



ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Φασματική ανάλυση μιγμάτων πετρελαίου

Εξεταστική επιτροπή: Καθ. Μιχάλης Ζερβάκης (επιβλέπων)

Καθ. Πασαδάκης Νικόλαος

Καθ. Κώστας Μπάλας

Jaohar Osman
A.M 2003030004

Χανιά, 2013

Περιεχόμενα

1.Εισαγωγή.....	5
2. Η φασματοσκοπία.....	8
3. Υπέρυθρη φασματοσκοπία(Infrared Spectroscopy).....	10
4. Η Φασματοσκοπία εγγύς υπερύθρου (NIR Spectroscopy) και η βενζίνη.....	16
4.1 Οι τρεις περιοχές υπερύθρων (Near-IR, Mid-IR, Far-IR).....	16
4.2 Η σημασία της NIR.....	19
4.3 Η βενζίνη.....	20
4.4 Αρωματικοί υδρογονάνθρακες.....	22
4.5 Δακτύλιο βενζολίου.....	23
4.6 Το βενζόλιο και τα παράγωγα βενζολίου.....	24
4.7 Πολυαρωματικοί υδρογονάνθρακες.....	26
4.8 Η ιστορία της NIR.....	30
4.9 Πλεονεκτήματα της NIR.....	31
4.10 Τα όργανα.....	33
5. Εφαρμογές της NIR.....	37
5.1 Εισαγωγή.....	37
5.2 Αστρονομική φασματοσκοπία.....	38
5.3 Γεωργία.....	39
5.4 Απομακρυσμένη παρακολούθηση.....	39
5.5 Επιστήμη των υλικών.....	39
5.6 Ιατρικές χρήσεις.....	39
5.7 Μέτρηση σωματιδίων.....	42
5.8 Βιομηχανικές χρήσεις.....	42
6. PARTIAL LEAST-SQUARES REGRESSION.....	43
6.1 Η σημασία της PLS μεθόδου.....	43
6.2 Η PLS μέθοδος, ξεκινώντας από άλλες τεχνικές πολλαπλών μεταβλητών.....	44

6.2.1 Πολλαπλή γραμμική παλινδρόμηση (MULTIPLE LINEAR REGRESSION (MLR)).....	44
6.2.2 Πολλαπλή γραμμική παλινδρόμηση (MULTIPLE LINEAR REGRESSION (MLR)) με περισσότερες από μία εξαρτημένη μεταβλητή.....	45
6.2.3 Ανάλυση Κύριων Συνιστωσών (PRINCIPAL COMPONENT ANALYSIS (PCA)): ΜΕΘΟΔΟΣ NIPALS.....	46
6.2.4 Παλινδρόμηση Κύριων Συνιστωσών.....	49
6.2.5 Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων (PARTIAL LEAST-SQUARES REGRESSION).....	49
6.2.6 Η πρόβλεψη.....	52
6.3 Πώς λειτουργεί η PLS.....	53
6.4 Βήματα Βαθμονόμησης (κατάρτισης) και πρόβλεψης (δοκιμής) (Calibration (training) and prediction (test) steps).....	55
6.5 Κεντράρισμα και η κλιμάκωση των μεταβλητών (Mean-centering and scaling of variables).....	55
7. Ανάπτυξη του προτεινόμενου μοντέλου.....	58
7.1 Εισαγωγή.....	58
7.2 Ξεκίνημα της διαδικασίας ανάπτυξης και εξέταση διαφόρων περιπτώσεων.....	59
7.3 INTERACTIVE VARIABLE SELECTION (IVS) FOR PLS.....	62
7.4 Ενσωμάτωση των περιοχών συγκλίσεων των καλύτερων προβλέψεων στη συνάρτηση plsDip και ολοκλήρωση του μοντέλου.....	66
Συμπεράσματα.....	70

1. Εισαγωγή

Στην παρούσα διπλωματική εργασία χρησιμοποιούμε τη φασματοσκοπία NIR (Near-IR Spectroscopy) για την ανάπτυξη του προτεινόμενου μοντέλου πρόβλεψης τιμών ποσοτήτων του αρωματικού περιεχομένου (Aromatic Content) σε διάφορα είδη δειγμάτων βενζίνης με τη βοήθεια της τεχνικής βαθμονόμησης Partial Least Squares.

Η περιοχή υπέρυθρων αποτελείται από τρεις περιοχές, Near-IR, Mid-IR, Far-IR. Η φασματοσκοπία NIR (Near-IR Spectroscopy) είναι μια δονητική μέθοδος φασματοσκοπίας (vibrational spectroscopic method) που ανήκει στο υπέρυθρο φάσμα φωτός που είναι πολύ κοντά στην ορατή περιοχή (από περίπου 750 nm έως 2500 nm ή 12820 cm^{-1} έως 4000 cm^{-1}), όπου οι περισσότερες οργανικές και επίσης μερικές ανόργανες ενώσεις έχουν εξαιρετική ανακλαστικότητα ή ιδιότητες διαπερατότητας, πράγμα που σημαίνει ότι είναι σε θέση να απορροφήσουν πολύ φως στην ακτινοβολία NIR.

Η χρήση της οπτικής μέτρησης χρονολογείται από τη δεκαετία του 1940. Οι πρώτες δοκιμές των οπτικών εφαρμογής έτειναν προς τη μέτρηση των βασικών προϊόντων αντί τη μονότονη διαδικασία της μέτρησης, η οποία ήταν πολύπλοκη, χρονοβόρα και δαπανηρή. Με τον καιρό, ιδιαίτερα συμφέροντα στη ορατή (VIS) και εγγύς υπέρυθρη (NIR) φασματοσκοπία έχουν επεκταθεί σημαντικά και έχουν κατευθυνθεί προς την μέτρηση on-line και τον έλεγχο της ποιότητας των διαφόρων διαδικασιών στον τομέα της γεωργίας και της βιομηχανίας τροφίμων. Κατά την ανάλυση συνθετικών και φυσικών υλικών η φασματοσκοπία NIR έχει δείξει πρωτοφανή βιομηχανική επιτυχία σε πολλαπλές εφαρμογές χρησιμοποιώντας δεκάδες χιλιάδες όργανα για σπόρους, ζωοτροφές, ψήσιμο των προϊόντων, αλεύρι, ποτά, τροφές, φαρμακευτικά προϊόντα, γαλακτοκομικά προϊόντα, υδρογονάνθρακες και πετροχημικά, χημικά προϊόντα, ραδιενεργά και επικίνδυνα υλικά, ιατρική απεικόνιση και διαγνωστικά.

Τα κυρίαρχα φασματικά χαρακτηριστικά της NIR περιλαμβάνουν: τις δονήσεις τάσης (stretching vibrations) C-H του μεθυλίου (CH_3), τις δονήσεις τάσης (stretching vibrations) C-H του μεθυλενίου (CH_2), δονήσεις τάσης C-H του αρωματικού (aromatic C-H stretching vibrations) και τις O-H δονήσεις τάσης (stretching vibrations). Ελάχιστα αλλά εξίσου σημαντικά φασματικά χαρακτηριστικά περιλαμβάνουν: μεθόξυ ($\text{O}-\text{CH}_3$) δόνηση τάσης C-H, δόνηση τάσης C-H συσχετιζόμενη με καρβονύλιο ($\text{C}=\text{O}$) (carbonyl associated C-H stretching), N-H από τα πρωτογενή αμίδια και δευτεροταγή αμίδια αμφότερα, αλκύλια ($\text{C}_n\text{H}_{2n+1}$) και αρυλ ομάδες ενώσεων, N-H από τις πρωτοταγείς, δευτεροταγείς, τριτοταγείς αμίνες, και N-H από άλατα αμίνης.

Οι βασικές χρήσεις της φασματοσκοπίας NIR είναι για τον έλεγχο της διαδικασίας, για την αξιολόγηση της ποιότητας, για την ταυτοποίηση των πρώτων υλών και την επεξεργασία υποπροϊόντων και για τη χημική ποσοτική ανάλυση των πολύπλοκων μειγμάτων.

Κατά τη μέτρηση μιγμάτων υδρογονανθράκων, όπως είναι τα καύσιμα ή μείγματα διαλυτών, η NIR δεν μετράει άμεσα κλάσεις υδρογονανθράκων όπως ολεφίνες ή ναφθένια, αλλά μετρά απορροφήσεις λειτουργικής ομάδας, όπως μεθύλιο, μεθυλένιο, μεθίνης, δονήσεις τάσης αρωματικών και δονήσεις παραμόρφωσης. Οι αναλογίες των διαφόρων αυτών απορροφήσεων όταν συσχετίζονται χρησιμοποιώντας διάφορες πολυμεταβλητές

τεχνικές βαθμονόμησης σε γνωστές φυσικές ή συνθετικές παραμέτρους για ένα σετ μάθησης παρέχουν μια εκτίμηση συσχέτισης της σύνθεσης κατηγορίας υδρογονανθράκων από διάφορα άγνωστα πολύπλοκα μίγματα.

