγαλύτερη διακύμανση. Έτσι, κατά την επεξεργασία των δεδομένων, μπορεί να έχει πολύ μεγάλη επίδραση σε κάποιες μεταβλητές και λιγότερη σε άλλες.

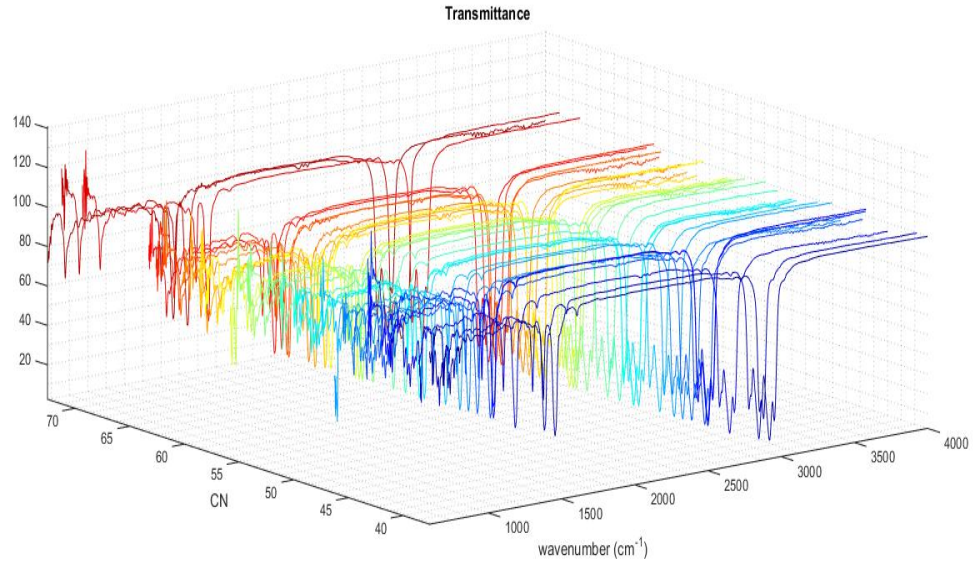


Figure 6.5: Transmittance

Εάν επιθυμείται να μην υπάρχει αυτό το πρόβλημα, και η δράση της PCA να είναι ανεξάρτητη από αυτήν την κλιμάκωση των δεδομένων, θα πρέπει να γίνει κανονικοποίηση των δεδομένων. Εάν η κλιμάκωση αυτή των δεδομένων αλλοιώνει το αποτέλεσμα τότε δεν θα πρέπει να γίνει κανονικοποίηση.

Η κανονικοποιημένες τιμές προκύπτουν από την σχέση :

$$StandardizedValues = (x - mean(X))/std(X))$$

Οι κανονικοποιημένες τιμές παρουσιάζουν μορφή κατανομής Gauss. Στο πρόγραμμα αυτό κατασκευάζονται 2 διαγράμματα κατανομής των δεδομένων με κανονικοποίηση και χωρίς.

Αρχικά παρουσιάζεται ο αλγόριθμος που απαλείφει 2 τιμές σύφωνα με την αξιολόγηση των δεδομένων από τον παρατηρητή, οι οποίες παρουσιάζουν μεγάλα σφάλματα, δηλαδή τις τιμές (13,31)
Στη συνέχεια δίνεται ο αλγόριθμος που κανονικοποιεί τα δεδομένα της απορρόφησης.

Στο πρόγραμμα συμπεριλαμβάνεται ένα κομμάτι το οποίο κάνει ομαλοποίηση στα δεδομένα (Normalization).

Θα πρέπει να διευκρινιστεί η διαφορά ανάμεσα στην κανονικοποίηση και την ομαλοποίηση, ώστε να μην συσχετίζονται.

Η Ομαλοποίηση στη στατιστική ανάλυση και στη γραμμική άλγεβρα, έχει αρκετές έννοιες και εφαρμογές. Στις πιο απλές εφαρμογές, με ομαλοποίηση των τιμών, γίνεται εφικτό να γίνει μια προσαρμογή των τιμών που ανήκουν σε διαφορετικές κλίμακες, σε μια ≪πλασματική≫ κοινή κλίμακα.

Σε πολλές εφαρμογές η ομαλοποίηση έχει πρόθεση να φέρει το σύνολο των κατανομών πιθανότητας των προσαρμοσμένων τιμών σε ευθυγράμμιση, τείνοντας σε μια κανονική κατανομή.

```
%% [SCRIPT 4] RECOMENDED - DELETE SAMPLES TRANSMITTANCE/CN( 13,:) &
    (31,:)
close all
clc

     Absorbance(13,:)=[];
     Absorbance(30,:)=[];
     CN(13,:)=[];
     CN(30,:)=[];
     Labels(13)=[];
     Labels(30)=[];

     [n,m]=size(Absorbance );
     [dummy,h2]=sort(CN);
oldorder=get(gcf,'defaultAxesColorOrder');
set(gcf,'DefaultAxesColorOrder',jet(size(Absorbance,1)));

plot3(wavenumber,repmat(CN(h2),1,size(Absorbance,2))',Absorbance(h2,:)
    ')
title('{\bf Absorbance}');xlabel('wavenumber (cm^{-1})'); ylabel('CN');
     axis('tight');
grid on

clearvars -except Absorbance wavenumber CN  Data Labels
```

```

%% [SCRIPT 5] Pretreatment of Absorbance Data
% This step is recommended.
clc
close all

% Calculation of Mean Value & Standard Deviation Vectors
Mean_of_Absorbance =mean(Absorbance );
Standard_Deviation_of_Absorbance =std(Absorbance);

% [1] Standardization
Standardised_Data_Matrix=(Absorbance-repmat(Mean_of_Absorbance ,[(size(
    Absorbance ,1)) 1]))./repmat(Standard_Deviation_of_Absorbance ,[(
    size(Absorbance ,1)) 1]);

% [2]  Absorbance/max
%Scaled_Data=(Standardised_Data_Matrix/max(max(Standardised_Data_Matrix
    )));
```

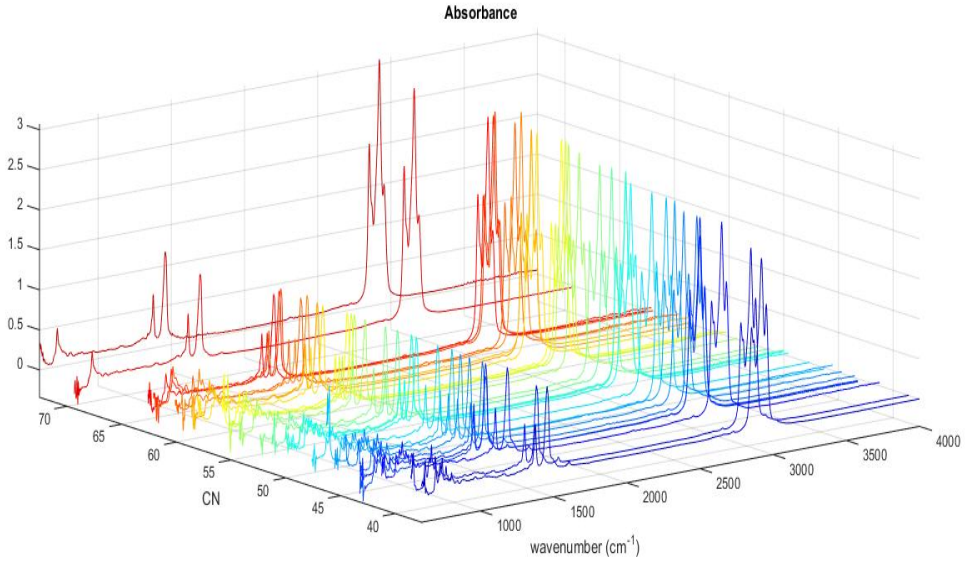


Figure 6.6: Absorbance with Deleted Samples

```
16
17 % [3]  Normalizing
18 %Normalized_Data=(Absorbance-min(min(Absorbance)))/((max(max(Absorbance
       ))-(min(min(Absorbance)))));
19
20 % [4] otherwise we can use the below function to Standardize
21 %Standardised_Data_Matrix=zscore(Absorbance );
22
23 % Change this script
24 Training_Set=Standardised_Data_Matrix;
25 %Training_Set=Scaled_Data;
26
27 clearvars -except Absorbance wavenumber CN  Data Labels Training_Set
```

Μετά την επεξεργασία των δεδομένων, σημαντικό είναι να παρασταθούν σε ένα διάγραμμα, τα καινούρια δεδομένα. Το διάγραμμα των επεξεργασμένων αυτών δεδομένων, όπως και το προηγούμενο με τα αρχικά δεδομένα, κατασκευάστηκε με αυξανόμενο αριθμό κετανίου.

```
1  %% [SCRIPT 6] Plot Training Absorbance Data, in ascending order
2 %  Scaling of Absorbance Data
3 clc
4 close all
5
```

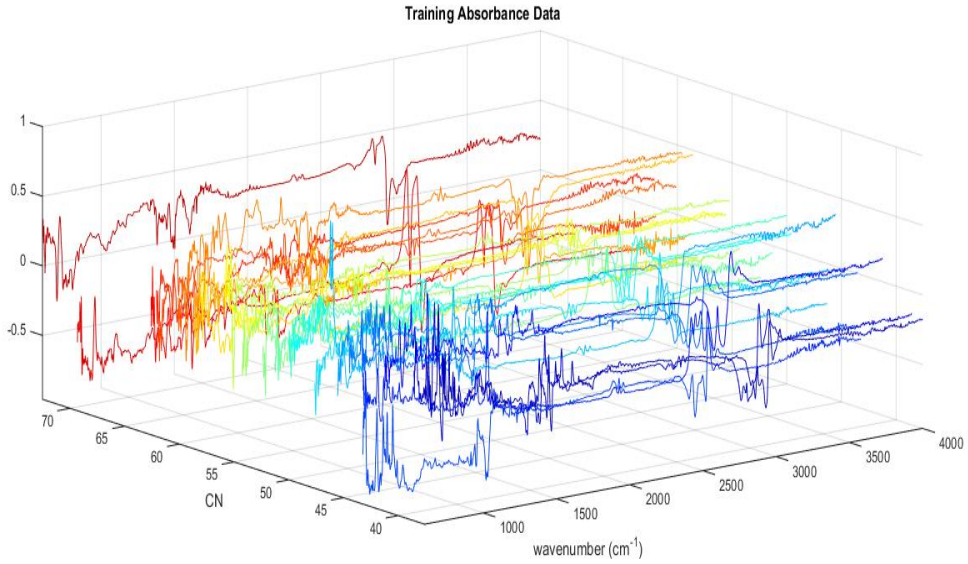


Figure 6.7: Standardized -Scaled Samples

```
    [~,h]=sort(CN);
    oldorder=get(gcf,'defaultAxesColorOrder');
    set(gcf,'DefaultAxesColorOrder',jet(size(Training_Set,1)));

% plot3(wavenumber,CN, Transmittance)
% Training_Set.

plot3(wavenumber,repmat(CN(h),1,size(Training_Set,2))',(Training_Set(h
    ,:))')
xlabel('wavenumber (cm^{-1})'); ylabel('CN'); axis('tight');
title('Training Absorbance Data')
grid on

Training_Set=Training_Set/max(max((Training_Set)));
clearvars -except Absorbance wavenumber CN Data Labels Training_Set
```

Το διάγραμμα είναι δύσκολο να ερμηνευτεί λόγω της πολυπλοκότητάς του. Μπορούν όμως, να σημειωθούν κάποια σημεία, όπου οι μετρήσεις θα μπορούσαν να χαρακτηριστούν προβληματικές.

6.5 Μεθοδος PCR- Διάγραμμα PCA

Όπως αναφέρθηκε και στο θεωρητικό μέρος, το scatter των πρώτων δύο κύριων συνιστωσών (με τις δύο μεγαλύτερες τιμές στις διακυμάνσεις των στηλών του πίνακα δεδομένων), δηλαδή τα PC1 και PC2 δίνει το διάγραμμα PCA.

Ο υπολογιστικός αλγόριθμος που δίνεται στη συνέχεια, υπολογίζει με την εντολή princomp της Matlab, τις παραμέτρους [COEFF SCORE LATENT] όπου το COEFF θα βοηθήσει στο να υπολογιστούν οι τιμές των Κυρίων Συνιστωσών.

Υπολογισμός Διακύμανσης και ποσοστό πληροφορίας που παραμένει υπολογίζονται αναλόγως.

```
%% [SCRIPT 7] PCA plot-PCR at Standardised Data
clc
close all

% finding the eigenfunctions of the
% sample covariance matrix, to calculate the coefficients
% of the principal components (V). The diagonal elements
% of D, store the variance of the respective
% principal components.
% Use of  below function

     [V, D] = eig(cov(Training_Set));
     Principal_Components=diag(D);

%The coefficients and respective variances
%of the principal components could also be found
%using the princomp function
    [COEFF, SCORE, LATENT] = princomp(Training_Set);

%To calculate the principal components
    PCs=Training_Set*COEFF;

%Original_Data=((Training_Set*COEFF)*COEFF').*repmat(
    Standard_Deviation_of_Absorbance ,[(size(Absorbance ,1)) 1]) +
    repmat(Mean_of_Absorbance ,[(size(Absorbance ,1)) 1]);
% set PCs in order according to Variance
    [o,p]=size(PCs);
    A=sort(var(PCs));

% Variance of PC1 ,PC2
    total_Variance=sum(var(PCs));
    Variance_PC1=sprintf('%.1f%%',round((round(A(p)*100/total_Variance))
        ,1))
    Variance_PC2=sprintf('%.1f%%',round((round(A(p-1)*100/total_Variance
```

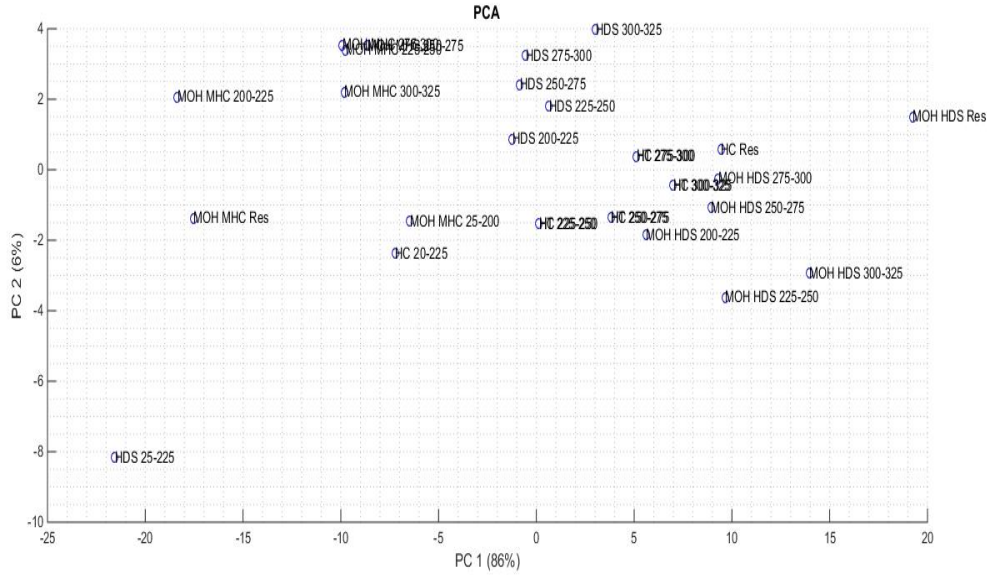


Figure 6.8: PCA Plot(ALL-29-St-Sc)

```
        ))),1))
32
33
34  % PCA plot with Labels
35  scatter(PCs(1:end,1),PCs(1:end,2))
36  text(PCs(1:end,1),PCs(1:end,2),Labels(1,:),'ro')
37  title('{\bf PCA}'); xlabel('PC 1'); ylabel('PC 2')
38  grid minor
39
40  clearvars -except p COEFF PCs Absorbance wavenumber CN Data Labels
        Training_Set
```

Η μεθοδολογία της PCA μεταφέρει στα 2 πρώτα PCs το 92% της συνολικής πληροφορίας.

6.6 Cross Validation

Το επόμενο κομμάτι του κώδικα συγκρίνει τη μεθοδολογία της μεθόδου PLS με την PCR στο ίδιο διάγραμμα και προσαρμόζει ευθεία ελαχίστων τετραγώνων στα ανάλογα σημεία.

Το πρόγραμμα αρχικά δίνει στο χρήστη να επιλέξει τον αριθμό των Κυρίων Συνιστωσών που επιθυμεί να χρησιμοποιήσει η μέθοδος και ύστερα υπολογίζει το συντελεστής συσχέτησης και παραθέτει ποια μεθοδολογία παρουσιάζει το λιγότερο σφάλμα.