Η βενζίνη απαιτεί κάποιες ελάχιστες ιδιότητες. Ωστόσο, οι τυποποιημένες μετρήσεις για αυτές τις ιδιότητες είναι δαπανηρές, χρονοβόρες και περίπλοκες. Για παράδειγμα, αριθμός οκτανίου (octane number), η πιο σημαντική ιδιότητα, είναι η ιδιότητα απόδοσης του καυσίμου της βενζίνης που έντονα σχετίζεται με την ικανοποίηση του καταναλωτή. Είναι αποφασισμένο από το πρότυπο ASTM-CFR δοκιμής κινητήρων. Άλλο ένα παράδειγμα είναι το αρωματικό περιεχόμενο (aromatic content). Οι αρωματικοί υδρογονάνθρακες χρειάζονται ιδιαίτερη προσοχή εξαιτίας του ότι προκαλούν μεταλλάξεις και καρκινογένεσεις. Αυτές οι ενώσεις εξετάζονται από τις αρχές διότι επηρεάζουν την ποιότητα του αέρα και υπόκεινται στους κανονισμούς περί τοξικότητας του αέρα. Εξάλλου ως πρόσθετο βενζίνης, οι αρωματικοί υδρογονάνθρακες αυξάνουν τον βαθμό οκτανίων και μειώνουν τα χτυπήματα. Το αρωματικό περιεχόμενο του καυσίμου είναι σημαντικό στον προσδιορισμό των χαρακτηριστικών ανάφλεξης του καυσίμου, αφού τα αρωματικά αναφλέγονται βραδύτερα από τις παραφίνες στον κινητήρα ντίζελ (diesel engine). Τα όργανα που απαιτούνται για τη μέτρηση είναι δαπανηρά, απαιτούν συνεχή συντήρηση, η διαδικασία διαρκεί 20 λεπτά ανά δείγμα και δεν είναι κατάλληλα για εφαρμογές on-line, όπως λειτουργίες ανάμιξης. Φασματική ανάλυση σε συνδυασμό με πολυμεταβλητή βαθμονόμηση παρέχει μη καταστρεπτική μέτρηση πολλών χημικών ενώσεων και έχει αποδείξει την αποτελεσματικότητά της για εργαστηριακές και βιομηχανικές εφαρμογές, ειδικά στη βιομηχανία πετρελαίου. Από το 1989, εγγύς υπέρυθρη (NIR) φασματοσκοπία έχει χρησιμοποιηθεί εκτενώς για τις ιδιότητες δοκιμής βενζίνης, χάρη στις απλές και ξεκάθαρες ιδιότητες μετάδοσης οπτικών ινών και την αντίσταση των οπτικών ινών της NIR προς την περιεκτικότητα του νερού. Μέθοδοι βαθμονόμησης σε συνδυασμό με φασματική ανάλυση, παρέχουν ένα ισχυρό εργαλείο για την πρόβλεψη ιδιοτήτων βενζίνης. Πολλές μέθοδοι βαθμονόμησης χρησιμοποιούνται, όπως Partial Least Squares (PLS) και άλλες γραμμικές και μη γραμμικές μέθοδοι βαθμονόμησης. Από αυτές τις μεθόδους, η PLS έχει αποδειχθεί ότι είναι μια ισχυρή τεχνική για τη μοντελοποίηση της διαδικασίας και τη βαθμονόμηση, όπου οι μεταβλητές πρόβλεψης (predictor variables) είναι συγγραμμικές (collinear), τα δεδομένα μέτρησης περιέχουν θόρυβο, οι μεταβλητές έχουν μεγάλη διάσταση και όπου υπάρχουν λιγά δείγματα.

Η PLS είναι μια γενικευμένη μέθοδος παλινδρόμησης (generalized regression method) που χρησιμοποιείται για την κατασκευή ενός μοντέλου πρόβλεψης μεταξύ των δύο μπλοκ μεταβλητών: το Χ-μπλοκ των μεταβλητών πρόβλεψης (predictor variables) και το Υ-μπλοκ των μεταβλητών απόκρισης (response variables). Η Partial least squares (PLS) αποκτά μεγαλύτερη σημασία σε πολλούς τομείς της χημείας, της αναλυτικής, φυσικής, κλινικής χημείας και της βιομηχανικής.

Με βάση τα παραπάνω, μέτρησα κάποια δείγματα βενζίνης στο εργαστήριο με τη βοήθεια της NIR(Near-IR) του εργαλείου Avantas, και προχώρησα για να αναλύσω τα φάσματα των μετρήσεων για να δω το περιεχόμενο τους, για να το κάνω αυτο χρησιμοποιήσα τη PLS μέθοδο και τα ανέλυσα σε συνιστώσεις, και μετά μέτρησα το περιεχόμενο αυτών των συνιστωσών και κράτησα τις συνιστώσες των οποίων τα περιεχόμενα είναι τα κοντινότερα στις αληθινές τιμές ποσοτήτων του αρωματικού περιεχομένου (Aromatic Content) των

διαφόρων ειδών δειγμάτων βενζίνης. Αυτές οι συνιστώσες μπορούν να χρησιμοποιηθούν για πρόβλεψη του αρωματικού περιεχομένου καινούριων δειγμάτων βενζίνης με άγνωστες τιμές αρωματικού περιεχομένου.

Ακολουθεί συντομή περιγραφή των επόμενων κεφαλαίων:

Στο κεφάλαιο 2 αναλύεται η φασματοσκοπία και οι χρήσεις της. Στο κεφάλαιο 3 εστιάζουμε στην υπέρυθρη φασματοσκοπία, στις εφαρμογές και στις δυνατότητές της. Στο κεφάλαιο 4 παρουσιάζονται τρεις περιοχές του υπέρυθρου φάσματος (Near-IR, Mid-IR, Far-IR) με ιδιαίτερη έμφαση στην περιοχή Near-IR (NIR). Εξηγούνται τα κυρίαρχα φασματικά χαρακτηριστικά της NIR για το περιεχόμενο της βενζίνης, τους Αρωματικούς υδρογονάνθρακες, το βενζόλιο, τα παράγωγα βενζολίου και τούς πολυαρωματικούς υδρογονάνθρακες. Στη συνέχεια εξηγούνται οι κορυφές NIR (NIR overtones) που αναμένονται στα φάσματα βενζίνης και εξηγείται η πειραματική διαδικασία που ακολουθήθηκε για να λάβουμε φάσματα NIR. Στο κεφάλαιο 5 παρουσιάζονται εφαρμογές της NIR σε διάφορους τομείς έρευνας. Το κεφάλαιο 6 παρουσιάζει τη μέθοδο PARTIAL LEAST-SQUARES REGRESSION (PLS) που χρησιμοποιήσαμε για την ανάλυση δειγμάτων. Αναλύεται η συσχέτιση της μεθόδου με άλλες τεχνικές πολλαπλών μεταβλητών, όπως, principal components regression analysis, multiple linear regression και maximum redundancy analysis. Επίσης σε αυτό το κεφάλαιο αναλύονται τα βήματα Βαθμονόμησης και πρόβλεψης καθώς και το κεντράρισμα και η κλιμάκωση των μεταβλητών (Mean-centering and scaling of variables). Στο κεφάλαιο 7 γίνεται η ανάπτυξη του προτεινόμενου μοντέλου πρόβλεψης ποσοτήτων του αρωματικού περιεχομένου (Aromatic Content) σε διάφορα είδη δειγμάτων βενζίνης. Χρησιμοποιώντας τα 47 φάσματα βενζίνης που έχουμε πάρει στο εργαστήριο, Το προτεινόμενο μοντέλο βασίζεται στις μεθόδους Cross-Validation, Interactive Variable Selection for PLS, Partial Least Squares (PLS). Η εργασία κλείνει με συμπεράσματα και μελλοντικές κατευθύνσεις.

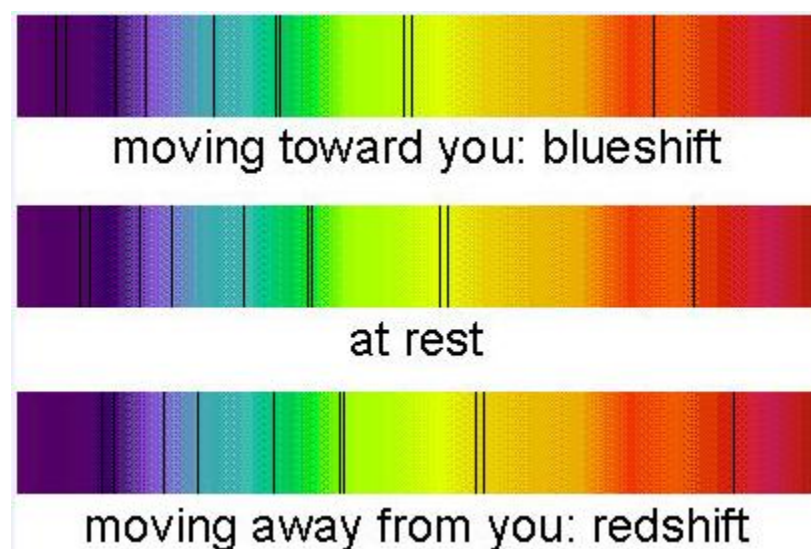
2. Η φασματοσκοπία

Η φασματοσκοπία είναι η μελέτη της αλληλεπίδρασης μεταξύ της ακτινοβολίας (είτε είναι ηλεκτρομαγνητική είτε είναι ακτινοβολία σωματιδίων) και της ύλης της οποίας περιλαμβάνει άτομα και μόρια , ενώ οι φασματικές μετρήσεις είναι η μέτρηση των αλληλεπιδράσεων αυτών σαν τη διαδικασία της απορρόφησης , εκπομπής ή σκέδασης του ηλεκτρομαγνητικού φάσματος.

Η φασματοσκοπία χρησιμοποιείται στην φυσικοχημεία και στην αναλυτική χημεία για την ποιοτική και ποσοτική ανάλυση των χημικών ουσιών, είτε είναι ατομικά χρησιμοποιώντας τα ατομικά φάσματα αυτών των στοιχείων είτε είναι μόρια Χρησιμοποιώντας οπτικές ακτίνες , UV ακτίνες ορατού υπεριώδους ή ακτίνες Υπερύθρων.

Τα φασματοσκόπια χρησιμοποιούνται επίσης στην αστρονομία και τη τηλεπισκόπηση. Τα μεγάλα τηλεσκόπια έχουν πάντα φασματόμετρο για τη μέτρηση , τη χημική σύνθεση , τις φυσικές ιδιότητες των αστρονομικών αντικειμένων ή τη μέτρηση των ταχυτήτων ανάλογα με τη μετατόπιση Doppler .

Σχήμα 1. μέτρηση των ταχυτήτων των αστρονομικών αντικειμένων με τη μετατόπιση Doppler, στο πρώτο φάσμα το αντικείμενο μετακινείται προς το μέρος του παρατηρητή, στο δεύτερο φάσμα το αντικείμενο παραμένει σταθερό στην διεύθυνση του παρατηρητή και στο τρίτο φάσμα το αντικείμενο απομακρύνεται από τον παρατηρητή .[8]



Στο Σχήμα 2 φαίνονται οι περιοχές του ηλεκτρομαγνητικού φάσματος και οι αντίστοιχες φασματοσκοπικές τεχνικές ανάλυσης .

Για τη διπλωματική χρησιμοποιήσαμε τη τεχνική Near Infrared Spectroscopy(NIR Spectroscopy) του εργαλείου Avantes για να πάρουμε τα φάσματα απορρόφησης των διάφορων ειδών βενζίνης. Η Near Infrared Spectroscopy(NIR Spectroscopy) ανήκει στην Υπέρυθρη φασματοσκοπία (Infrared Spectroscopy) .

Σχήμα 2, οι περιοχές του ηλεκτρομαγνητικού φάσματος και οι αντίστοιχες φασματοσκοπικές τεχνικές ανάλυσης .[9]

Οι περιοχές του ηλεκτρομαγνητικού φάσματος και οι αντίστοιχες φασματοσκοπικές τεχνικές ανάλυσης.

Περιοχή	Μήκος κύματος [Å]	Συχνότητα [Hz]	Ενέργεια [eV]	Είδος μεταβολής	Φασματοσκοπικές τεχνικές ανάλυσης
Ραδιοκύματα (Radio waves)	$> 10^9$	$< 3 \times 10^9$	$< 10^{-5}$	Διέγερση πυρήνα	Φασματοσκοπία πυρηνικού μαγνητικού συντονισμού (Nuclear Magnetic Resonance)
Μικροκύματα (Micro waves)	10^9 έως 10^6	3×10^9 έως 3×10^{12}	10^{-5} έως 0.01	Περιστροφή μορίων	Φασματοσκοπία περιστροφής (Rotational Spectroscopy)
Υπέρυθρη (Infrared)	10^6 έως 7000	3×10^{12} έως 4×10^{14}	0.01 έως 2	Δόνηση δεσμών	Υπέρυθρη φασματοσκοπία (Infrared Spectroscopy)
Ορατή (Visible)	7000 έως 4000	4×10^{14} έως 7×10^{14}	2 έως 3	Ηλεκτρονιακές διεγέρσεις	Φασματοσκοπία ορατού υπεριώδους (UV-Vis spectroscopy)
Υπεριώδη (Ultraviolet)	4000 έως 10	7×10^{14} έως 3×10^{17}	3 έως 10^3	Φθορισμός	
Ακτίνες X (X-Rays)	10 έως 0.1	3×10^{17} έως 3×10^{19}	10^3 έως 10^5	Περίθλαση ακτίνων X & εκπομπή φωτοηλεκτρονίων	Περίθλαση ακτίνων X (XRD) Φασματοσκοπία φωτοηλεκτρονίων (XPS) Φασματοσκοπία ηλεκτρονίων Auger (AES) Φασματοσκοπία φθορισμού (XRF)

3. Υπέρυθρη φασματοσκοπία(Infrared Spectroscopy)

Στην περιοχή του υπέρυθρου φάσματος βρίσκει εφαρμογή η Υπέρυθρη φασματοσκοπία(Infrared Spectroscopy), η οποία θεωρείται μια από τις πλέον δυναμικές τεχνικές ανάλυσης αφού παρέχει σημαντικές πληροφορίες τόσο για τη μοριακή δομή όσο και για τις επιφανειακές ιδιότητες των υλικών. Όταν η υπέρυθρη ακτινοβολία προσπέσει σε ένα υλικό τότε παρατηρούνται μεταβολές στο μήκος των δεσμών (δονήσεις) καθώς και στην περιστροφή των μορίων .

Οι μοριακές δονήσεις είναι ένα από τρία διαφορετικά είδη κίνησης των μορίων που περιλαμβάνουν μεταγραφική κίνηση (όταν το σύνολο μόριο πηγαίνει στην ίδια κατεύθυνση) και περιστροφική κίνηση (όταν το μόριο περιστρέφεται γύρω από το ίδιο). Η συχνότητα της περιοδικής κίνησης είναι γνωστό ως συχνότητα δόνησης, οι τυπικές συχνότητες μοριακών δονήσεων κυμαίνονται λιγότερο από 10^{12} μέχρι 10^{14} Hz.

Η κίνηση των μοριακών δονήσεων είναι όταν οι δεσμοί μεταξύ των ατόμων μέσα σε ένα μόριο κινούνται και ταλαντεύονται. Οι μεταβολές αυτές μπορούν να γίνουν κατανοητές προσομοιώνοντας το χημικό δεσμό με ένα ελατήριο, στα άκρα του οποίου βρίσκονται δυο σφαιρικά σωματίδια τα οποία και επιτελούν το ρόλο των ατόμων.