Παρατηρείται ότι όσες περισσότερες Κύριες Συνιστώσες χρησιμοποιεί το πρόγραμμα τόσο καλύτερα προσαρμόζονται τα σημεία πάνω στις ευθείες ελαχίστων τετραγώνων. Ο υντελεστής συσχέτησης τείνει στη μονάδα και οι δύο ευθείες τείνουν να συσχετιστούν. Η επιλογή βέλτιστης επιλογής αριθμού Κυρίων Συνιστωσών παρουσιάζεται στη συνέχεια. Ο αλγόριθμος ο οποίος κατασκευάζει το διάγραμμα είναι ο παρακάτω.

```
%% [SCRIPT 8] Cross Validation: PCR vs PLSR
%  Fitting No of components manually
% As more components are added in PCR , it will necessarily do a better
    job
% of fitting original data |y|, simply because at some point most of
    the
% important predictive in formation in |X| will be present in the
    principal
% components.

clc
close all

    NumberofComponents=input('enter number of principal components=');
    [Xloadings,Yloadings,Xscores,Yscores,betaPLSn,PLSctVar]=plsregress(
        Training_Set,CN,NumberofComponents);
    [PCALoadings,PCAScores,PCAVar]=pca(Training_Set,'Economy',false);

    yfitPLSn=[ones(size(Training_Set,1),1) Training_Set]*betaPLSn;
    betaPCRn=regress(CN-mean(CN),PCAScores(:,1:NumberofComponents));
    betaPCRn=PCALoadings(:,1:NumberofComponents)*betaPCRn;
    betaPCRn=[mean(CN)-mean(Training_Set)*betaPCRn; betaPCRn];
    yfitPCRn=[ones(size(Training_Set,1),1) Training_Set]*betaPCRn;

plot(CN,yfitPLSn,'bo',CN,yfitPCRn,'rx');
title('Fitting more Components')
xlabel('observed Response');
ylabel('Fitted Response');
lsline
legend({'PLSR ' 'PCR '},'location','NW');

```

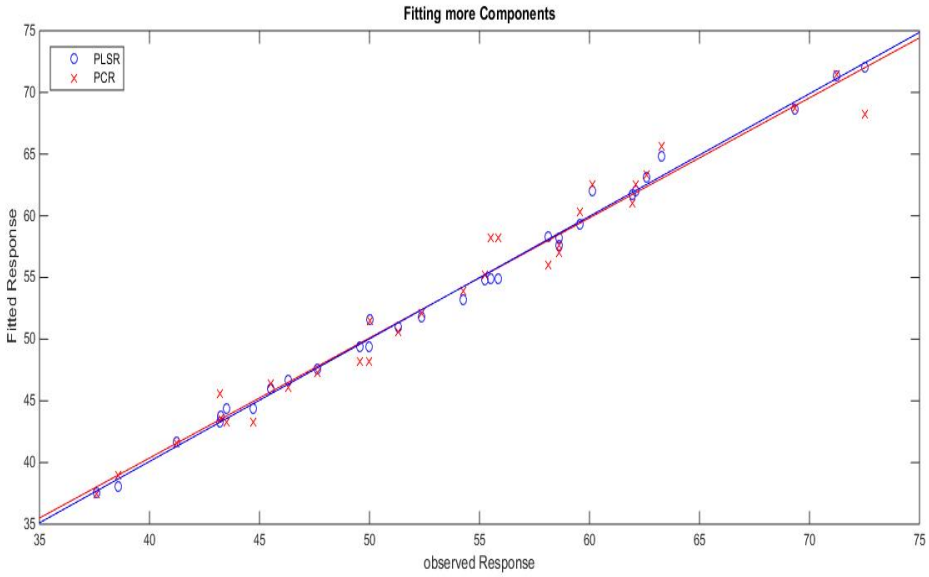


Figure 6.9: Cross Validation PCR-PLS (10 PCs-ALL-31-St)

```
% Correlation_Coeficient

   TSS=sum((CN-mean(CN)).^2);
   RSS_PLS=sum((CN-yfitPLSn).^2);
    Correlation_Coeficient=1-RSS_PLS/TSS
   RSS_PCR=sum((CN-yfitPCRn).^2);
    Correlation_Coeficient=1-RSS_PCR/TSS

clearvars -except Xscores PCAScores n yfitPCRn Yscores betaPLSn PCs
    Absorbance wavenumber CN  Data Labels Training_Set
```

Μία ακόμα μέθοδος για την σύγκριση των δύο μεθόδων είναι να κατασκευαστούν διαδιαγράμματα της μεταβλητής απόκρισης (CN), σε σχέση με τα αντίστοιχα Scores.

```
%%  [SCRIPT 9]
% Another way to compare the efficiency of the prediction
% bettwin the PLSR and PCR.
% plot the response variable against the 2 predictors in both cases.

clc
close all

% PLSR PLOT
plot3(Xscores(:,1),Xscores(:,2),CN-mean(CN),'bo');
legend('PLSR');
grid on; view(-30,30);

figure()
plot3(PCAScores(:,1),PCAScores(:,2),CN-mean(CN),'rx')
legend('PCR')
grid minor ; view(-30,30);

clearvars -except yfitPCRn PCAScores n Yscores betaPLSn PCs Absorbance
    wavenumber CN  Data Labels Training_Set
```

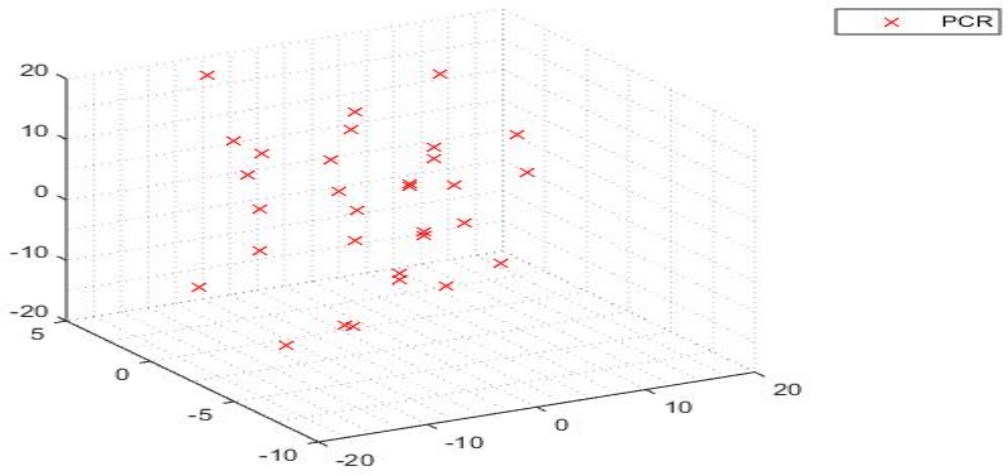


Figure 6.10: PCR

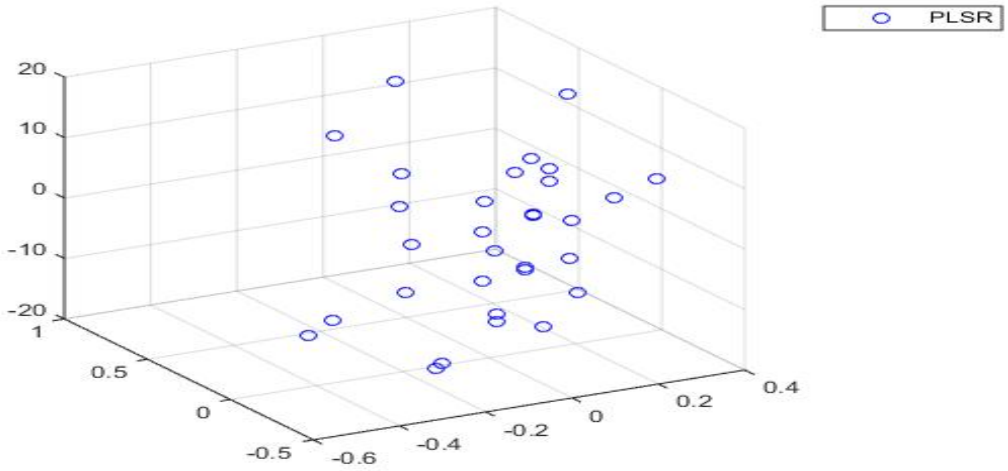


Figure 6.11: PLSR

6.7 Διακύμανση εκφρασμένη με τον αριθμό των Κυρίων Συνιστωσών, για μεθοδολογίες PCA - PLS

Ο παρακάτω αλγόριθμος έχει σκοπό να συγκρίνει και να παρουσιάσει σε μια γραφική παράσταση την αθροιστική διακύμανση που υπάρχει κατά τις μεθοδολογίες των PCA και PLS.

Αυξάνοντας τον αριθμό των Κυρίων Συνιστωσών, η αθροιστική διακύμανση - όπως είναι λογικό - αυξάνεται κι αυτή καθώς στο σύστημα εισέρχεται μεγαλύτερο ποσοστό της συνολικής πληροφορίας και στις δύο περιπτώσεις.

Το διάγραμμα αυτό βοηθάει στο να ξεχωρίσει η μέθοδος η οποία χρησιμοποιεί μεγαλύτερο ποσοστό πληροφορίας, για το καλύτερο δυνατό αποτέλεσμα.

```
%% %% [SCRIPT 10] Plot: Variance Explained in Training_Set  vs Number of Principal Components
% This program presents the Variance calculating by PCR and PLS
% simultaneusly.
clc
close all

    principalcomponents=input('add Number of principal components=')

    [PCALoadings, PCAScores,PCAVar]=pca(Training_Set ,'Economy' ,false);
    [Xloadings, Yloadings, Xscores, Yscores, betaPLS10, PLSPctVar]=plsregress(Training_Set ,CN, principalcomponents);

plot(1:principalcomponents,100*cumsum(PLSPctVar(1,:)),'b-o');
xlabel('Number of Principal Components');
ylabel('Percent Variance Explained in Training_Set ');
legend({'PLSR'}, 'location','SE');
grid on

hold on
     [LOADINGS SCORE LATENT]=princomp(Training_Set);
     x=var(SCORE);
     y=100*x/(sum(var(SCORE)));

plot((1:principalcomponents),cumsum(y(1,1:principalcomponents)),'r-x')
grid minor
xlabel('Number of PCs')
ylabel('persentage of total variance')

clearvars -except Xloadings Yloadings Xscores Yscores betaPLS10 PLSPctVar PCAVar PCALoadings yfitPCRn PCAScores n Absorbance wavenumber CN  Data Labels Training_Set
```

6.8 Επιλογή κατάληλλου αριθμού κύριων συνιστωσών για μείωση σφάλματος

Η επιλογή του αριθμού των Κυρίων Συνιστωσών είναι ένας πολυσήμαντος παράγοντας για την επεξεργασία τέτοιων πολυμεταβλητών συστημάτων. Όταν γίνεται η κατασκευή των Z-Scores ενός πίνακα δεδομένων οι στήλες του πίνακα παρατάσσονται με σειρά μειούμενης διακύμανσης. Αυτό σημαίνει ότι οι πρώτες στήλες έχουν και το μεγαλύτερο ποσοστό της συνολικής διακύμανσης. Έτσι, στις μεθόδους αυτές γίνεται χρήση των πρώτων στηλών των Z-Scores.

Ένας μικρός αριθμός από κύριες συνιστώσες χρησιμοποιεί μικρό ποσοστό της συνολικής πληροφορίας του συστήματος με κίνδυνο να επιφέρει κακά αποτελέσματα, αλλά ταυτόχρονα μειώνει τον υπολογιστικό χρόνο. Το αντίθετο συμβαίνει με μεγάλο αριθμό Κυρίων Συνιστωσών.

Το παρακάτω πρόγραμμα παρουσιάζει τον αριθμό των Κυρίων Συνιστωσών σε συνάρτηση με το μέσο τετραγωνικό σφάλμα που προκύπτει. Γίνεται η επιλογή του μικρότερου αριθμού Κυρίων Συνιστωσών με το μικρότερο δυνατό σφάλμα.

```
%% [SCRIPT 11] CHOOSE THE NUMBER OF COMPONENTS
% This part of the code is to CHOOSE THE NUMBER OF COMPONENTS for the
%minimization of the expected error. This is about to prevent an
%overffiting problem.(Required).

% [plsregress] has an option to estimate the mean squared
% prediction error (MSEP) by cross-validation,
% in THIS CASE USING 10-FOLD CV)
close all
clc

    Number_of_PCS=input('Number of Components=')
    [X1,Y1,Xs,Ys,beta,pctVar,PLSmsep]=plsregress(Training_Set ,CN,
        Number_of_PCS,'CV',Number_of_PCS);

plot(0:Number_of_PCS,PLSmsep(2,:),'b-o');
title('Choosing Number of PCs');
xlabel('Number of components');
ylabel('Estimated Mean Squared Prediction Error');
legend({'PLSR'},'location','NE');
grid minor

% Calculate the minimum error by PLSR & PLS Method
    min_PLSmsep=min(PLSmsep(2,:))

hold on
% PCR_ERROR  This calculates error for more PCs for PCR

```

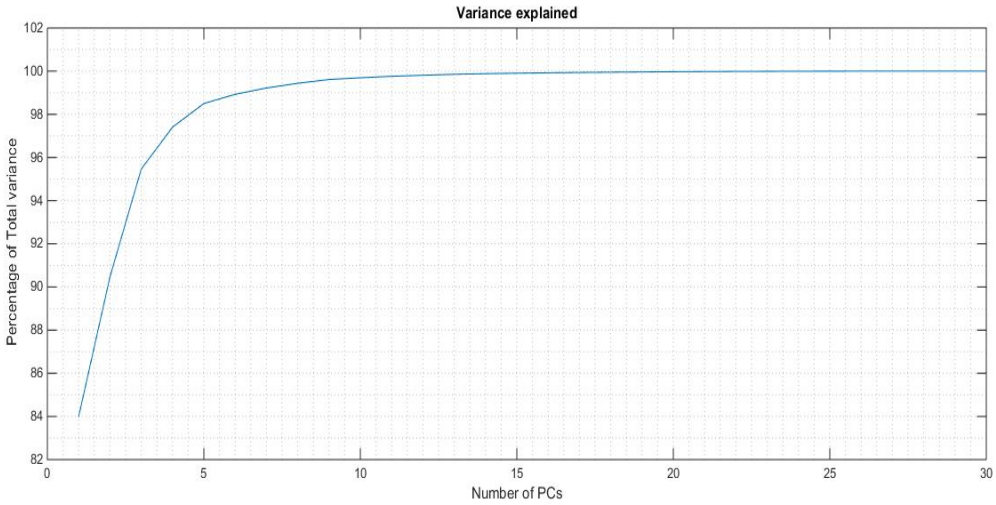


Figure 6.12: PCA Variance Explained (All-31-St)

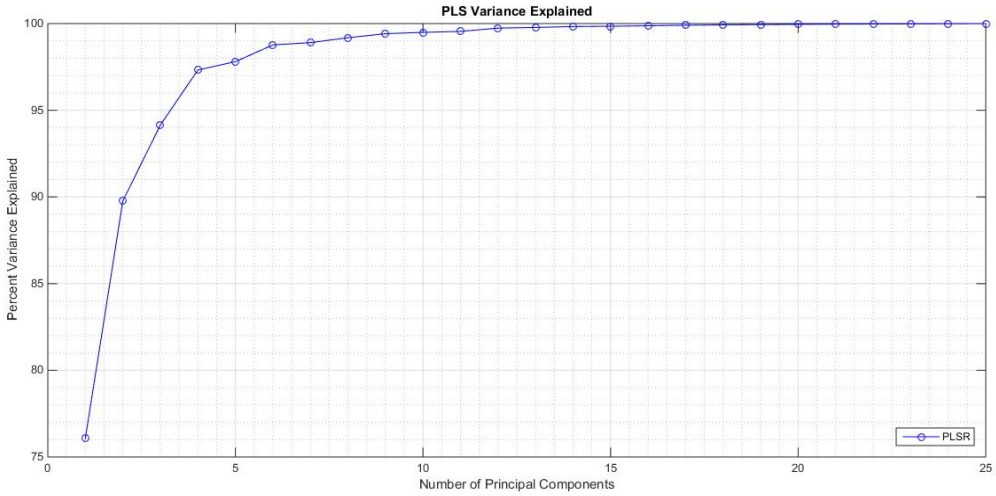


Figure 6.13: PLS Variance Explained(All-31-St)

```
28 for i=1:Number_of_PCS;
29 [PCALoadings, PCAScores,PCAVar]=pca(Training_Set,'Centered',false,'
       Economy',false,'NumComponents',i);
30 betaPCRn=regress(CN-mean(CN), PCAScores(:,1:i));
31 yfitPCRn=PCAScores*betaPCRn+mean(CN);
32 TSS=sum((CN-mean(CN)).^2);
33 RSS=sum((CN-yfitPCRn).^2);
34 Correlation_Coeficientn(i)=RSS/TSS;
35 end
36
37 plot(0:Number_of_PCS-1,100*Correlation_Coeficientn,'r-^');
38 legend({'PLS','PCR'},'location','NE');
39
40
41 clearvars -except X1 Y1 Xs Ys beta pctVar PLSmsep PCALoadings PCAScores
       PCAVar Number_of_PCS n Absorbance wavenumber CN Data Labels
      Training_Set
```

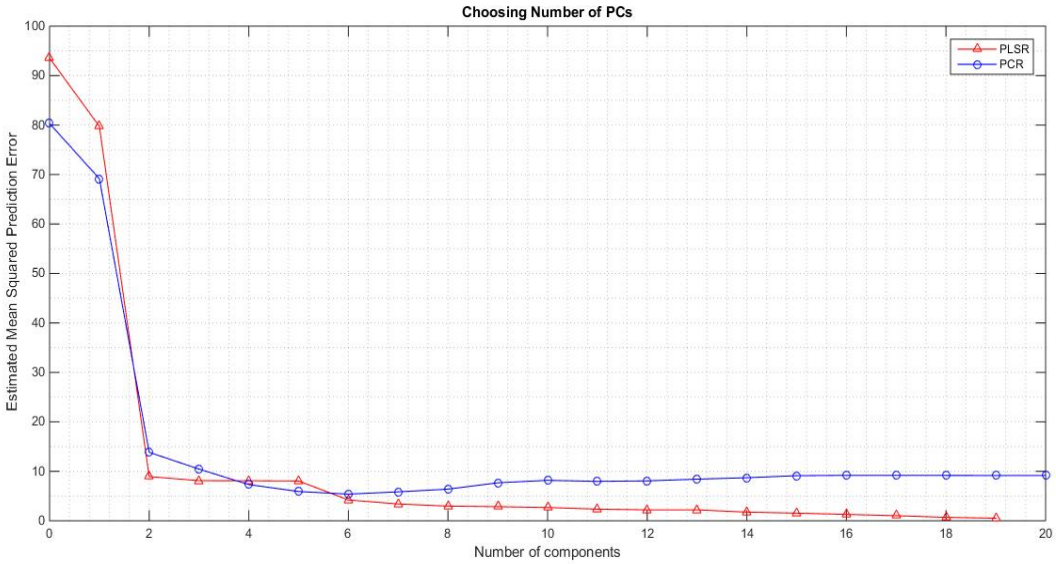


Figure 6.14: Number of Components Selection (All-29-NSt)

Μελετώντας τα διαγράμματα διακύμανσης και του αθροιστικού τετραγωνικού σφάλματος βγαίνει το συμπέρασμα ότι η καλύτερη επιλογή αριθμού Κυρίων Συνιστωσών είναι περίπου 15-20.