Σχήμα 3. Προσομοίωση του χημικού δεσμού με ένα ελατήριο στα άκρα του οποίου βρίσκονται δυο σφαιρικά σωματίδια, τα οποία και επιτελούν το ρόλο των ατόμων.[9]

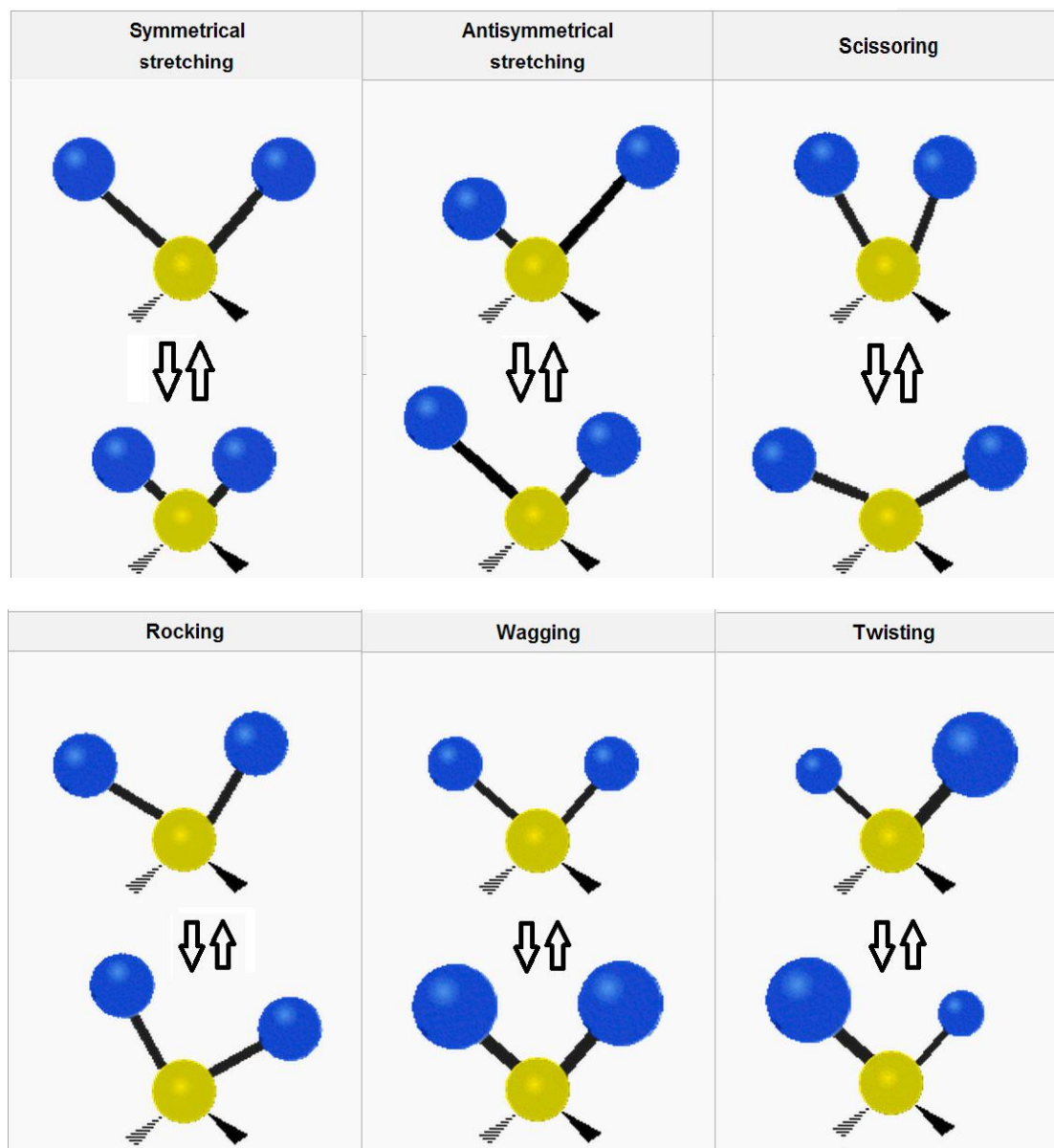


Η θεώρηση αυτή δεν απέχει πολύ από την πραγματικότητα αφού και τα άτομα έχουν σχήμα σφαιρικό και το ηλεκτρονικό νέφος τους(δεσμός) μπορεί να έχει μια σχετική ευκαμψία. Ένα παράδειγμα αυτής της κίνησης είναι "δόνηση τάσης (stretching)". Το πιο απλό παράδειγμα μίας δόνησης για ένα μόριο λαμβάνει χώρα μόνο μεταξύ δύο ατόμων. Μερικά παραδείγματα των μορίων που είναι σαν αυτό είναι του υδρογόνου H_2 , του αζώτου N_2 και του οξυγόνου O_2 .

Αν το μόριο έχει περισσότερα από δύο άτομα τα πράγματα γίνονται πιο περίπλοκα. Υποθέτουμε τώρα ότι προστίθεται ένα ακόμη άτομο. Έτσι τώρα υπάρχουν τρία άτομα, όπως στο νερό H_2O , όπου και τα δύο άτομα υδρογόνου είναι συνδεδεμένα με το κεντρικό άτομο οξυγόνου. Θυμηθείτε με υδρογόνο υπήρχε ένα είδος "δόνησης τάσης (stretching)", αλλά στο νερό υπάρχουν δύο είδη "δόνησης τάσης (stretching)" τα οποία είναι: η συμμετρική και η αντισυμμετρική δόνηση τάσης (symmetrical antisymmetrical stretching). Άλλο ένα είδος δόνησης που υπάρχει ονομάζεται ψαλίδισμα (scissoring).

Τα άτομα σε μια ομάδα $-CH_2$ (βρίσκεται συνήθως σε οργανικές ενώσεις) όπως το νερό έχει τρία άτομα, όμως επειδή η ομάδα $-CH_2$ συνδέεται με άλλο άτομο μπορεί να δονείται σε έξι διαφορετικούς τρόπους: συμμετρική δόνηση τάσης, αντίσυμμετρική δόνηση τάσης (symmetrical, antisymmetrical stretching), ψαλίδισμα (scissoring), λικνίζοντας (rocking), κουνώντας (wagging) και τη συστροφή (twisting) :

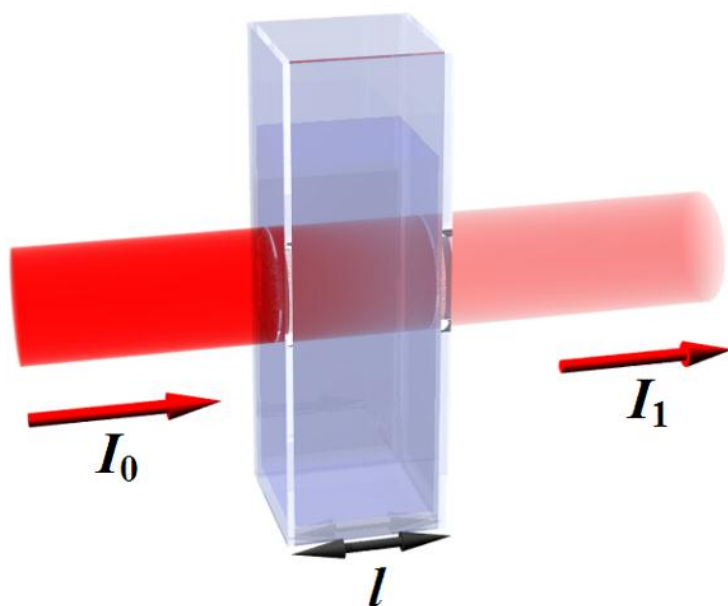
Σχήμα 4, οι έξι δονήσεις των ατόμων ομάδας $-CH_2$.[11]



Με την απορρόφηση της ακτινοβολίας τα άτομα του μορίου δονούνται κατά μήκος του δεσμού, όπως δονούνται οι δυο σφαίρες ενωμένες με ελατήριο. Αυτό είναι ο μηχανισμός με τον οποίο λειτουργεί φασματοσκοπία IR: υπέρυθρη ακτινοβολία διέρχεται μέσα από το δείγμα και η ένταση του μεταδιδόμενου φωτός είναι εντονότερη σε σύγκριση με εκείνη του

εξερχομένου φωτός. Μία μείωση της έντασης σε ένα δεδομένο μήκος κύματος του φωτός δείχνει την απορρόφηση της ενέργειας σε μια δονητική μετάβαση.

Σχήμα 5, απορρόφηση της ενέργειας του φωτός μέσω ενός δείγματος.[12]



η απορρόφητικότητα [absorbance = $-\log(I_1 / I_0)$] ενός δείγματος είναι ανάλογη με το πάχος του δείγματος και την συγκέντρωση των ειδών απορρόφησης στο δείγμα και είναι αντίστροφα ανάλογη με την διαπερατότητα [transmittance (I_1 / I_0)] ενός δείγματος.

Ο νόμος του Hooke εφαρμοζόμενος στο διατομικό μόριο AB δίνει για τη συχνότητα δόνησης του δεσμού A-B

$$\nu = \frac{1}{2\pi} \sqrt{\frac{\kappa}{\mu}}$$

όπου (κ) η σταθερά δύναμης του δεσμού (ή σταθερά ισχύος του δεσμού) σε $\frac{\text{dynes}}{\text{cm}}$ (1 dyne = 10^{-5} Newton) και (μ) η ανηγμένη μάζα (reduced mass), δηλαδή ο αρμονικός μέσος όρος των ατομικών μαζών m_A και m_B :

$$\frac{1}{\mu} = \frac{1}{m_A} + \frac{1}{m_B} \quad \gg \quad \mu = \frac{m_A \cdot m_B}{m_A + m_B}$$

Συνήθως στην παρουσίαση των φασμάτων χρησιμοποιείται για καθαρά πρακτικούς λόγους ο κυματαριθμός ($\bar{\nu}$), ο οποίος αντιστοιχεί στο αντίστροφο του μήκους κύματος:

$$\bar{\nu} = \frac{1}{\lambda} = \frac{\nu}{c} = \frac{1}{2\pi c} \sqrt{\frac{\kappa}{\mu}}$$

Όπου (c) είναι η ταχύτητα του φωτός στο κενό, η σταθερά δύναμης (κ) εξαρτάται από την ισχύ του δεσμού A-B και είναι τόσο μεγαλύτερη όσο ισχυρότερος είναι ο χημικός δεσμός.

Για παράδειγμα, για τον απλό δεσμό C-C είναι $\approx 5 \times 10^{-5} \frac{\text{dynes}}{\text{cm}}$, ενώ για τον διπλό δεσμό C=C που είναι ισχυρότερος είναι $\approx 10 \times 10^{-5} \frac{\text{dynes}}{\text{cm}}$. Η ανηγμένη μάζα (μ) εξαρτάται από την μάζα των ατόμων ή των χαρακτηριστικών ομάδων που συνδέονται μεταξύ τους μέσω του δεσμού.

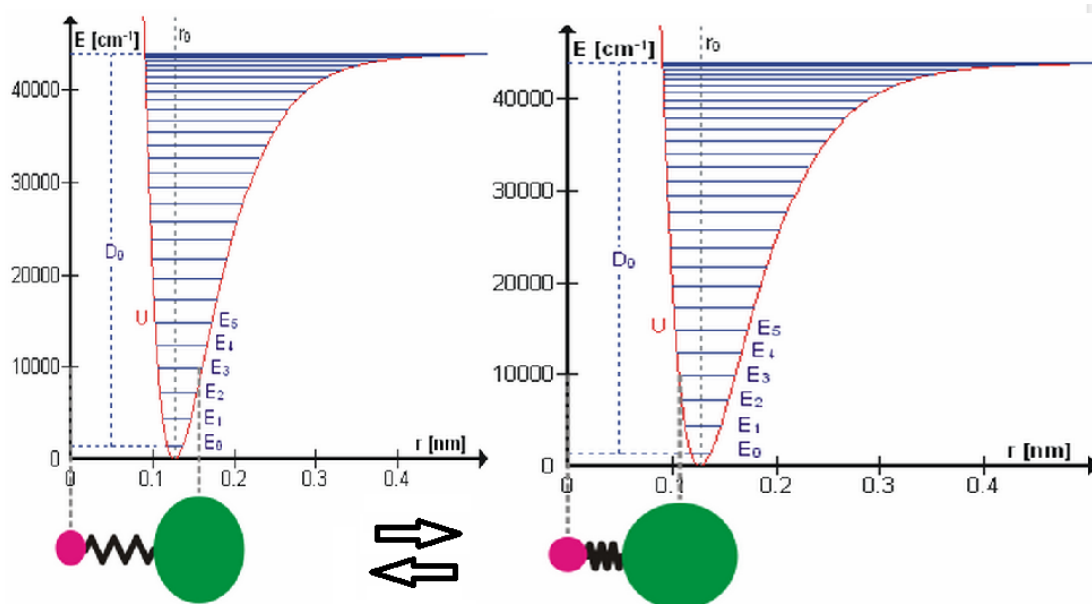
Η ενέργεια που απαιτείται για να προκαλέσει τη δόνηση δεσμού είναι:

$$E = h\nu = \frac{h}{2\pi} \sqrt{\frac{\kappa}{\mu}}$$

Όπου (h) σταθερά του Planck = 6.63×10^{-34}

Μια μοριακή δόνηση διεγείρεται όταν το μόριο απορροφά ένα κβαντικό ενέργειας (quantum of energy), E , που αντιστοιχεί στη συχνότητα της δόνησης, ν , σύμφωνα με τη σχέση $E = h\nu$ (όπου h είναι η σταθερά του Planck). Μια θεμελιώδης δόνηση διεγείρεται όταν ένα τέτοιο κβαντικό ενέργειας (quantum of energy) απορροφάται από το μόριο στην θεμελιώδη κατάσταση του. Όταν δύο κβάντα απορροφώνται η πρώτη κορυφή (overtone) αυξάνεται, όπως και οι υψηλότερες κορυφές (overtones).

Σχήμα 6. στάδια απορρόφησης ενέργειας ενός δεσμού.[11]



Συμπερασματικά, η συχνότητα δόνησης και επομένως η ενέργεια της ακτινοβολίας που απορροφάται, συνδέεται άμεσα τόσο με τη μάζα των ατόμων, όσο και την ισχύ του δεσμού. Αυτός είναι και ο λόγος που η ύπαρξη μιας ταινίας απορρόφησης σε συγκεκριμένη συχνότητα δόνησης (ή διαφορετικά σε συγκεκριμένο κυματαριθμό), αποτελεί σημαντικό και αδιαφιλονίκητο στοιχείο για την πιστοποίηση της ύπαρξης μιας συγκεκριμένης ομάδας ατόμων, τα οποία συνδέονται με καθορισμένο τρόπο (απλό, διπλό ή τριπλό δεσμό). Αυτό μπορεί να γίνει κατανοητό από τον επόμενο πίνακα στον οποίο παρουσιάζονται οι συχνότητες δόνησης κάποιων χαρακτηριστικών ομάδων. Είναι εμφανές ότι μεταβολές τόσο

στη μάζα όσο και στην ενέργεια του δεσμού συντελούν σε αντίστοιχες μεταβολές στη συχνότητα δόνησης.

Σχήμα 7, Συχνότητες δόνησης στο υπέρυθρο χαρακτηριστικών ομάδων.[9]

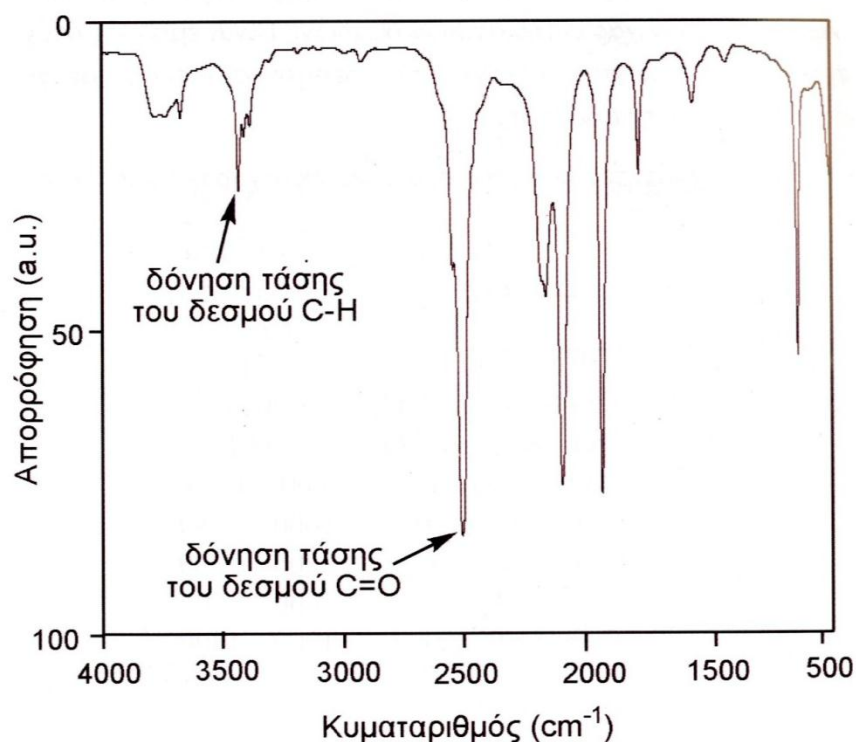
Δεσμός	Συχνότητα δόνησης [cm ⁻¹]
¹² C - ¹⁶ O	2143
¹³ C- ¹⁶ O (ισότοπο ¹³ C)	2096
¹² C- ¹⁸ O (ισότοπο ¹⁸ O)	2091
-C-C- (αλκάνια)	600 - 1500
-C=C- (αλκένια)	1600 - 1680
-C≡C- (αλκίνια)	2200 - 2260
C-Cl	600 - 800
-C-O- (αιθέρες)	1000 - 1300

Μόρια με περισσότερα από τρία άτομα είναι ακόμη πιο περίπλοκα και έχουν ακόμη περισσότερες δονήσεις τα οποία μερικές φορές ονομάζεται " δονητικές λειτουργίες (vibrational modes) ". Κάθε νέα δονητική λειτουργία είναι ουσιαστικά ένας διαφορετικός συνδυασμός των έξι που δείχνονται παραπάνω. Όσο περισσότερα άτομα στο μόριο τόσο περισσότεροι τρόποι που μπορούν να συνδυαστούν.