6.9 Διάγραμμα Βαρών PLS - PCA

Διάγραμμα βαρών των μεθόδων PLS-PCA. Εκτελεί ένα διάγραμμα για την εύκολη κατανόηση της κατάστασης του βάρους των μεταβλητών.

```
%% [SCRIPT 12]
% Model parsimony
% PLS Weights are linear combinations of the original variables.
% they describe how strongly each component in the PLSR depends
% on the original variables , and in what direction.

clc
close all

% Plot PLS Loadings
    [X1,Y1,Xs,Ys,beta,pctvar,mse,stats]=plsregress(Training_Set ,CN,3);

plot(1:(size(Training_Set ,2)),stats.W,'-');
xlabel('Variable');
ylabel('PLS Weight');
title('PLS Loadings')
legend({'1st Component' '2nd Component' '3rd Component'},'location','SE
    ');
grid minor


clearvars -except PCALoadings X1 Y1 Xs Xs Ys beta  n Absorbance
    Training_Set wavenumber CN  Data Labels

%: Plot PCA Loadings

figure()
clc
plot(1:(size(Training_Set ,2)),PCALoadings(:,1:4),'-');
xlabel('Variable');
ylabel('PCA Loading');
title('PCA Loadings')
legend({'1st Component' '1nd Component' '3rd Component' '4th Component'
    },'location','NW');
grid minor

clearvars -except n PCALoadings X1 Y1 Xs Xs Ys beta  n Absorbance
    Training_Set wavenumber CN  Data Labels
```

6.10 Μοντέλα Παλινδρόμησης με PLS

Ο επόμενος αλγόριθμος είναι κατασκευασμένος ώστε να υπολογίζει ένα Διάνυσμα παλινδρόμησης για μια διαδικασία PLS.

Γίνεται χρήση της εντολής -plsregress και υπολογίζει τους προβλεπόμενους αριθμούς κετανίου όπου παραθέτει και ένα διάγραμμα, το μέσο τετραγωνικό σφάλμα και τα σφάλματα των προβλέψεων.

```
%% [SCRIPT 13]: Regression - PLS
clc
close all

   [n,k]=size(CN);
   Number_of_PCs=input('Number_of_Components=');


% Selection of Samples for prediction
S=[1,16,20];
A=Training_Set;
A([S],:)=[];
B=CN;
B([S],:)=[];
[n,k]=size(B);

%Calculation of the Regression Vector X by PLS
    [Xloadings,Yloadings,Xscores,Yscores,betaPLS10,PLSPctVar]=plsregress
        (A,B,Number_of_PCs);
    yfitPLSn=[ones(size(A ,1),1) A ]*betaPLS10;


%scatter(CN,yfitPLSn,'ro')
scatter(B,yfitPLSn,'ro')
grid minor
lsline
xlabel('Cetane');
ylabel('Xscores*X');
title('PLS Regression efficiency')
hold on

    yfitPLSn2=[ones(size(Training_Set(S,:) ,1),1) Training_Set(S,:) ]*
        betaPLS10

scatter(CN(S),yfitPLSn2,'bo')


errors=B-yfitPLSn;
```

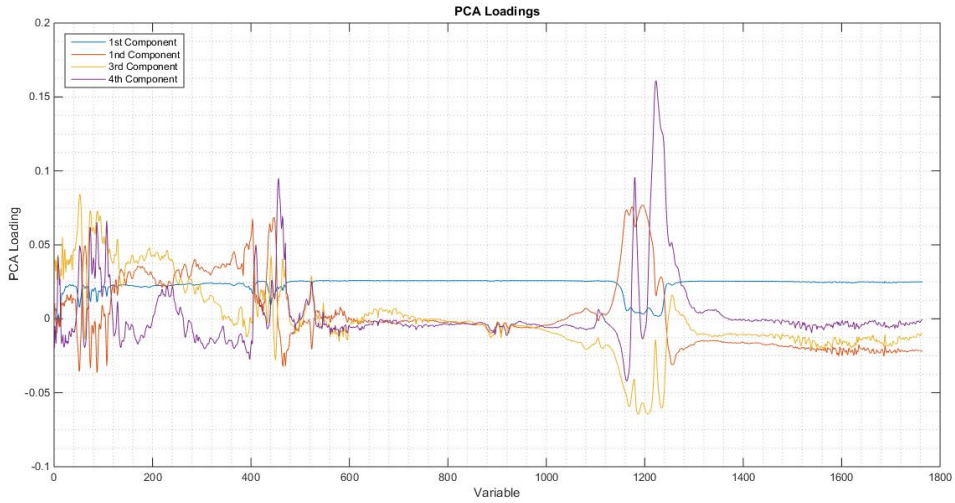


Figure 6.15: PCA Loadings

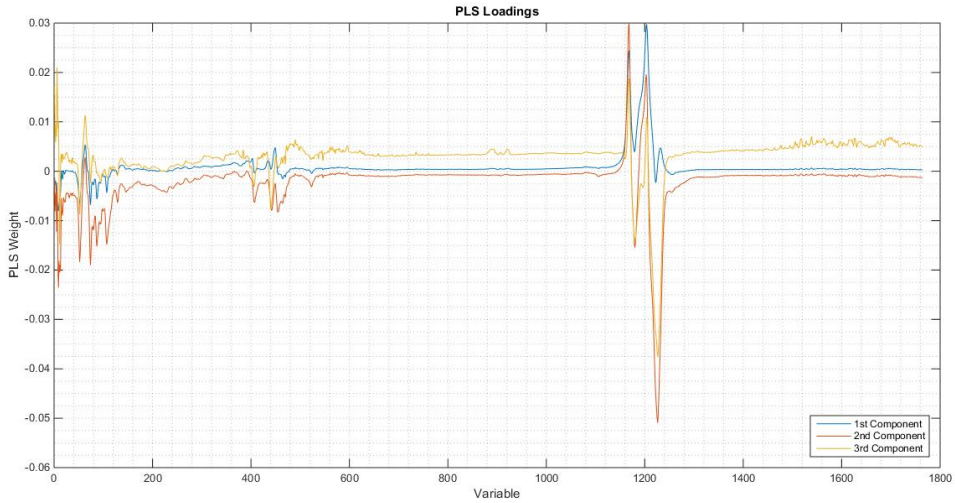


Figure 6.16: PLS Loadings

```
No=(1:1:n)';
table(No,B,yfitPLSn,errors)

TSS=sum((B-mean(B)).^2);
RSS=sum((B-yfitPLSn).^2);
Correlation_Coefficient=1-RSS/TSS

figure()

residuals = B - yfitPLSn;
stem(residuals)
xlabel('Samples');
ylabel('Residual');
grid on
title ('Residuals PLS ALL')


clearvars -except S A yfitPLSn B errors No Xloadings Yloadings Xscores
    Yscores betaPLS10 PLSPctVar n PCs Absorbance wavenumber CN  Data
    Labels Training_Set
```

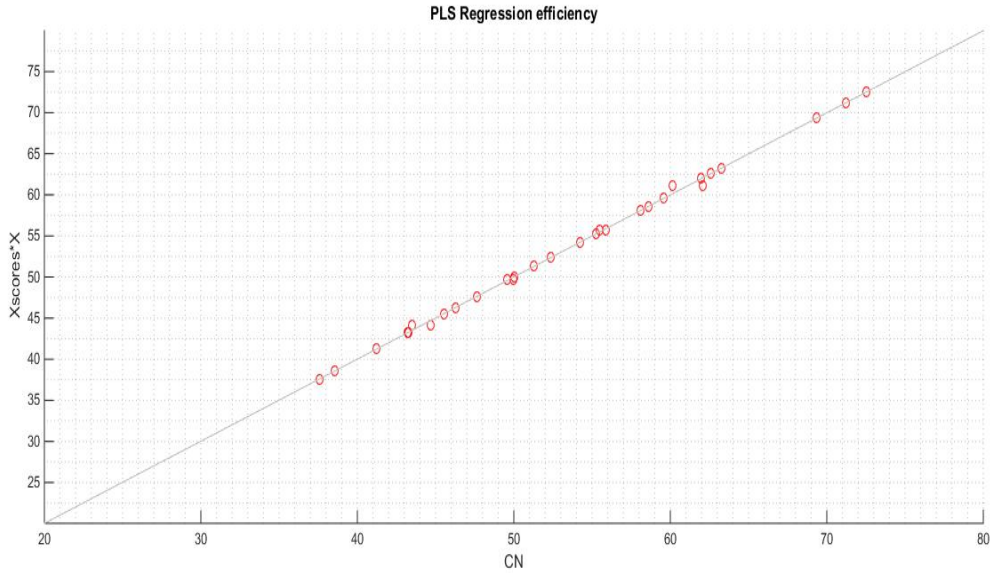


Figure 6.17: Regression PLS (20 PCs-All-31-St)

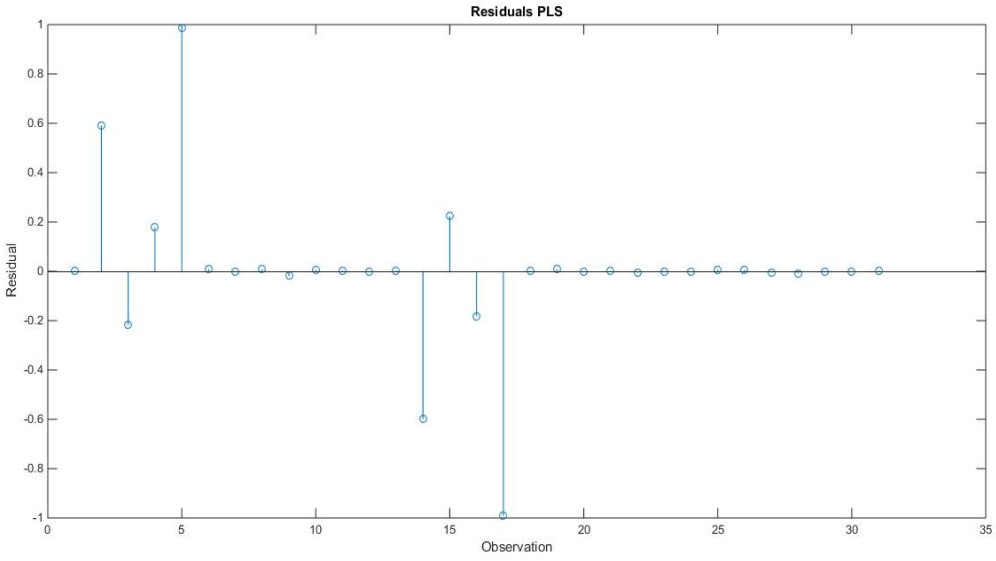


Figure 6.18: Regression PLS errors

6.11 Μοντέλα Παλινδρόμησης με PCR

Κατά την εκτέλεση μιας υπολογιστικής διαδικασίας, όπως και στην παρούσα εργασία, σωστό είναι στο μοντέλο της πρόβλεψης – με όποια μέθοδο κατασκευάζεται –να συνεισφέρει μόνο ένα μεγάλο ποσοστό του συνολικού δείγματος των δεδομένων (εδώ 90%) έτσι ώστε το μοντέλο της πρόβλεψης να υπολογίζει το υπόλοιπο 10%.

Αυτό συμβαίνει για λόγους επαλήθευσης της μεθόδου. Στην παρούσα εργασία ενώ υπάρχουν συνολικά 31 δείγματα, η πρόβλεψη με το μοντέλο της PCR θα γίνει στα 28 δεδομένα μόνο, ενώ θα προβλεφθούν τα υπόλοιπα 3.

Στην περίπτωση όπου έχουν απαλειφθεί 2 μετρήσεις, συνολικά θα υπάρχουν 29 δείγματα και το μοντέλο θα κατασκευαστεί για τα 26 από αυτά, ενώ τα άλλα 3 θα προβλεφθούν.

Ο αλγόριθμος που παρουσιάζεται παρακάτω είναι κατασκευασμένος ώστε να δίνει την ευχαίρια στον χρήστη να θέσει ο ίδιος τα δείγματα στα οποία θέλει να γίνει η πρόβλεψη. Ο χρήστης μπορεί να προσαρμόσει στον αλγόριθμο στη μεταβλητή S[...]; τα συγκεκριμένα δείγματα που επιθυμεί.

Σημαντικό είναι η επιλογή των τριών τιμών Αριθμού Κετανίου, όπως προαναφέρθηκε, να βρίσκονται σε όλο το εύρος των περιοχών.

No	1500-2500	2500-3500
1	HC 20-225	38.57
2	HC 225-250	44.70
3	HC 250-275	49.55
4	HC 275-300	55.87
5	HC 300-325	62.1
6	HC Res	72.52
7	HDS 25-225	43.24
8	HDS 200-225	47.62
9	HDS 225-250	51.29
10	HDS 250-275	54.26
11	HDS 275-300	58.63
12	HDS 300-325	61.95
13	HDS Res	63.26
14	HT 225-250	43.51
15	HT 250-275	49.99
16	HT 275-300	55.51
17	HT 300-325	60.12
18	MOH HDS 200-225	41.24
19	MOH HDS 225-250	45.53
20	MOH HDS 250-275	50.01
21	MOH HDS 275-300	55.23
22	MOH HDS 300-325	58.61
23	MOH HDS Res	59.55
24	MOH MHC 25-200	37.59
25	MOH MHC 200-225	43.21
26	MOH MHC 225-250	46.29
27	MOH MHC 250-275	52.37
28	MOH MHC 275-300	58.11
29	MOH MHC 300-325	62.59
30	MOH MHC 325-350	69.33
31	MOH MHC Res	71.23

Οι μπλέ τιμές είναι αυτές που χρησιμοποιούνται για πρόβλεψη παρακάτω ως άγνωστες παράμετροι

Το εύρος των αριθμών κετανίου είναι ελάχιστο 37.59, μέγιστο στο 72.52, με μέσο όρο 53.66. Οπότε για τα 28 δεδομένα επιλέγονται τα δείγματα S(1, 16, 31). Για τα 26 δείγματα επιλέγονται S(1, 11, 29), ώστε και στις 2 περιπτώσεις να προβλεφθούν τιμές σε όλο το εύρος.

Στο επόμενο διάγραμμα PCR οι τιμές με κόκκινο (x) είναι οι 3 τιμές που θεωρήθηκαν ώς άγνωστα δείγματα για την μελέτη πρόβλεψης του μονέλου αυτού.

```
%% [SCRIPT 14] Regression - PCR
close all
clc

```

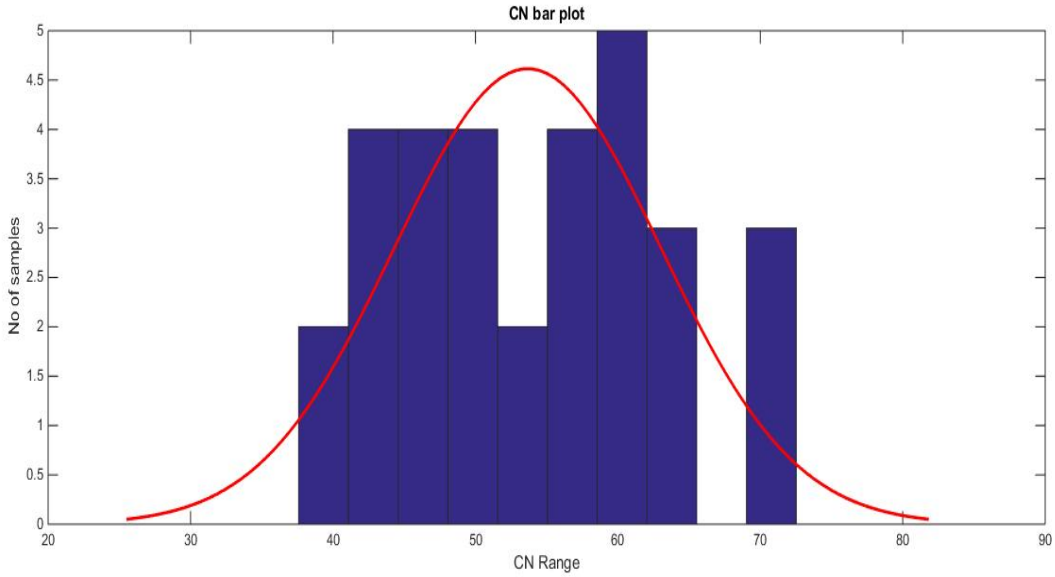


Figure 6.19: Data Cetane Number Bar Plot

```
5       Number_of_PCs=input('Number_of_PCs=');
6
7   % Selection of Number_of_Samples
8       [n,k]=size(CN);
9       cn_Data=(Data(1:end,2:end))';
10
11  % delete useless rows from PCAScores Matrix to make the new
        Wavenumberset
12      S=[2,16,24];
13      A=Training_Set;
14      A([S],:)=[];
15      B=CN;
16      B([S],:)=[];
17      C=(CN(S));
18
19  % PCA
20      [PCALoadings, PCAScores,PCAVar]=pca(A,'Centered',false,'Economy',
            false,'NumComponents',Number_of_PCs);
21
22  % Standardization of CN
23      cn=(B-min(B))/(max(B)-min(B));
24
25  % Calculation of regression vector
```

```
26      betaPCRn=regress(cn-mean(cn), PCAScores(:,1:Number_of_PCs));
27
28  % Calculation of CN_new
29      cn_new=(PCAScores*betaPCRn+mean(cn))*(max(B)-min(B))+min(B);
30
31  % Plot 90% of Data
32  scatter(B,cn_new,'bo')
33  lsline
34  xlabel('Cetane Number')
35  ylabel('PCAScores*x')
36  grid minor
37  title('PCR Efficiency of prediction')
38  hold on
39
40  % Adaptation of Data C
41  % This script is about to calculate the efectivenes of prediction.
42  AA=Training_Set(S,1:end);
43  zz=AA*PCALoadings;
44  cn_C_Predicted=(zz*betaPCRn+mean(cn))*(max(B)-min(B))+min(B)
45
46    TSS=sum((B-mean(B)).^2);
47    RSS=sum((B-cn_new).^2);
48     Correlation_Coeficient=1-RSS/TSS
49
50  % RMSEP
51      RMSEP=sqrt(sum((cn_new-B).^2)/size(B,1))
52      m=sum(B-cn_new)/size(B,1);
53      SEP=sqrt((sum((B-cn_new-m).^2))/(size(B,1)-1))
54       Error=(cn_new-B);
55
56  % Plot 10% of Data
57  hold on
58  scatter(C,cn_C_Predicted,'rx')
59  error=(C-cn_C_Predicted);
60  No=(1:1:3)';
61  table(No,C,cn_C_Predicted,error)
62
63  figure()
64  stem(B-cn_new)
65  ('betaPCRn = Regression Vector')
66  ('PCALoadings = Loadings')
67  xlabel ('Samples')
68  ylabel ('Error')
69  title ('PCR Cetane Residuals')
70  grid minor
```