Η πολυπλοκότητα των υπέρυθρων φασμάτων έγκειται στο γεγονός ότι σε ένα μόριο μπορούν να υπάρχουν ταυτόχρονα όλοι οι δυνατοί τρόποι δόνησης με αποτέλεσμα την εμφάνιση πολλών κορυφών απορρόφησης στα φάσματα. Θεωρητικά ένα γραμμικό μόριο με n άτομα (η διάταξη των τριών ή περισσότερων ατόμων τοποθετούνται σε μια αναμενόμενη γωνία δεσμού 180 °) έχει τη δυνατότητα $3n - 5$ βασικών τρόπων δόνησης, ενώ σε ένα μη γραμμικό μόριο με n άτομα οι δυνατές δονήσεις είναι $3n - 6$. Έτσι για παράδειγμα, το μόριο του νερού που είναι μη γραμμικό εμφανίζει 3 ($3 \times 3 - 6 = 3$) δυνατούς τρόπους δονήσεις, ενώ το μόριο του HCl που είναι γραμμικό εμφανίζει 1 ($3 \times 2 - 5 = 1$) δόνηση, η οποία και προκαλεί την εμφάνιση απορρόφησης στο φάσμα στα 2886 cm⁻¹. Αντίθετα το μόριο του H₂ παρόλο που είναι και αυτό διατομικό δεν εμφανίζει απορρόφηση. Αυτό οφείλεται στο γεγονός ότι για την εμφάνιση απορρόφησης στο υπέρυθρο απαιτείται μεταβολή της διπολικής ροπής του μορίου (οι διπολικές ροπές οφείλονται σε μη ομοιόμορφες κατανομές των θετικών και αρνητικών φορτίων επί των διαφόρων ατόμων).

Στο επόμενο σχήμα παρουσιάζεται το φάσμα απορρόφησης στο υπέρυθρο της ακετόνης (CH₃COCH₃), είναι εμφανής ο μεγάλος αριθμός των ταινιών απορρόφησης, ο οποίος και καθιστά το φάσμα υπέρυθρου της συγκεκριμένης ένωσης αρκετά πολύπλοκο.

Σχήμα 8, φάσμα απορρόφησης υπέρυθρου για την ακετόνη (CH_3COCH_3). [9]



Ακριβώς όμως αυτή η πολυπλοκότητα καθιστά το φάσμα μοναδικό επιτρέποντας την ταυτοποίηση της ένωσης με πλήρη βεβαιότητα και τη διάκριση της από άλλες ενώσεις παρόμοιας σύνταξης. Δεν θα ήταν υπερβολή να αναφέρουμε ότι το υπέρυθρο φάσμα αποτελεί το "δαχτυλικό αποτύπωμα" του κάθε μορίου.

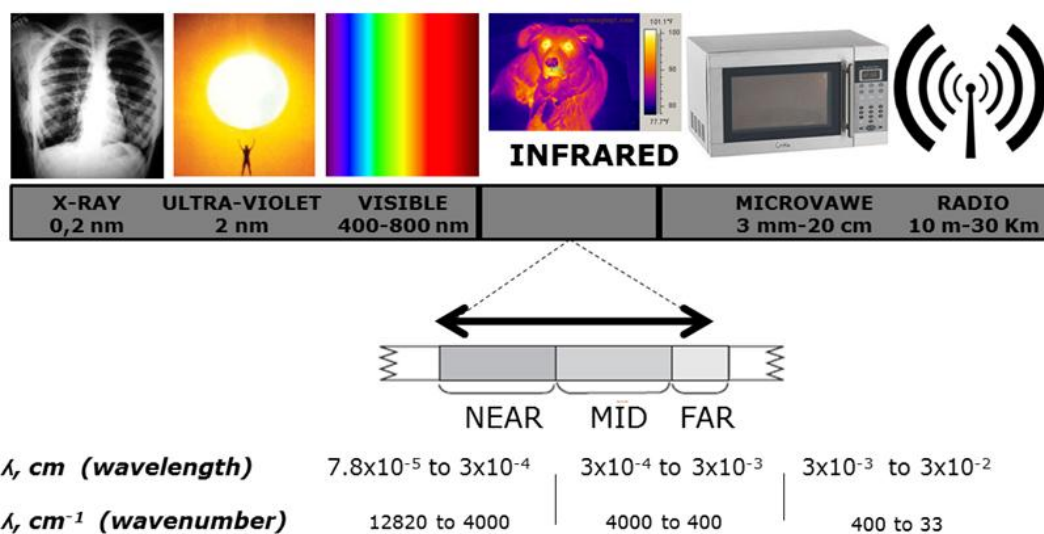
Πέρα από τον τομέα της ταυτοποίησης συγκεκριμένων μορίων και ενώσεων, σημαντική είναι η συμβολή της υπέρυθρης φασματοσκοπίας και στην κατανόηση των χημικών δράσεων που λαμβάνουν χώρα στην επιφάνεια των καταλυτών. Με την τεχνική αυτή μπορούμε να δούμε επιτόπου κατά τη διάρκεια μιας αντίδρασης, τα ενδιάμεσα χημικά είδη που σχηματίζονται πάνω στην επιφάνεια των καταλυτών και επομένως να αποφανθούμε για τα επιμέρους βήματα που επιτελούνται, για το μηχανισμό δηλαδή της αντίδρασης.

4. Η Φασματοσκοπία εγγύς υπερύθρου (NIR Spectroscopy) και η βενζίνη

4.1 Οι τρεις περιοχές υπερύθρων (Near-IR, Mid-IR, Far-IR)

Η περιοχή υπερύθρων αποτελείται από τρεις περιοχές, Near-IR, Mid-IR, Far-IR όπως φαίνεται παρακάτω.

Σχήμα 9 ,το ηλεκτρομαγνητικό φάσμα.[4]

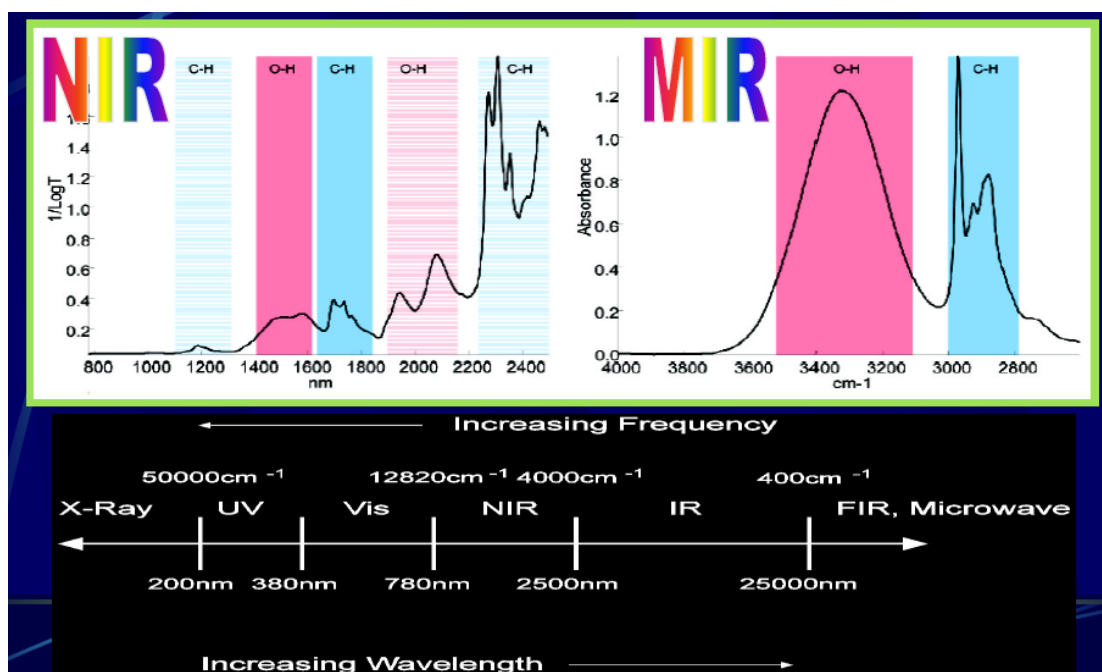


- Η περιοχή κάτω από 400 cm^{-1} ταξινομείται ως Far-IR , που χαρακτηρίζεται από δονήσεις χαμηλής συχνότητας. Χαρακτηρίζεται από χαμηλή ενέργεια, παραμορφωτικές δονήσεις και θεμελιώδεις λειτουργίες δόνησης τάσης των βαρέων ατόμων. Υπάρχει μόνο μία υπέρυθη δραστική θεμελιώδη δόνηση που εκτείνεται πέρα από 4000 cm^{-1} , και αυτή είναι η λειτουργία δόνησης τάσης του υδροφθορίου H-F (hydrogen fluoride) .