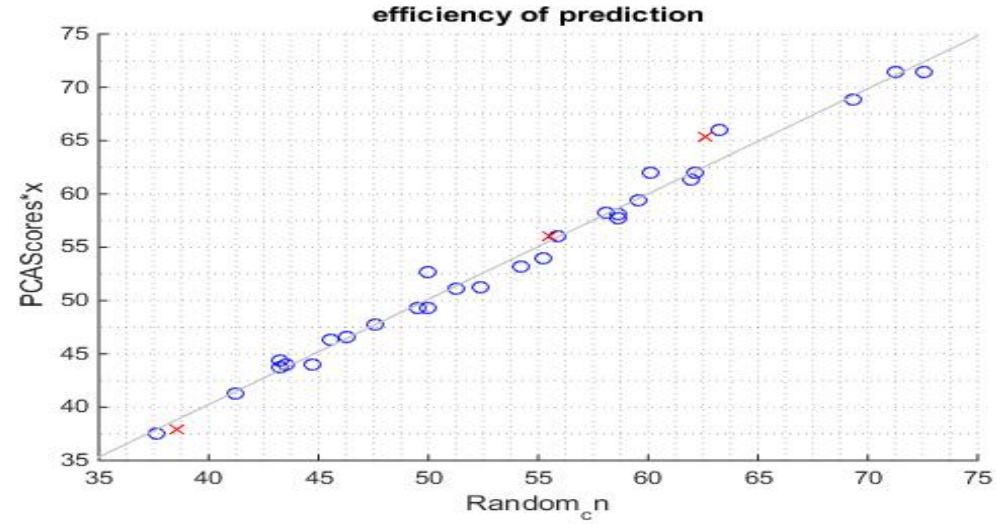


Figure 6.20: PCA Prediction Efficiency (20PCs-28-St)

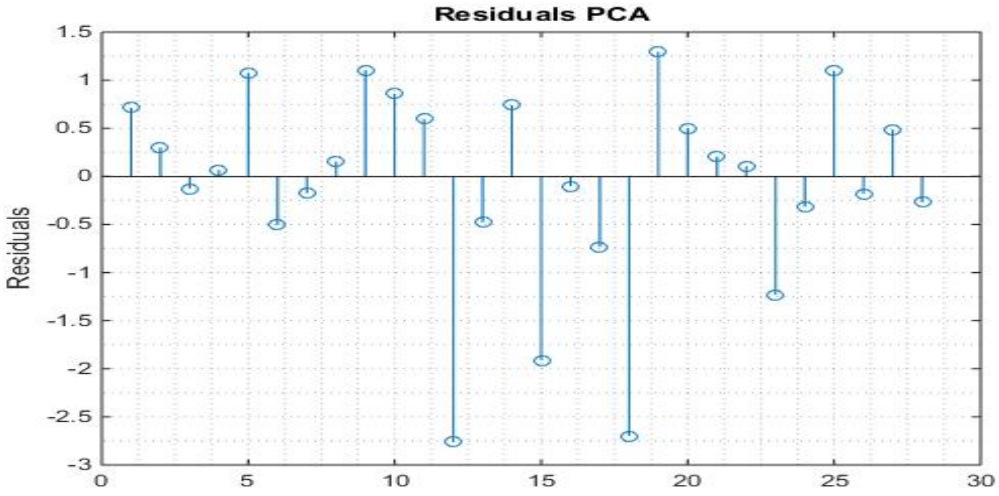


Figure 6.21: PCA Error Stem

Chapter 7

ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Ο αλγόριθμος θα τρέξει εισάγοντας διαφορετικά δεδομένα απορρόφησης, κακονιοποιημένα ή μη κανονικοποιημένα για διαδικασίες PLS και PCR.

Από το διάγραμμα του μέσου τετραγωνικού σφάλματος, σε σχέση με τον αριθμό των Κυρίων Συνιστωσών (εικόνα 7.14), γίνεται εμφανές ότι η καλύτερη επιλογή γίνεται ανάμεσα σε 15 με 20 Κύριες Συνιστώσες. Οι επόμενοι υπολογισμοί έγιναν με 10 και 20 Κύριες Συνιστώσες.

Σκοπός είναι να βρεθεί η κατάλληλη μέθοδος η οποία παράγει το μικρότερο δυνατό μέσο τετραγωνικό σφάλμα (RMSEP) και που εμφανίζει τον μεγαλύτερο συντελεστή συσχέτισης.

Ο υπολογισμός του συντελεστή συσχέτησης γίνεται με την εντολή.

```
TSS=sum((Y-mean(Y)).^2);\\
RSS=sum((Y- Y    ).^2);\\
rsquared=1-RSS/TSS\\
```

ενώ το μέσο τετραγωνικό σφάλμα

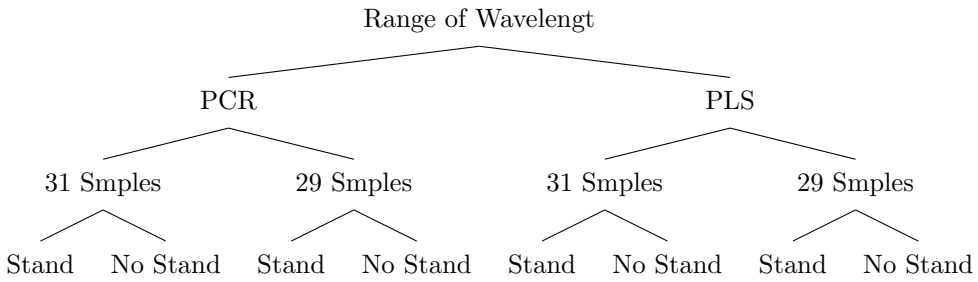
```
RMSEP=sqrt(sum((Y-Y').^2)/size(Y,1))
```

$$RMSEP = \sqrt{\sum_{i=1}^{n}(Y_{real} - Y_{pred})^2/n}$$

Η μελέτη των αποτελεσμάτων είναι συγκριτική μεταξύ των παρακάτω παραγόντων.

- PLS-PCR

- Standardized - No Standardized
- Με όλα τα δείγματα (31) – Με τα (29) δείγματα
- wavelength (1500-2500 – 2500-3500 – All Spectra)



Ως εκ τούτου υπάρχουν $3*8 = 24$ αποτελέσματα του συντελεστή συσχέστησης, τα οποία παρουσιάζονται παρακάτω.

VALIDATION SET-10 PCs

Data	1500-2500	2500-3500	All Spectra
(28/31)-PCR-Stand	0.9365	0.9765	0.9702
(28/31)-PCR-NoStand	0.9247	0.9754	0.9714
(26/29)-PCR-Stand	0.9299	0.9656	0.9622
(26/29)-PCR-NoStand	0.9270	0.9680	0.9568

Data	1500-2500	2500-3500	All Spectra
(28/31)-PLS-Stand	0.9873	0.9908	0.9947
(28/31)-PLS-NoStand	0.9842	0.9915	0.9949
(26/29)-PLS-Stand	0.9915	0.9899	0.9964
(26/29)-PLS-NoStand	0.9892	0.9923	0.9962

TRAINING SET-10 PCs.

Data	1500-2500	2500-3500	All Spectra
(31)-PCR-Stand	0.9396	0.9757	0.9732
(31)-PCR-NoStand	0.9298	0.9740	0.9687
(29)-PCR-Stand	0.9310	0.9735	0.9715
(29)-PCR-NoStand	0.9203	0.9723	0.9609

Data	1500-2500	2500-3500	All Spectra
(31)-PLS-Stand	0.98835	0.9894	0.9939
(31)-PLS-NoStand	0.9803	0.9914	0.9939
(29)-PLS-Stand	0.9831	0.9892	0.9956
(29)-PLS-NoStand	0.9793	0.9938	0.9955

VALIDATION SET-20 PCs

Data	1500-2500	2500-3500	All Spectra
(28/31)-PCR-Stand	0.9863	0.9937	0.9873
(28/31)-PCR-NoStand	0.9748	0.9935	0.9904
(26/29)-PCR-Stand	0.9923	0.9957	0.9965
(26/29)-PCR-NoStand	0.9908	0.9976	0.9978

Data	1500-2500	2500-3500	All Spectra
(28/31)-PLS-Stand	0.9988	0.9988	0.9988
(28/31)-PLS-NoStand	0.9988	0.9988	0.9988
(26/29)-PLS-Stand	0.9995	0.9995	0.9995
(26/29)-PLS-NoStand	0.9995	0.9995	0.9995

TRAINING SET-20 PCs.

Data	1500-2500	2500-3500	All Spectra
(31)-PCR-Stand	0.9740	0.9921	0.9881
(31)-PCR-NoStand	0.9588	0.9943	0.9894
(29)-PCR-Stand	0.9835	0.9942	0.9948
(29)-PCR-NoStand	0.9687	0.9957	0.9748

Data	1500-2500	2500-3500	All Spectra
(31)-PLS-Stand	0.9987	0.9987	0.9987
(31)-PLS-NoStand	0.9987	0.9987	0.9987
(29)-PLS-Stand	0.9987	0.9987	0.9987
(29)-PLS-NoStand	0.9987	0.9987	0.9987

- Όσο πιο κοντά στην μονάδα βρίσκεται το αποτέλεσμα, τόσο πιο αξιόπιστη είναι η πρόβλεψη.
- Συνεπώς το καλύτερο μοντέλο πρόβλεψης παρουσιάζεται όταν αφαιρούνται οι μή αξιόπιστες μεταβλητές και υπάρχουν συνολικά 29 δείγματα για το Μοντέλο επικύρωσης, με τη μεθοδολογία της PLS, σε Κανονικοποιημένα δεδομένα από το φάσμα 2500-3500 cm^{-1} .
- Η μεθοδολογία της PLS παρουσιάζει καλύτερα αποτελέσματα κατα 3.75% στα δεδομένα των 10 Κυρίων Συνιστωσών και 0.74% στα δεδομένα 20 Κυρίων Συνιστωσών.
- Τα κανονικοποιημένα δεδομένα υπολογίζουν καλύτερα από τα μη επεξεργασμένα κατά 0.17% στα δεδομένα των 10 Κυρίων Συνιστωσών και κατά 0.058% στα δεδομένα 20 Κυρίων Συνιστωσών.
- Επίσης, το εύρος κυματάριθμων από $2500 - 3500cm^{-1}$, δίνει σχετικά με τους υπόλοιπους, πρόβλεψη μεγαλύτερης ακίβειας, με οποιονδήποτε αριθμό Κυρίων Συνιστωσών.

7.1 Στοιχεία για την πρόβλεψη Ιδιοτήτων αγνώστου δείγματος

Εκτελώντας το παραπάνω πρόγραμμα παράγονται τα Διανύσματα Βαρών σε ένα διάνυσμα
size(Loading Vector) = Number of PCs, Number of Variables.

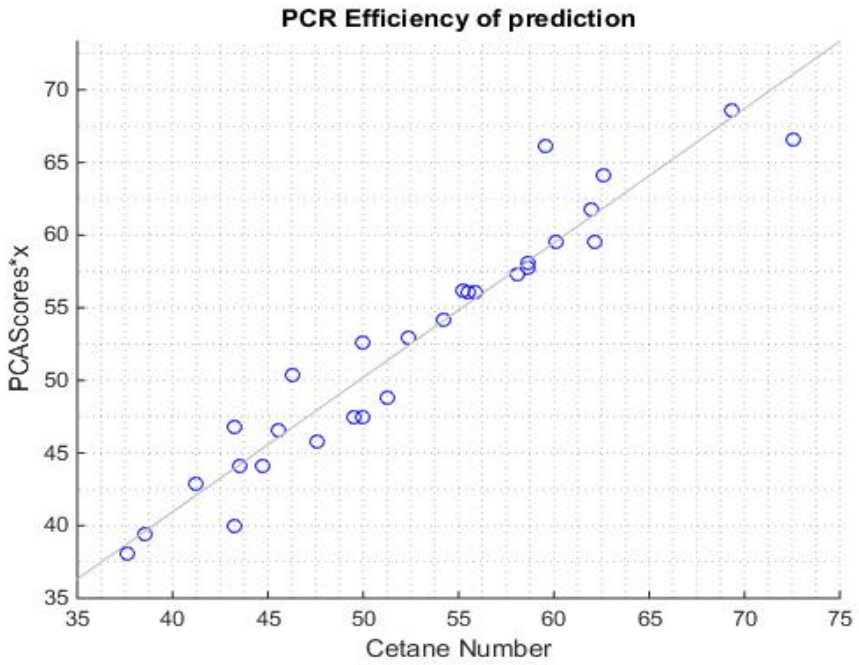
Για τον υπολογισμό του Αριθμού Κετανίου ενός αγνώστου δείγματος από φασματοσκοπικά δεδομένα υπέρυθρης φασματοσκοπίας, θα πρέπει να πολλαπλασιαστεί το διάνυσμα των τιμών του φασματοσκοπήματος με τα Διανύσματα Βαρών.
Στην παρούσα εργασία τα Διανύσματα Βαρών υπολογίζονται στη μεταβλητή με όνομα PCALoadings για την μεθοδολογία PCR και Yloadings για την μεθοδολογία της PLS. Εν συνεχεία το αποτέλεσμα θα πολλαπλασιαστεί με το Διάνυσμα παλινδρόμησης (betaPCRn).
Τα αποτελέσματα που παράγονται είναι με κανονικοποιημένα δεδομένα του Trainings Set του αριθμού κετανίου. Για να υπολογιστούν τα πραγματικά δεδομένα πρέπει να εκτελεστεί η εντολή που αναγράφεται στον κώδικα.

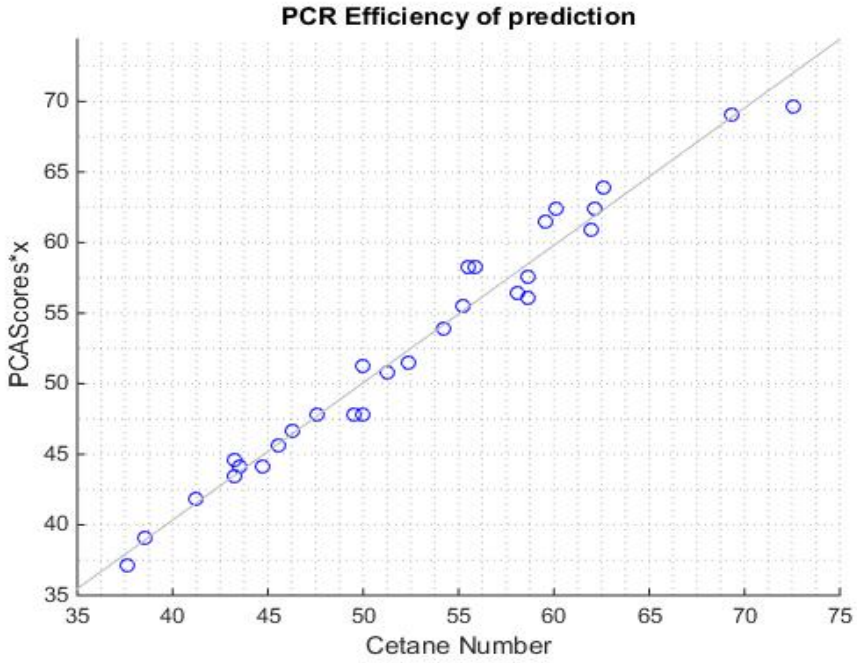
$$CNPredicted = (New{-}IR{*}PCALoadings{*}betaPCRn{+}mean(cn)){*}(max(B){-}min(B)){+}min(B).$$

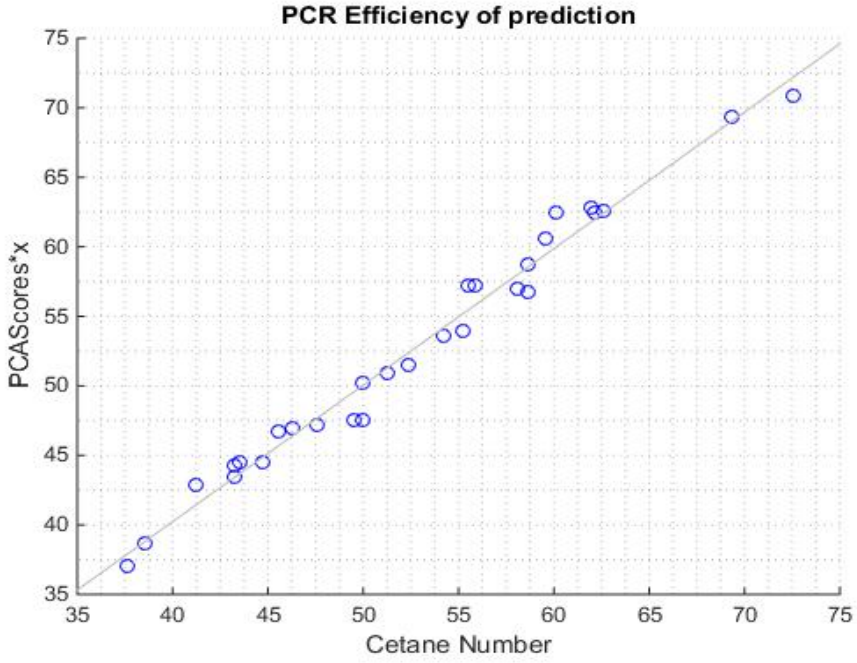
Το φασματογράφημα του αγνώστου δείγματος θα πρέπει να είναι ίδιας μορφής και μεγέθους με τα πρότυπα του Μοντέλου εκπαίδευσης.

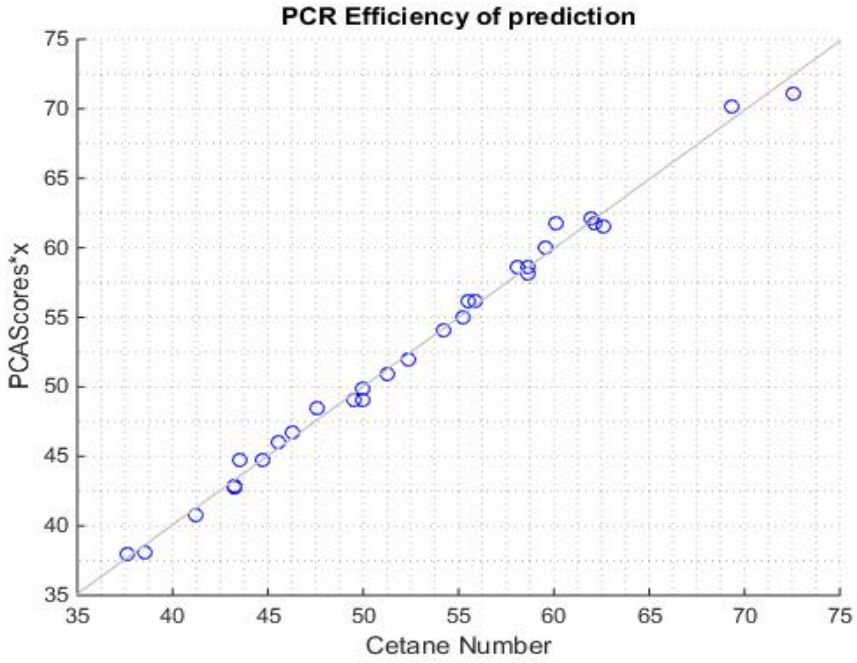
7.2 Προσαρμογή περισσοτέρων Κυρίων Συνιστωσών στο Μοντέλο εκπαίδευσης

Διαγράμματα 5,10,15,20 PCs




PCR Efficiency of prediction
PCAScores*x
Cetane Number
35
40
45
50
55
60
65
70
75


PCR Efficiency of prediction
PCAScores*x
Cetane Number
35
40
45
50
55
60
65
70
75



PCA for n-PCs
Διάνυσμα παλινδρόμησης

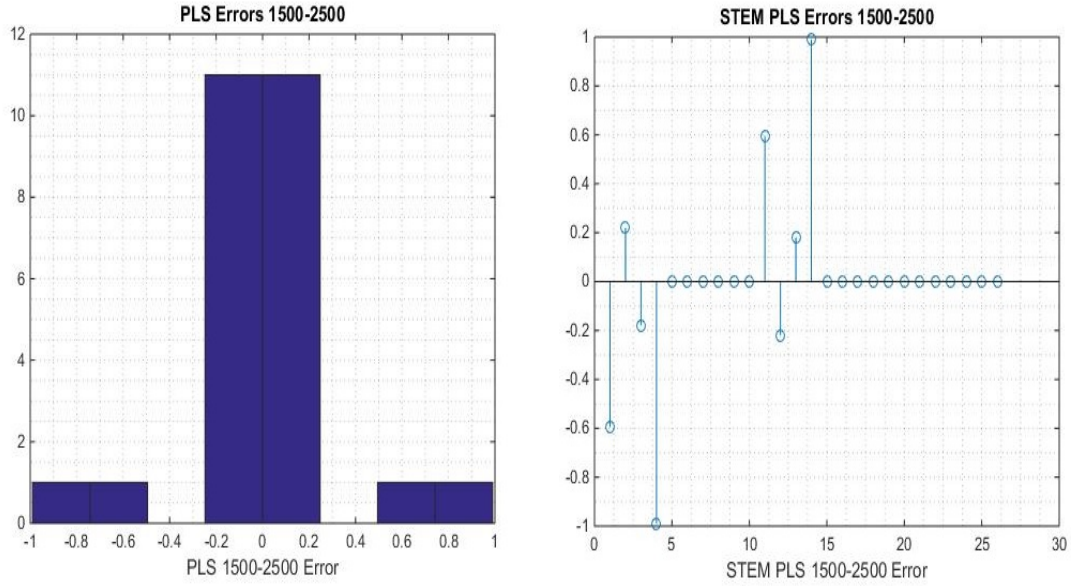
- Reg-V-20PCs-St=(0.0024 0.0098 -0.0234 -0.0041 0.0081 -0.0052 0.0160 -0.0108 -0.0124 0.0005 -0.0209 -0.0090 0.0323 -0.0101 0.0160 -0.0082 0.0325 0.0114 -0.0393 0.0379)

7.3 Μελέτη σφαλμάτων - Αριθμού Κετανίου

Στον παρακάτω πίνακα παρουσιάζεται το μέσο τετραγωνικό σφάλμα και η τυπική απόκληση απόλυτων τιμών σφάλματος, για τις διαδικασίες PLS και PCR με κανονικοποιημένα δεδομένα.

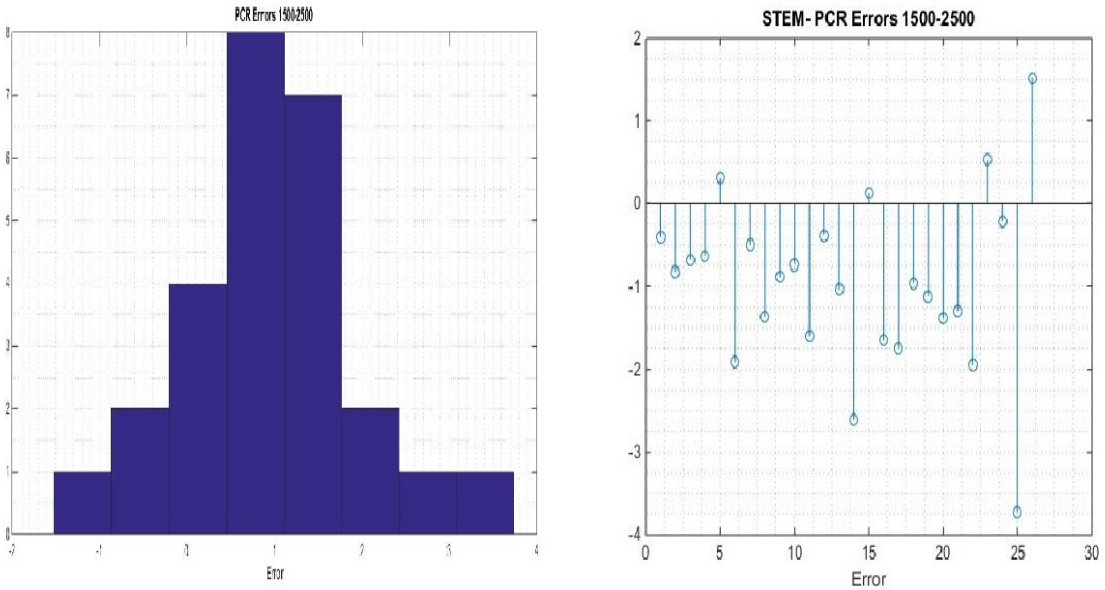
Method/Range	RMSEP	SEP
PCR 1500-2500	1.4068	1.0423
PCR 2500-3500	0.7511	0.5400
PCR ALL	0.6911	0.6285
PLS 1500-2500	0.3180	0.3238
PLS 2500-3500	0.3179	0.3237
PLS ALL	0.3179	0.3237

Εν συνεχεία παρουσιάζονται τα ιστογράμματα των σφαλμάτων.



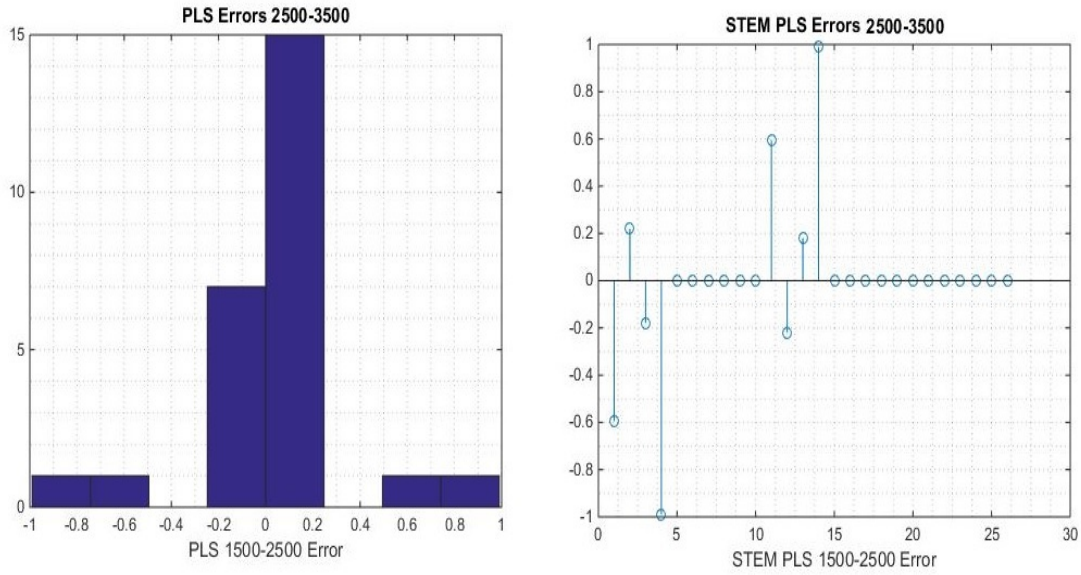
1500.jpg

Figure 7.1: PLS 1500-2500



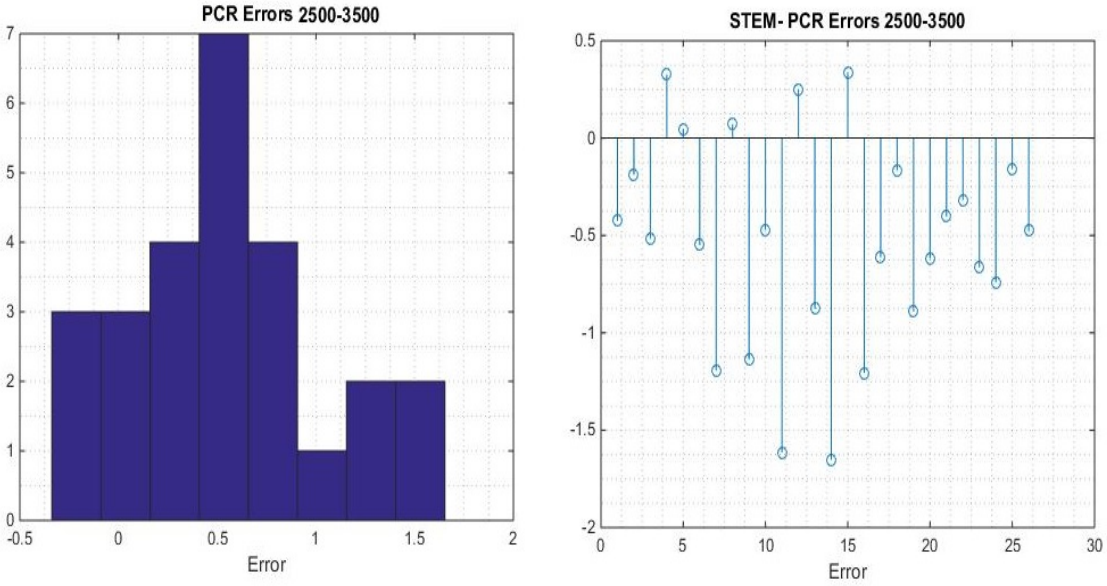
1500.jpg

Figure 7.2: PCR 1500-2500



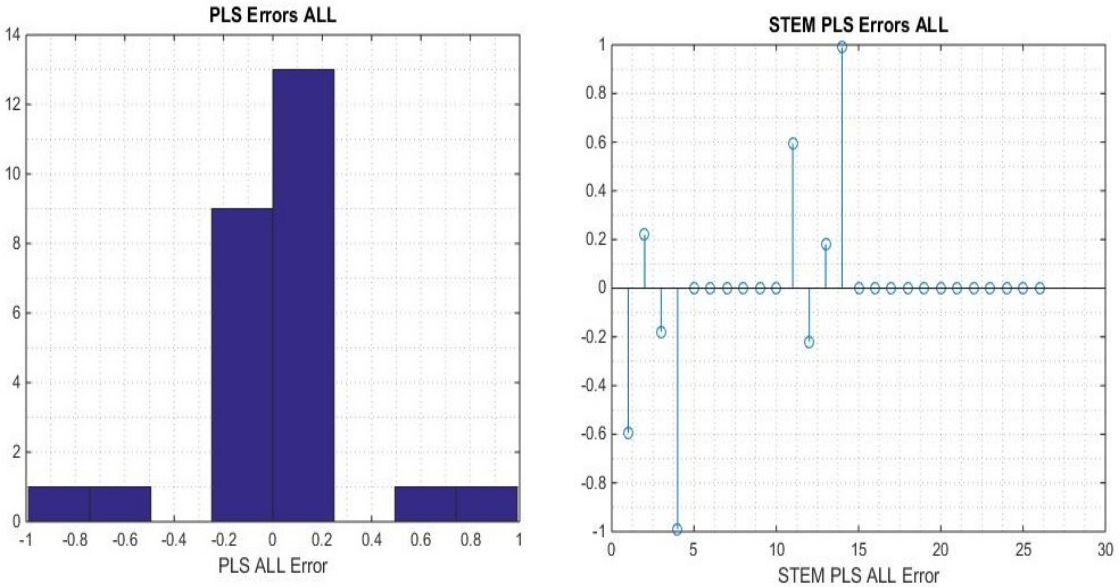
2500.jpg

Figure 7.3: PLS 2500-3500



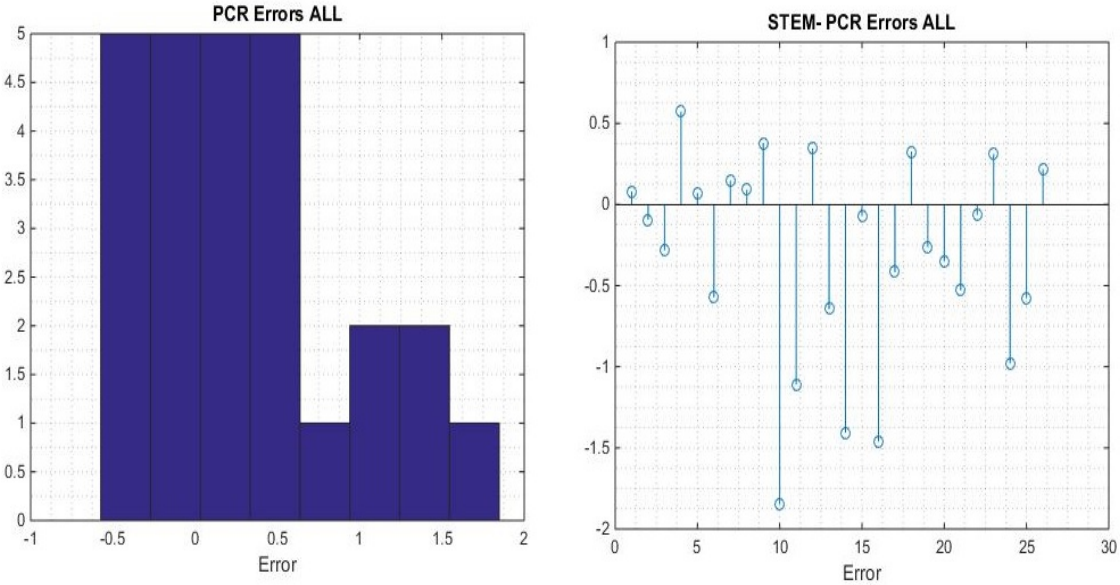
2500.jpg

Figure 7.4: PCR 2500-3500



ALL.jpg

Figure 7.5: PLS ALL



ALL.jpg

Figure 7.6: PCR ALL

7.4 Ακρίβεια υπολογισμών - Σύγκριση με αποτελέσματα άλλων ερευνών

Η ακρίβεια των υπολογισμών χαρακτηρίζεται ως ικανοποιητική, ειδικά για την περίπτωση των 20 Κυρίων Συνιστωσών, με ακρίβεια υπολογισμών μικρότερη του 1 για τη μεθοδολογία της PLS και για την PCR από (-2.5 - 1).