- Η Mid-IR περιοχή συνήθως ορίζεται ως η περιοχή συχνοτήτων των 4000 cm^{-1} έως 400 cm^{-1} . Το ανώτερο όριο είναι περισσότερο ή λιγότερο αυθαίρετα, και επιλέχθηκε αρχικά ως ένα πρακτικό όριο με βάση τα χαρακτηριστικά απόδοσης των παλαιών μέσων. Το κατώτερο όριο, σε πολλές περιπτώσεις, ορίζεται από ένα ειδικό οπτικό συστατικό, όπως ένα διαχωριστή δέσμης με βρωμιούχο κάλιο (KBr)υπόστρωμα, το οποίο έχει αποκοπή φυσικής μετάδοσης ακριβώς κάτω από 400 cm^{-1} .

- Οι κορυφές της NIR(NIR overtones) και μπάντες συνδυασμού (combination bands) προέρχονται από τις θεμελιώδεις μπάντες που παρατηρούνται στο MID-IR.

Σχήμα 10, σχετικές μπάντες- κορυφές (overtones) και μπάντες συνδυασμού (combination bands).[5]

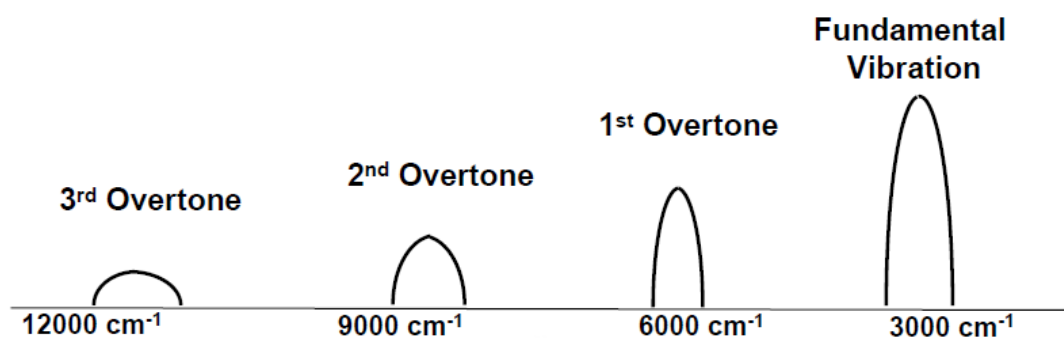


Οι κορυφές (overtones) είναι περίπου 2 , 3 και 4 φορές πολλαπλάσιοι της συχνότητας της θεμελιώδους δόνησης.

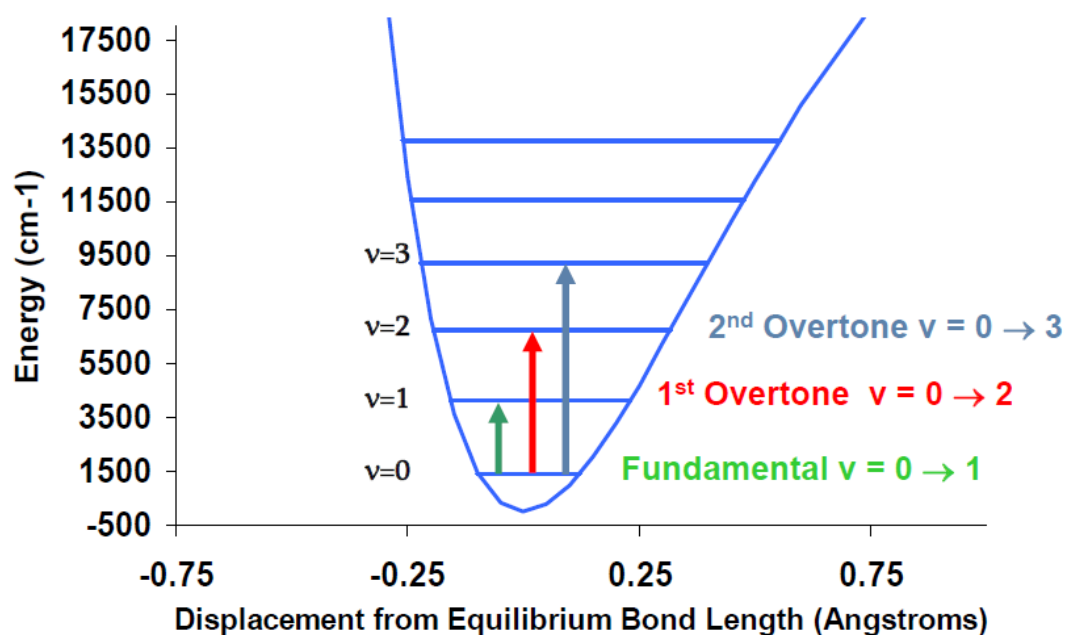
Η ένταση απορρόφησης μειώνεται με την αύξηση των κορυφών.

Η επικάλυψη μπάντας αυξάνεται με την αύξηση των κορυφών.

Σχήμα 11, οι κορυφές (overtones) σχετικά με την θεμελιώδη δόνηση.[6]



Σχήμα 12, στάδια απορρόφησης ενέργειας και εμφάνιση των κορυφών (overtones).[6]

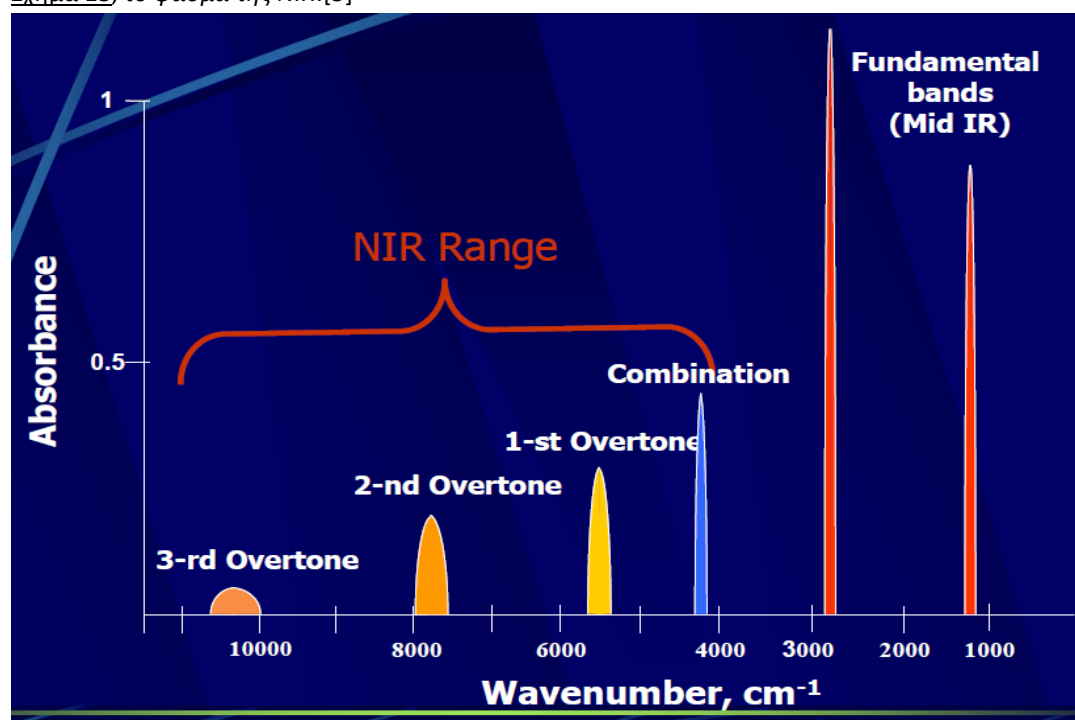


Υπάρχουν και άλλες δονήσεις στη NIR περιοχή εκτός από κορυφές των θεμελιωδών, οι οποίες καλούνται μπάντες συνδυασμού (combination bands).