- Αντίστοιχες προβλέψεις έχουν γίνει με ακρίβεια 1.6 (mean absolute error)[**?**]
- Ακρίβεια μεγαλύτερη από 92% χρησιμοποιώντας τεχνητά νευρωνικά δίκτυα και Multiple Linear Regression [**?**]
- Με Standared error μικρότερο του 1.5 με πρόβλεψη από τις φυσικές ιδιότητες[**?**]
- Πρόσπάθεια πρόβλεψης του CN από την σύνθεση λιπαρών οξέων με μεθόδους Multiple Linear regression εδωσε ένα $R^2 = 0.953$ [**?**]

7.5 Μοντέλα πρόβλεψης Ιξώδους

Η εφαρμογή των μεθόδων αυτών μπορεί να εκτελεστεί και για την πρόβλεψη άλλων ιδιοτήτων. Στα δεδομένα που δόθηκαν στην παρούσα εργασία, όπως αναφέρθηκε, υπάρχουν δεδομένα για το ιξώδες, το ποσοστό θείου, τον δείκτη διάθλασης και το ποσοστό αρωματικών συστατικών.
Ο Αριθμός Κετανίου έχει καλή γραμμική σχέση με το ιξώδες. Αυτό σημαίνει ότι η μέθοδος αυτή θα μπορούσε να δώσει καλά αποτελέσματα και για την πρόβλεψη του ιξώδους.
Η μέθοδος της PLS και PCR εφαρμόζεται για την πρόβλεψη του ιξώδους από φασματοσκόπιμα και προκύπτουν οι επόμενες εικόνες.
Παρατηρείται ότι η καλύτερη πρόβλεψη για το ιξώδες, γίνεται με την μεθοδολογία της PLS για τους κυμματάριθμους από 2500-3500 $cm^{-}1$. Η μέτρηση των δειγμάτων για το ιξώδες δεν θα πρέπει να διαφέρει περισσότερο από 5% από το πραγματικό απότέλεσμα σύμφωνα με τις περισσότερες μεθόδους όπως ASTM 975 και ASTM D445 και ISO 3104. Οι μεθοδολογία της PLS θα μπορούσε να εφαρμοστεί για τον προσδιορισμό του ιξώδους.
Για την επεξεργασία των δεδομένων για την πρόβλεψη του ιξώδους χρησιμοποιήθηκαν 21 δεδομένα ιξώδους, διαφορετικών δειγμάτων, με τα αντίστοιχα φασματοσποπίματα, όπως ακριβώς και με την προβλεψη του Αριθμού Κετανίου.
Τα μοντέλα πρόβλεψης παρουσιάζονται εν συνεχεία στις επόμενες σελίδες.

Method/Range	RMSEP	SEP
PCR 1500-2500	0.1326	0.1358
PCR 2500-3500	0.2144	0.2197
PCR ALL	0.2015	0.2065
PLS 1500-2500	0.0354	0.0363
PLS 2500-3500	0.0319	0.0327
PLS ALL	0.0319	0.0327

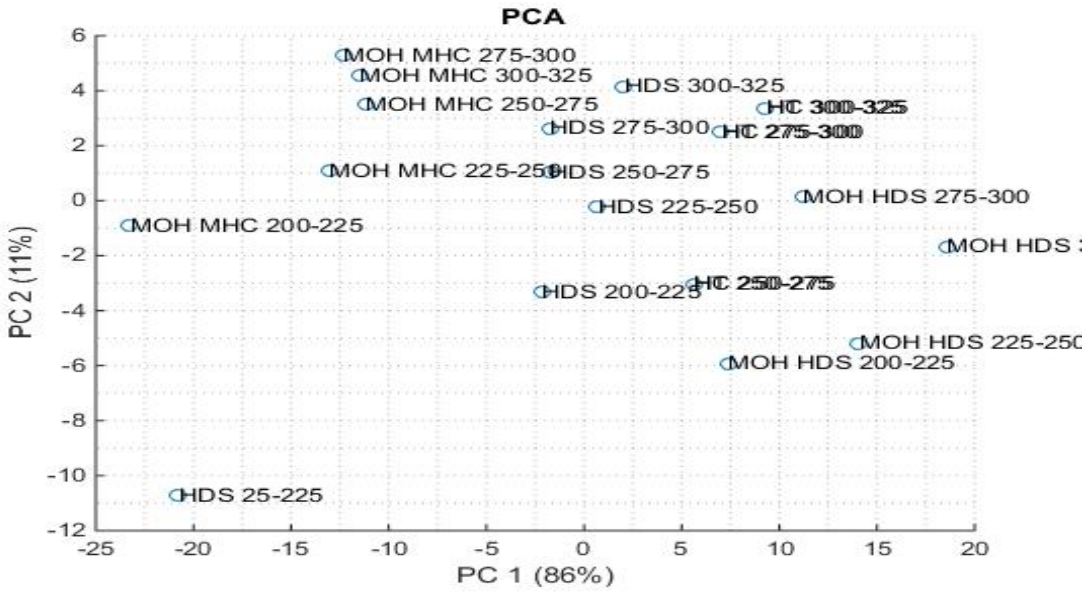


Figure 7.7: PCA Viscosity

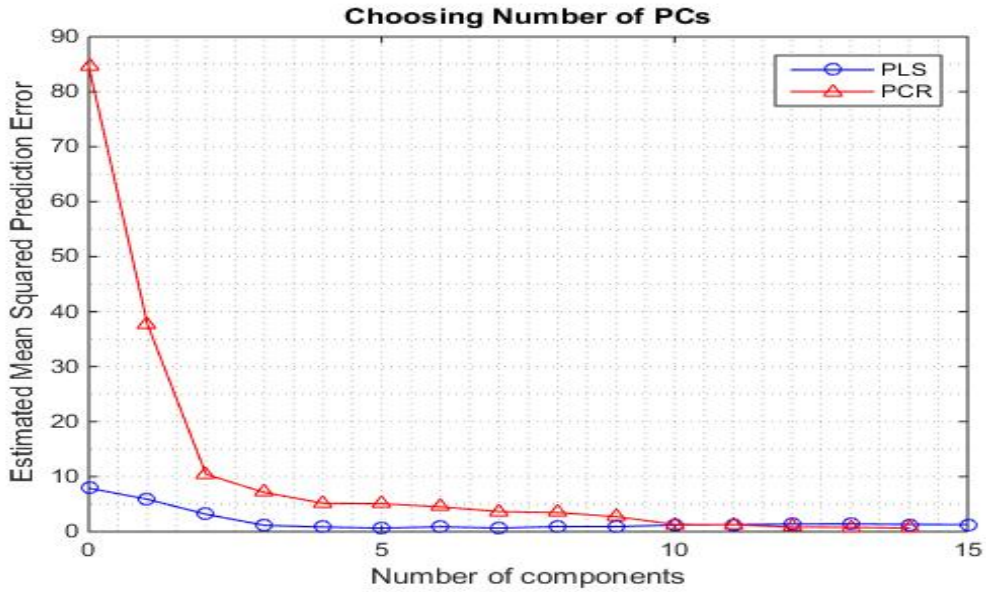


Figure 7.8: Viscosity error PLS PCR

St PLS.jpg

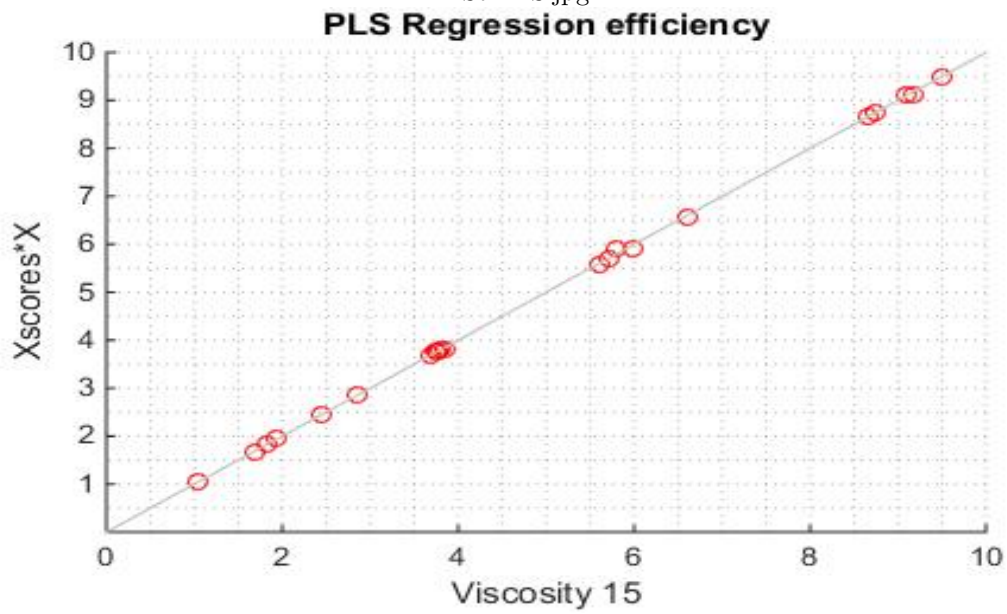


Figure 7.9: 1500-2500 St PLS

St PLS Residuals.jpg

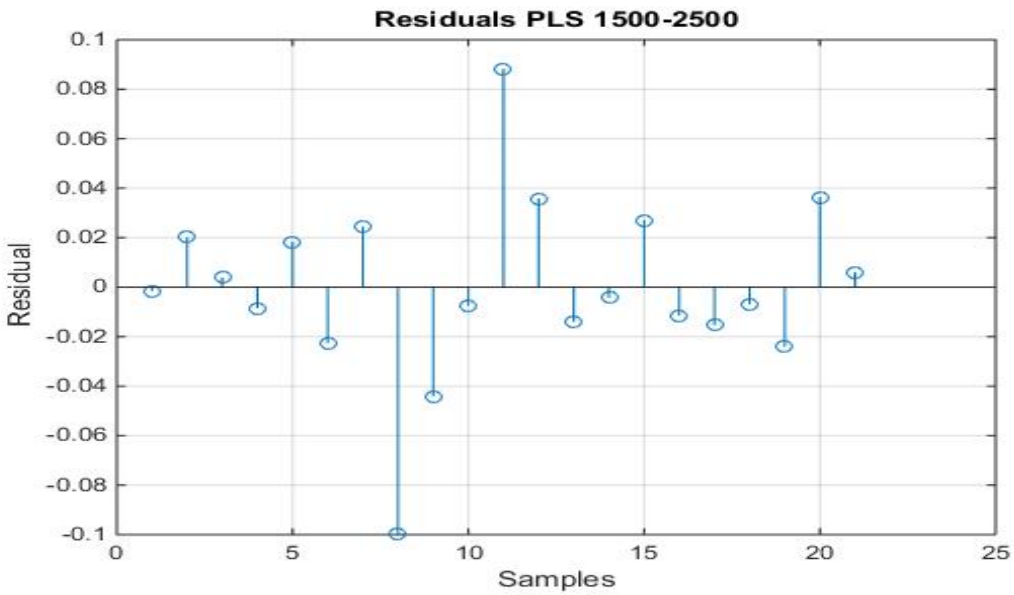


Figure 7.10: 1500-2500 St PLS Residuals

Figure 7.11: 1500-2500 St PLS Viscosity

St PCR.jpg

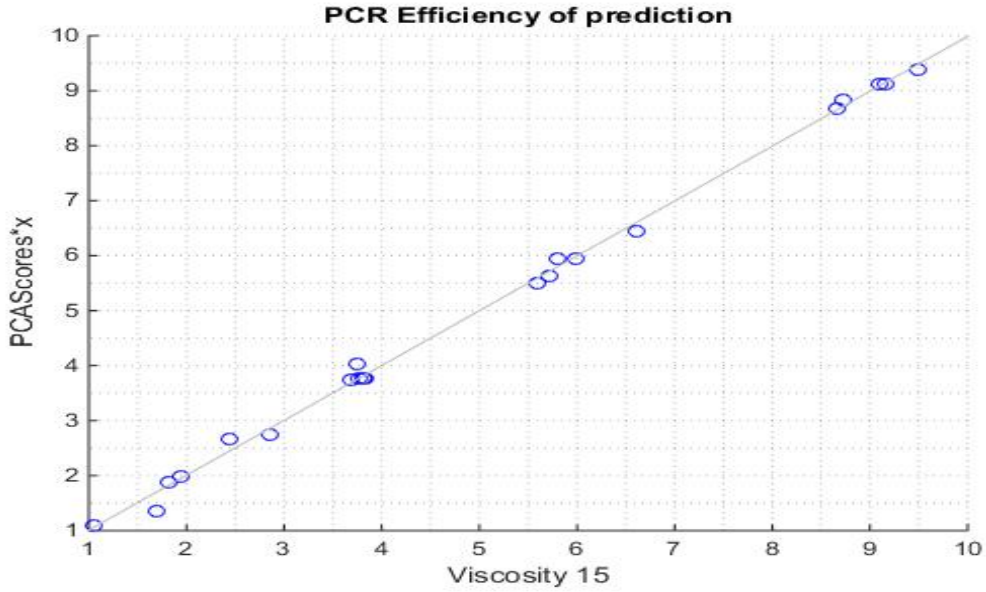


Figure 7.12: 1500-2500 St PCR

St PCR Residuals.jpg

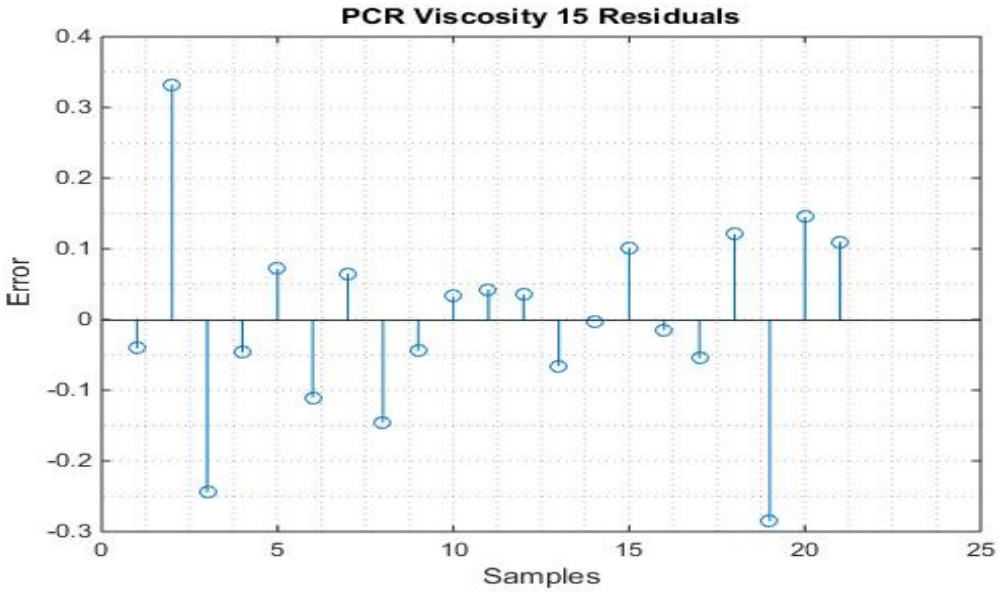


Figure 7.13: 1500-2500 St PCR Residuals

Figure 7.14: 1500-2500 St PCR Viscosity

St PLS.jpg

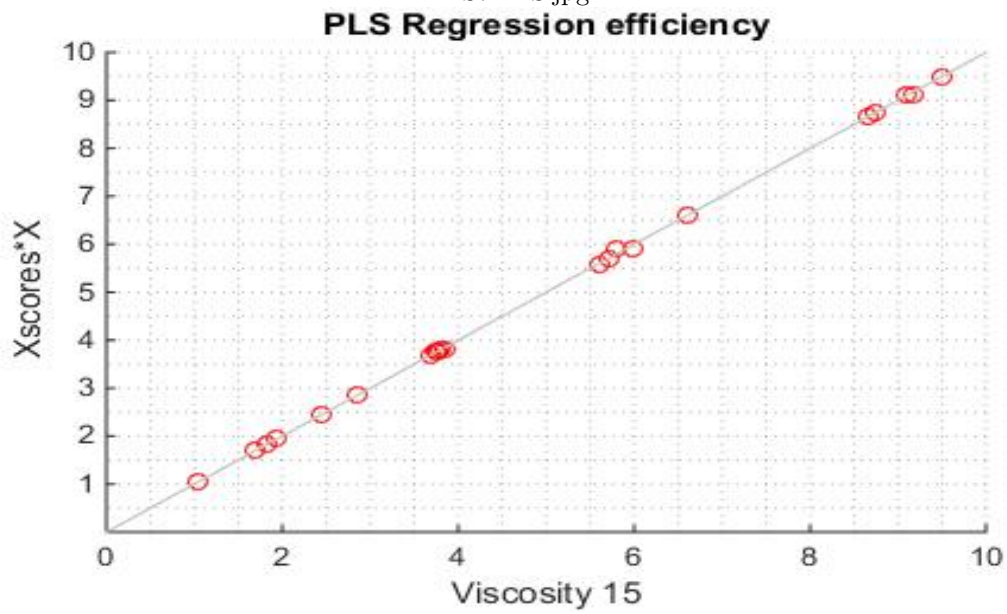


Figure 7.15: 2500-3500 St PLS

St PLS Residuals.jpg

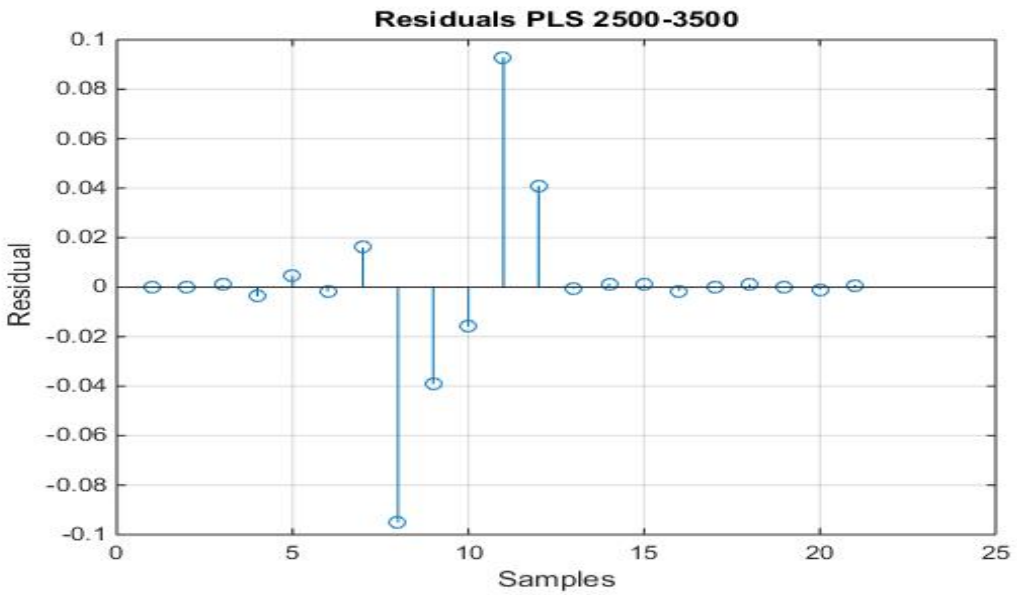


Figure 7.16: 2500-3500 St PLS Residuals

Figure 7.17: 2500-3500 St PLS Viscosity

St PCR.jpg

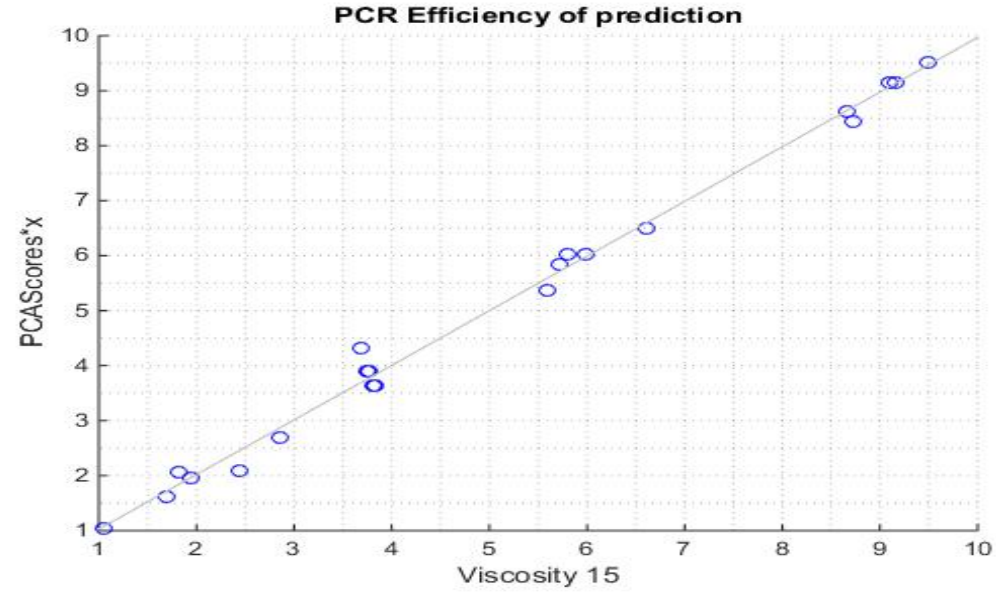


Figure 7.18: 2500-3500 St PCR

St PCR Residuals.jpg

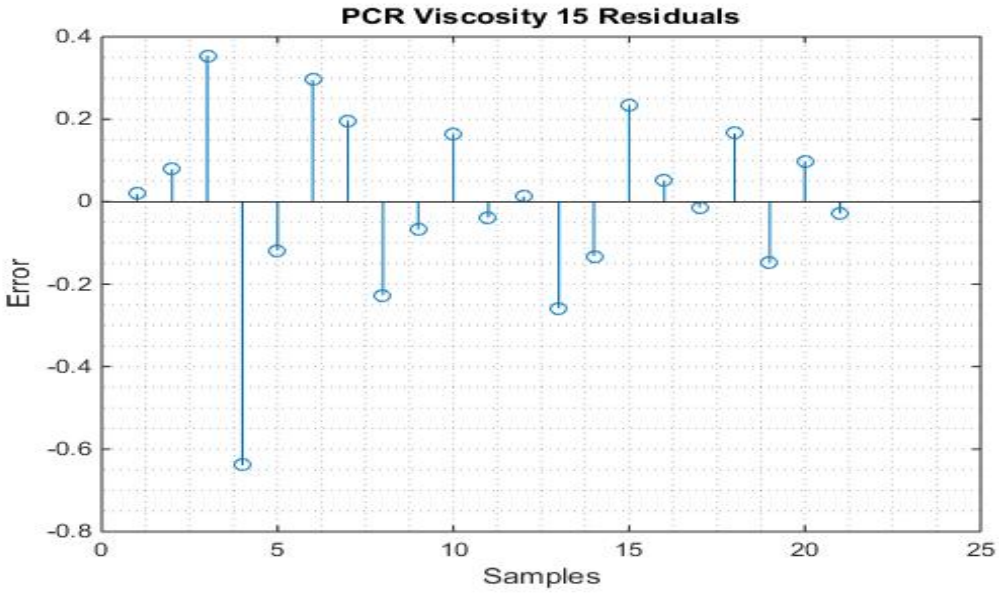


Figure 7.19: 2500-3500 St PCR Residuals

Figure 7.20: 2500-3500 St PCR Viscosity

St PLS.jpg

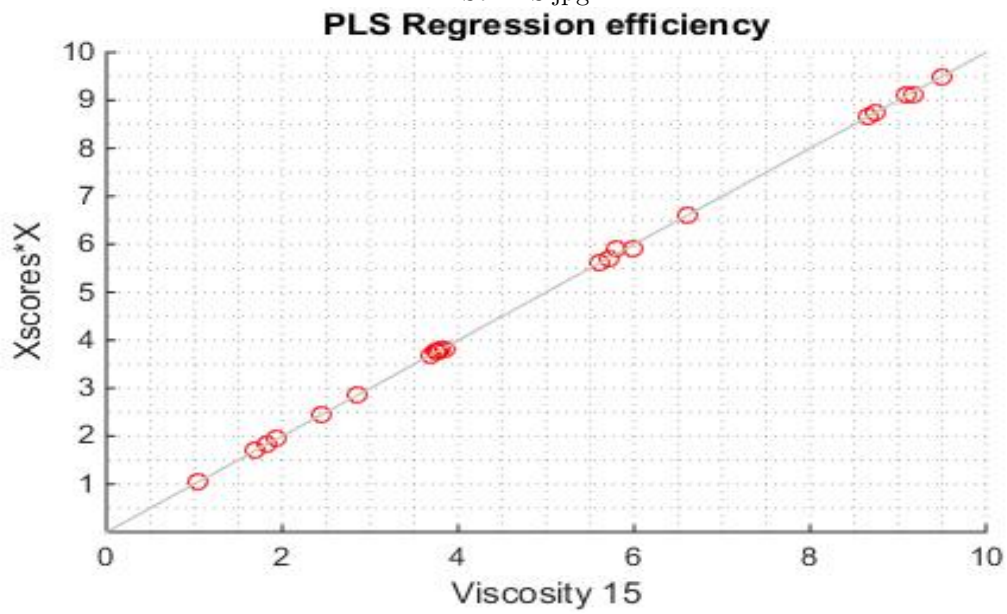


Figure 7.21: ALL St PLS

St PLS Residuals.jpg

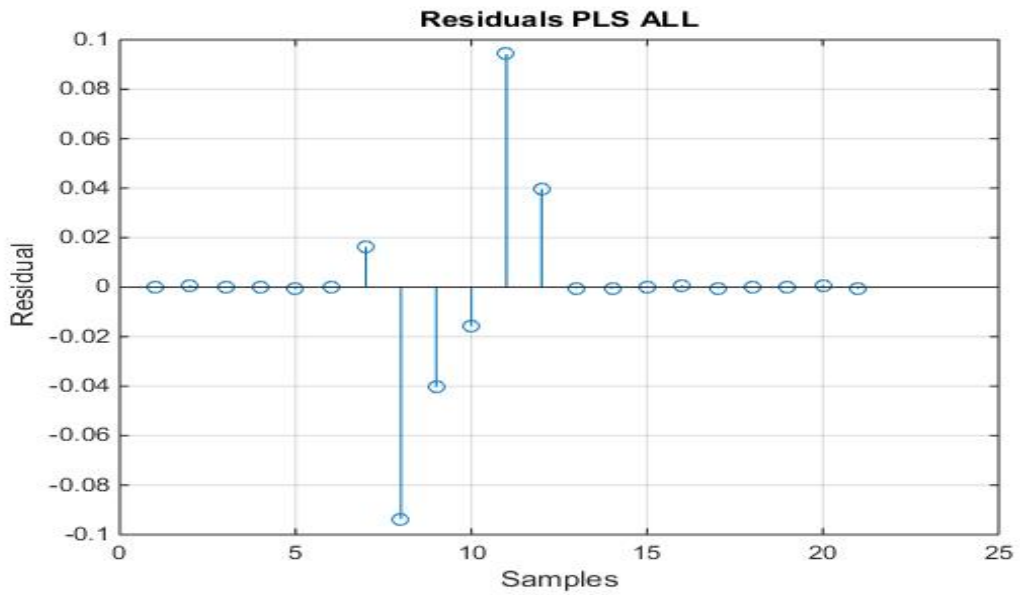


Figure 7.22: ALL St PLS Residuals

Figure 7.23: ALL St PLS Viscosity

St PCR.jpg

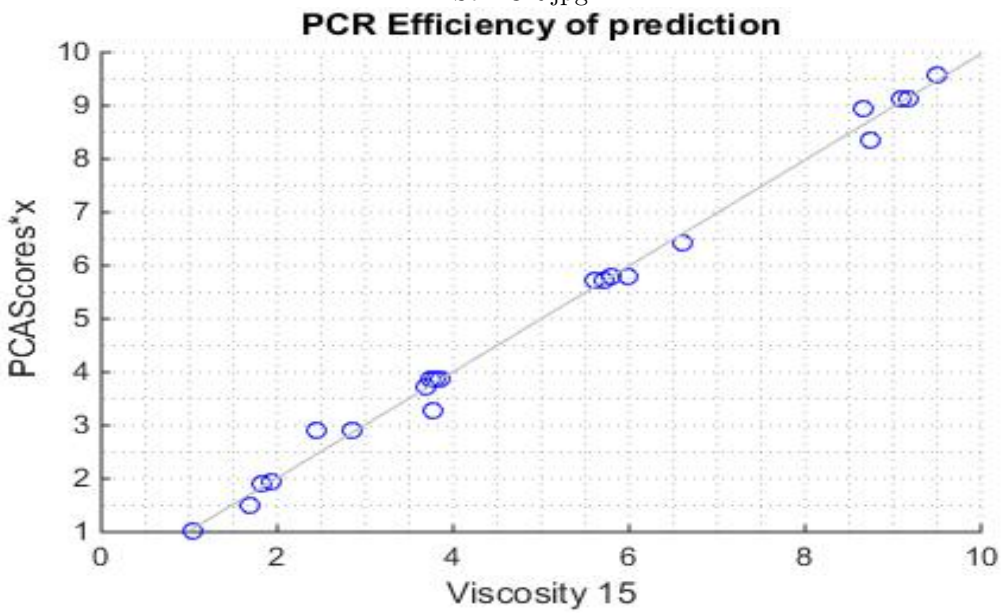


Figure 7.24: ALL St PCR

St PCR Residuals.jpg

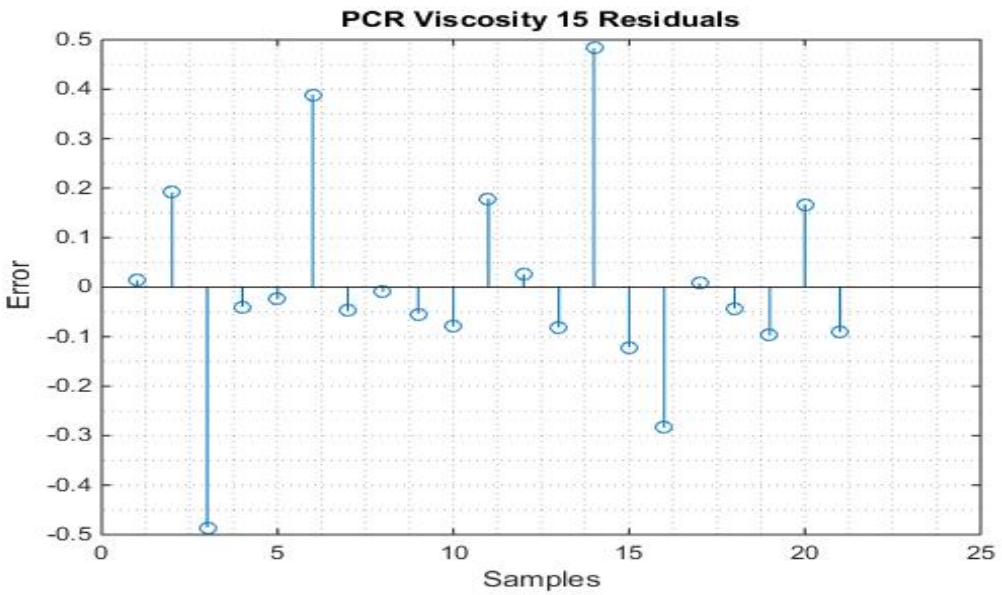


Figure 7.25: ALL St PCR Residuals

Figure 7.26: ALL St PCR Viscosity

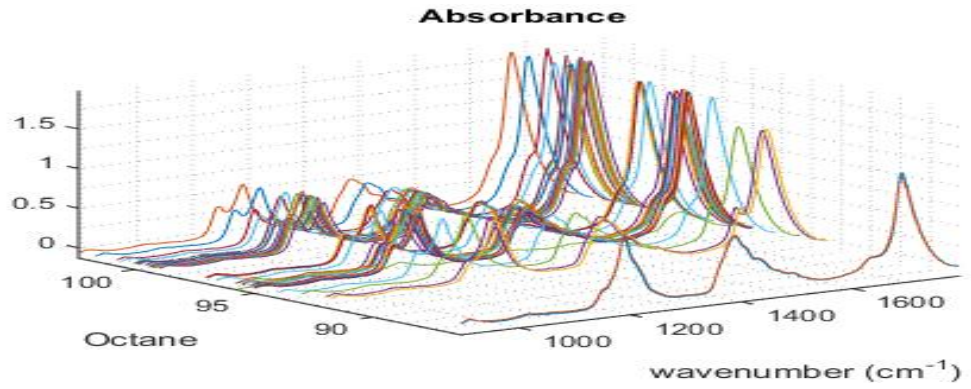


Figure 7.27: Absorbance Plot Octane

7.6 Μοντέλα πρόβλεψης Αριθμού Οκτανίων

Δόθηκαν δεδομένα από Φασματοσκοπία κοντά στην υπέρυθρη ακτινοβολία (NIR-Spectroscopy) για τον υπολογισμό του αριθμού Οκτανίου από κλάσματα βενζίνης. Συνολικά δόθηκαν 44 δεδομένα φασματοσκοπίας με τα αντίστοιχα δεδομένα Οκτανίων. Το φάσμα κυμαίνεται από 895-1792 cm^{-1} με βήμα 3.5. Τα δεδομένα που δώθηκαν για τον αριθμό οκτανίου είναι υπολογισμένα με την μέθοδο (RON- Research Octane Number).
Μετά από βιβλιογραφική μελέτη και την μελέτη που παρουσιάζεται παρακάτω συμπεραίνεται ότι το καλύτερο μοντέλο με το μικρότερο μέσο τετραγωνικό σφάλμα υπάρχει στο μοντέλο της PLS με κανονικοποιημένα δεδομένα ενώ έχει αφαιρεθεί η μέση τιμή και ταυτόχρονα να διαιρεθεί με την τυπική απόκλιση των δεδομένων.
Εν συνεχεία απεικονίζονται διαγράμματα PLS αλλά και PCR. Ο αλγόριθμος όπως αποφαίνεται μπορεί να έχει πάρα πολύ καλά αποτελέσματα πρόβλεψης για τον αριθμό Οκτανίων, καθώς το μέσο τετραγωνικό σφάλμα που προβλέπεται είναι RMSEP=0.0034.
Η πρόβλεψη που υπολογίζεται από το παρακάτω πρόγραμμα έρχεται ταύτηση με τα βιβλιογραφικά δεδομένα.
Επίσης παρατίθενται τα διάγραμματα διακύμανσης, μέσου τετραγωνικού σφάλματος, PCA αλλά και τα μοντέλα πρόβλεψης με PCR και PLS. Για όλο το φάσμα.
Τα δεδομένα απορρόφησης Οκτανίων θα χωριστούν σε 3 ομάδες για την μελέτη τους, για τις μεθοδολογίες PLS και PCR. Απο 900-1400, 1400-1790 και σε όλο το φάσμα.

Method/Range	RMSEP	SEP
PLS 900-1400	0.0043	0.0044
PLS 1400-1790	0.0206	0.0208
PLS ALL	0.0034	0.0034
PCR 900-1400	0.6348	0.6421
PCR 1400-1790	0.9212	0.9319
PCR ALL	0.8119	0.8213

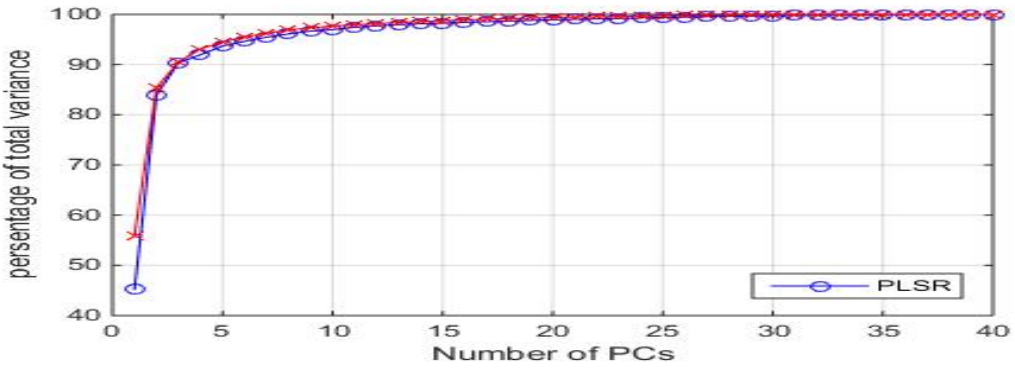


Figure 7.28: PLS-PCR Variance Octane

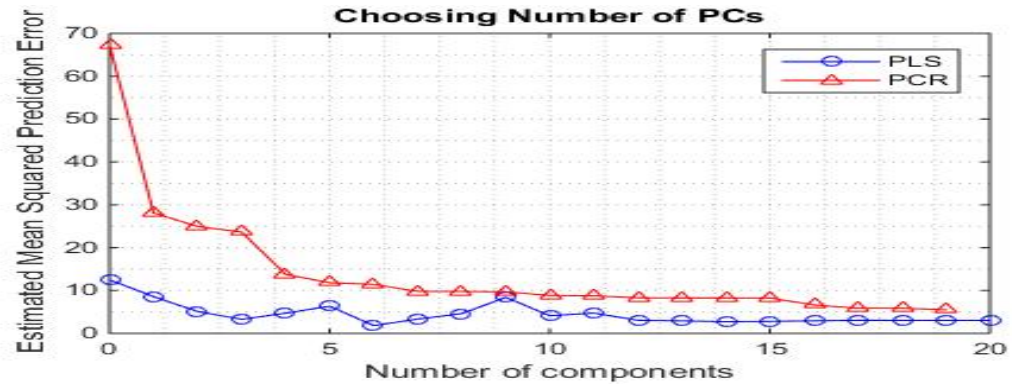


Figure 7.29: RMSEP Octane

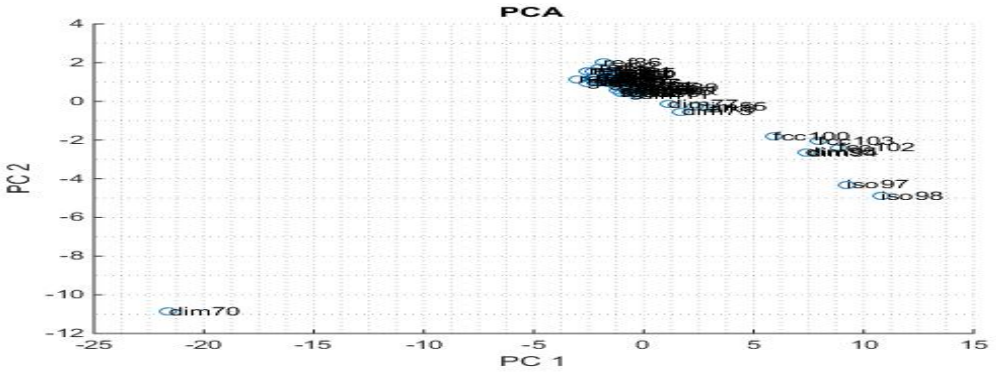


Figure 7.30: PCA Octane

Figure 7.31: PLS Efficiency Octane

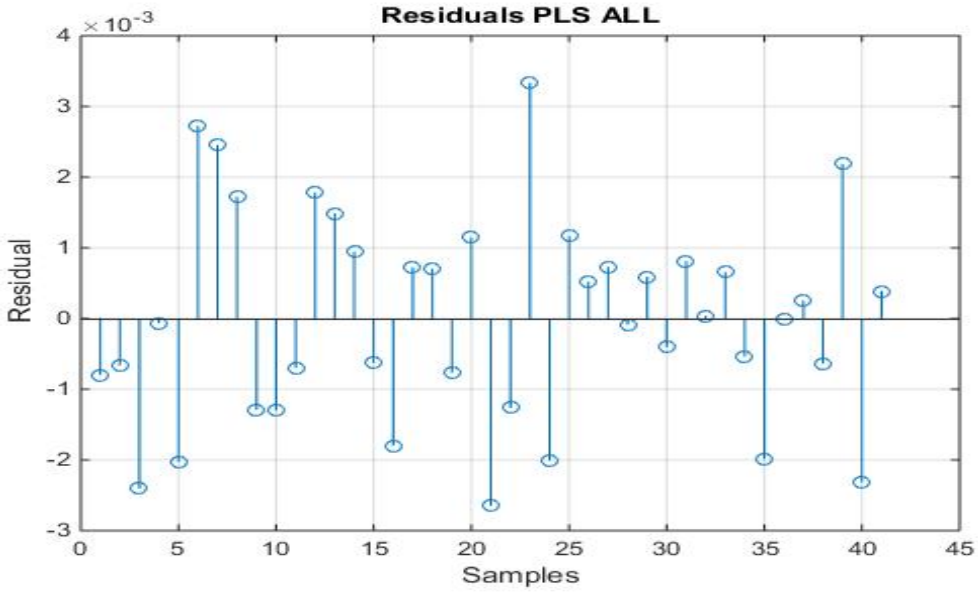


Figure 7.32: PLS Residuals Octane

Figure 7.33: PLS Octane

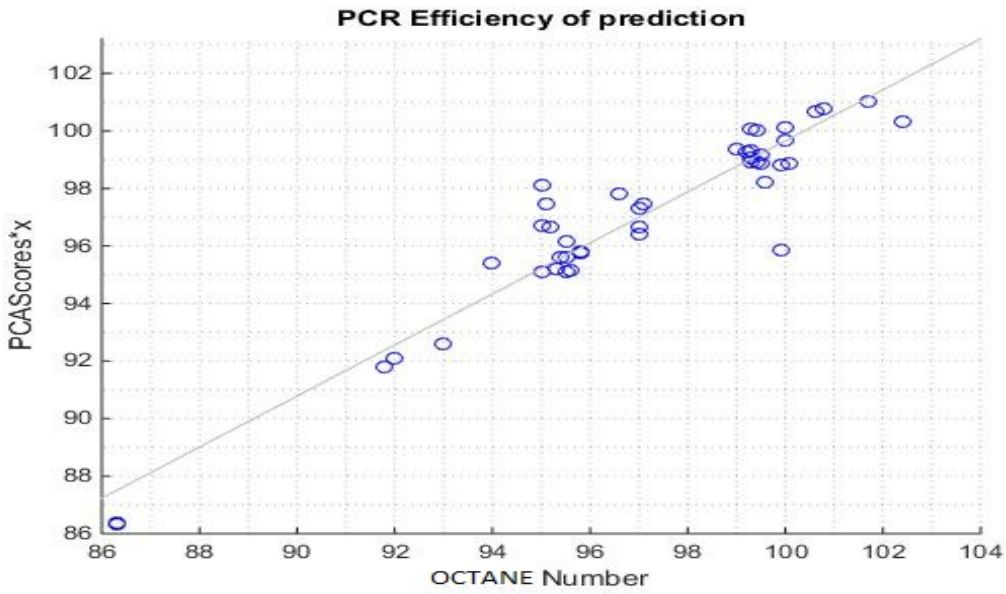


Figure 7.34: PCR Efficiency Octane

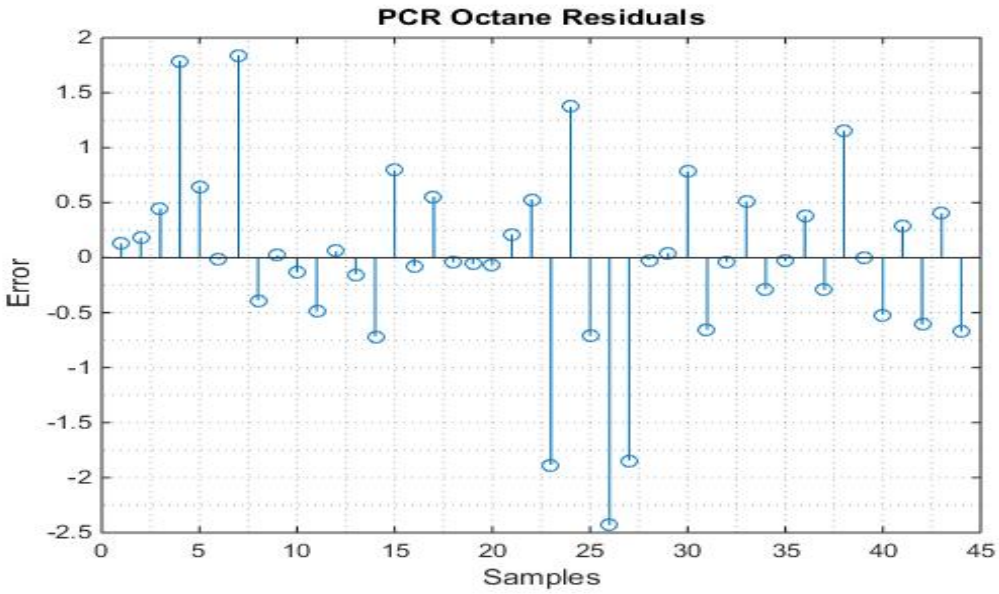


Figure 7.35: PCR Residuals

Figure 7.36: PCR Octane

7.7 Προτάσεις για βελτίωση μεθοδοδολογιών για καλύτερα αποτελέσματα

- Εισαγωγή περισσοτέρων δειγμάτων στο Μοντέλο εκπαίδευσης των δεδομένων για αύξηση της ακρίβειας.
- Εισαγωγή σωστών μεταβλητών που προέρχονται από μεθοδολογίες Variable Selection όπως η iPLS ή από εμπειρογνώμονες.
- Διόρθωση αρχικών τιμών φασματοσκοπήματος, καλύτερο Calibration του οργάνου φασματογραφίας.
- Προσοχή στην επιλογή σωστού αριθμού κύριων συνιστωσών.
- Σωστή μελέτη της επεξεργασίας δεδομένων, εφαρμογή Κανονικοποίησης, Ομαλοποίησης, και Κλιμάκωση ή και αφαίρεση εσφαλμένων τιμών από το σύνολο των δεδομένων, και στις εξαρτημένες και στις ανεξάρτητες μεταβλητές.
- Ευθυγράμμιση, εάν χρειάζεται, των κορυφών του φασματοσκοπήματος στους σωστούς κυματάριθμους.
- Εφαρμογή προγραμμάτων για μείωση θορύβου του αρχικού Μοντέλου εκπαίδευσης.
- Ορθή χρήση και σωστή επιλογή δειγμάτων για φασματοσκοπικές διαδικασίες.
- Εν τέλει, θα πρέπει να γίνονται πολλές τέτοιες πειραματικές εφαρμογές με διαφορετικές μεθοδολογίες και οργανολογία, ώστε να γίνεται ένα cross examination των αποτελεσμάτων.

List of Figures

1 TUC 1

3.1 Partial Least Squares 13
3.2 Dimensionality Reduction 14
3.3 PCA Classification example 15
3.4 Principal Components 16
3.5 Singular Value Decomposition 17

5.1 images/Cetane 23
5.2 Infra Red Spectrum 26

6.1 All- Spectra Variables 30
6.2 1500-2500 - Spectra Variables 30
6.3 2500-3500 - Spectra Variables 30
6.4 Absorbance 32
6.5 Transmittance 33
6.6 Absorbance with Deleted Samples 35
6.7 Standardized -Scaled Samples 36
6.8 PCA Plot(ALL-29-St-Sc) 38
6.9 Cross Validation PCR-PLS (10 PCs-ALL-31-St) 40
6.10 PCR 42
6.11 PLSR 42
6.12 PCA Variance Explained (All-31-St) 45
6.13 PLS Variance Explained(All-31-St) 45
6.14 Number of Components Selection (All-29-NSt) 47
6.15 PCA Loadings 50
6.16 PLS Loadings 50
6.17 Regression PLS (20 PCs-All-31-St) 52
6.18 Regression PLS errors 52
6.19 Data Cetane Number Bar Plot 55
6.20 PCA Prediction Efficiency (20PCs-28-St) 57
6.21 PCA Error Stem 57

7.1 PLS 1500-2500 64
7.2 PCR 1500-2500 64
7.3 PLS 2500-3500 65

7.4 PCR 2500-3500 . . . 65
7.5 PLS ALL . . . 66
7.6 PCR ALL . . . 66
7.7 PCA Viscosity . . . 68
7.8 Viscosity error PLS PCR . . . 68
7.9 1500-2500 St PLS . . . 69
7.10 1500-2500 St PLS Residuals . . . 69
7.11 1500-2500 St PLS Viscosity . . . 69
7.12 1500-2500 St PCR . . . 70
7.13 1500-2500 St PCR Residuals . . . 70
7.14 1500-2500 St PCR Viscosity . . . 70
7.15 2500-3500 St PLS . . . 71
7.16 2500-3500 St PLS Residuals . . . 71
7.17 2500-3500 St PLS Viscosity . . . 71
7.18 2500-3500 St PCR . . . 72
7.19 2500-3500 St PCR Residuals . . . 72
7.20 2500-3500 St PCR Viscosity . . . 72
7.21 ALL St PLS . . . 73
7.22 ALL St PLS Residuals . . . 73
7.23 ALL St PLS Viscosity . . . 73
7.24 ALL St PCR . . . 74
7.25 ALL St PCR Residuals . . . 74
7.26 ALL St PCR Viscosity . . . 74
7.27 Absorbance Plot Octane . . . 75
7.28 PLS-PCR Variance Octane . . . 76
7.29 RMSEP Octane . . . 76
7.30 PCA Octane . . . 76
7.31 PLS Efficiency Octane . . . 77
7.32 PLS Residuals Octane . . . 77
7.33 PLS Octane . . . 77
7.34 PCR Efficiency Octane . . . 78
7.35 PCR Residuals . . . 78
7.36 PCR Octane . . . 78

Bibliography

[1] Marini, Federico, et al. "Artificial neural networks in chemometrics: History, examples and perspectives." Microchemical journal 88.2 (2008): 178-185.

[2] Skorupska, N.M., "Coal Specifications-Impact on Power Station Performance", IEACR/52,IEA Coal Research, London 1993.

[3] Reris, Robert, and J. Paul Brooks. "Principal Component Analysis and Optimization: A Tutorial." (2015): 212.

[4] Springer Series in Statistics. I.T Jolliffe, Principal Component Analysis (Second Edition)

[5] Reris, Robert, and J. Paul Brooks. "Principal Component Analysis and Optimization: A Tutorial." (2015): 212.

[6] Tanskanen, Antti, Jani Lukkarinen, and Kari Vatanen. "Random factor approach for large sets of equity time-series." arXiv preprint arXiv:1604.05896 (2016).

[7] Guyon, Isabelle; Elisseeff, André (2003). "An Introduction to Variable and Feature Selection. JMLR. 3

[8] web.as.uky.edu/statistics/users/pbreheny/764-F11/notes/8-30.pdf

[9] Migon, Helio S., Dani Gamerman, and Francisco Louzada. Statistical inference: an integrated approach. CRC press, 2014.

[10] Song, Chunsham. Chemistry of diesel fuels. CRC Press, 2000.

[11] Diesel fuel characteristics and resources". ufa.com. UFA. 2009. Retrieved 18 July 2014.

[12] International Organization for Standardization - Site

[13] Bamgboye, A. I., and A. C. Hansen. "Prediction of cetane number of biodiesel fuel from the fatty acid methyl ester (FAME) composition." International Agrophysics 22.1 (2008): 21.

[14] Cairns, T., McWilliam, I. G., Pecsok, R. L., Shields, L. D. (1976). Modern methods of chemical analysis. Pecsok. New York: John Wiley Sons.

[15] Sánchez-Borroto, Yisel, et al. "Prediction of cetane number and ignition delay of biodiesel using Artificial Neural Networks." Energy Procedia 57 (2014): 877-885.

[16] Piloto-Rodríguez, Ramón, et al. "Prediction of the cetane number of biodiesel using artificial neural networks and multiple linear regression." Energy Conversion and Management 65 (2013): 255-261.

[17] Ladommatos, Nicos, and John Goacher. "Equations for predicting the cetane number of diesel fuels from their physical properties." Fuel 74.7 (1995): 1083-1093.

[18] Gopinath, A., Sukumar Puhan, and G. Nagarajan. "Relating the cetane number of biodiesel fuels to their fatty acid composition: a critical study." Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering 223.4 (2009): 565-583.