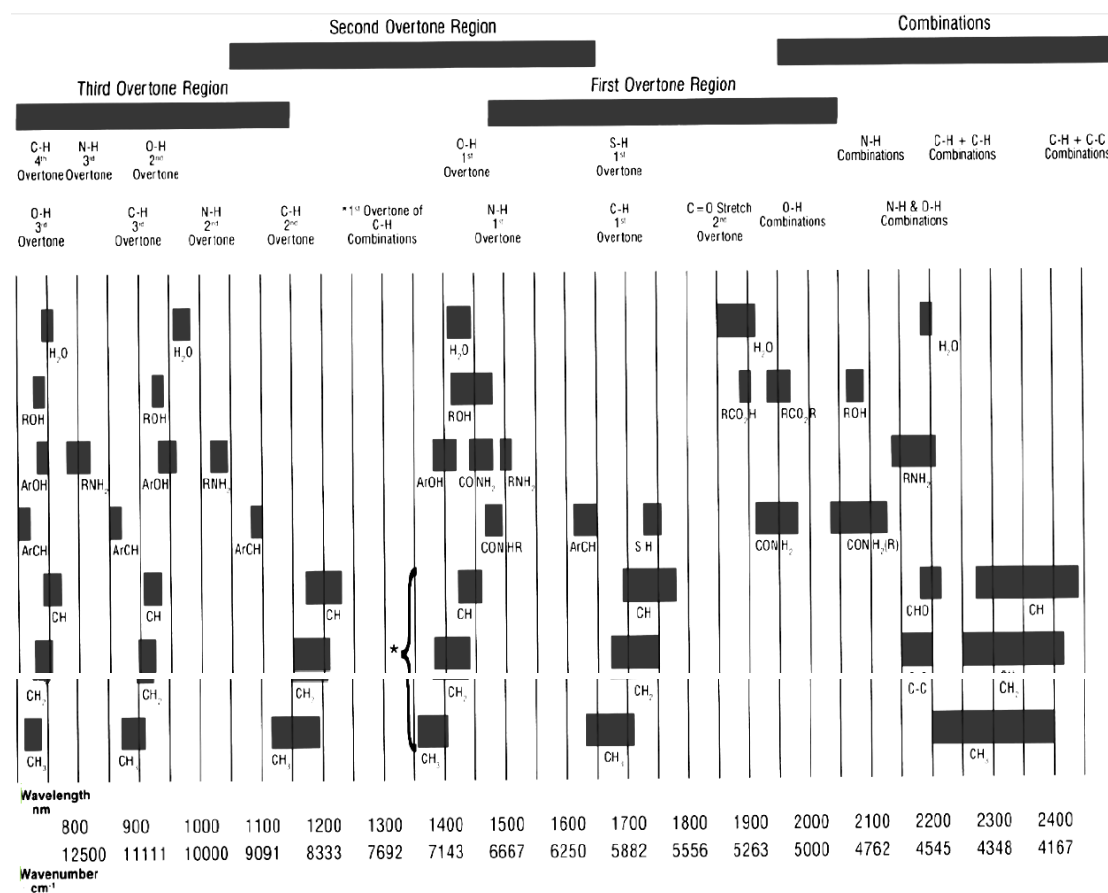
Μπάντες συνδυασμού (combination bands) είναι το άθροισμα των πολλών θεμελιωδών από διαφορετικές δονήσεις. Συνήθως βρίσκονται σε χαμηλότερες ενέργειες από αυτές των κορυφών.

Η μπάντα συνδυασμού εμφανίζονται κοντά στο άθροισμα των δύο ή τριών θεμελιωδών μπαντών.

Σχήμα 13, το φάσμα της NIR.[5]



Σχήμα 14,ο κυματαριθμός ή μήκος κύματος των κορυφών(overtones) και μπάντων συνδυασμού (combination bands) διάφορων στοιχείων που παρατηρούνται στο φάσμα της NIR.[5]



Το πρωτότυπο έργο για την NIR(Near-IR) είναι ότι ήταν μια επιπλέον εκτεταμένη περιοχή στα UV-Vis φασματόμετρα. Αυτό υποδεικνύει ότι η MID και NIR πρέπει να θεωρούνται στο ίδιο πεδίο. Για να ενισχυθεί αυτή η άποψη θα πρέπει να συνειδητοποιηθεί ότι περισσότερο από το ήμισυ του MID-IR φάσματος περιέχει κορυφές(overtones) και τις συχνότητες συνδυασμού των θεμελιωδών απορροφήσεων που συμβαίνουν κάτω από τα 2000 cm^{-1} .

4.2 Η σημασία της NIR

Η φασματοσκοπία NIR είναι μια δονητική μέθοδος φασματοσκοπίας (vibrational spectroscopic method) που ανήκει στο υπέρυθρο φάσμα φωτός που είναι πολύ κοντά στην ορατή περιοχή (από περίπου 750 nm έως 2500 nm ή 12820 cm^{-1} έως 4000 cm^{-1}), όπου οι περισσότερες οργανικές και επίσης μερικές ανόργανες ενώσεις έχουν εξαιρετική ανακλαστικότητα ή ιδιότητες διαπερατότητας, πράγμα που σημαίνει ότι είναι σε θέση να απορροφήσουν πολύ φως στην ακτινοβολία NIR.

Τα κυρίαρχα φασματικά χαρακτηριστικά της NIR περιλαμβάνουν: τις δονήσεις τάσης(stretching vibrations) C-H του μεθυλίου (CH₃), τις δονήσεις τάσης(stretching vibrations) C-H του μεθυλενίου(CH₂), δονήσεις τάσης C-H του αρωματικού (aromatic C-H stretching vibrations) και τις O-H δονήσεις τάσης (stretching vibrations). Ελάχιστα αλλά εξίσου σημαντικά φασματικά χαρακτηριστικά περιλαμβάνουν: μεθόξυ(O-CH₃) δόνηση τάσης C-H, δόνηση τάσης C-H συσχετιζόμενη με καρβονύλιο(C=O) (carbonyl associated C-H

stretching), N-H από τα πρωτογενή αμίδια και δευτεροταγή αμίδια αμφότερα, αλκύλια(C_nH_{2n+1}) και αρυλ ομάδες ενώσεων, N-H από τις πρωτοταγείς, δευτεροταγείς, τριτοταγείς αμίνες, και N-H από άλατα αμίνης.

Παραδοσιακά φασματοσκοπία της NIR χρησιμοποιείται συχνά για την ανάλυση των πολυμερών λιγνίνης (lignin polymers) (2270 nm , 4405.29 cm^{-1}), παραφινών και μακράς αλυσίδας πολυμερών αλκανίου(paraffins and long alkane chain polymers) (2310 nm , 4329 cm^{-1}), γλυκόζης με βάση πολυμερή όπως κυτταρίνη(glucose based polymers such as cellulose) (2336 nm , 4280.82 cm^{-1}), πολυμερών αμινοξέων ως πρωτεΐνες (amino acid polymers as proteins) (2180 nm , 4587.15 cm^{-1}), υδατανθράκων(carbohydrates) (2100 nm , 4761.9 cm^{-1}), και της υγρασίας ($1440\text{ and }1940\text{ nm}$, 6944.4 cm^{-1} and 5154.6 cm^{-1}).

Κατά την ανάλυση συνθετικών και φυσικών υλικών η φασματοσκοπία της NIR έχει δείξει πρωτοφανή βιομηχανική επιτυχία σε πολλαπλές εφαρμογές χρησιμοποιώντας δεκάδες χιλιάδες όργανα για σπόρους, ζωοτροφές, ψήσιμο των προϊόντων, αλεύρι, ποτά, τροφές, φαρμακευτικά προϊόντα, γαλακτοκομικά προϊόντα, υδρογονάνθρακες και πετροχημικά, χημικά προϊόντα, ραδιενεργά και επικίνδυνα υλικά, ιατρική απεικόνιση και διαγνωστικά.

Οι βασικές χρήσεις της φασματοσκοπίας NIR έχουν για τον έλεγχο της διαδικασίας, για την αξιολόγηση της ποιότητας, για την ταυτοποίηση των πρώτων υλών και την επεξεργασία υποπροϊόντων και για τη χημική ποσοτική ανάλυση των πολύπλοκων μειγμάτων.

Κατά τη μέτρηση μιγμάτων υδρογονανθράκων, όπως είναι τα καύσιμα ή μείγματα διαλυτών, η NIR δεν μετράει άμεσα κλάσεις υδρογονανθράκων όπως ολεφίνες ή ναφθένια, αλλά μετρά απορροφήσεις λειτουργικής ομάδας, όπως μεθύλιο, μεθυλένιο, μεθίνης, και δονήσεις τάσης αρωματικών και δονήσεις παραμόρφωσης. Οι αναλογίες των διαφόρων αυτών απορροφήσεων όταν συσχετίζονται χρησιμοποιώντας διάφορες πολυμεταβλητές τεχνικές βαθμονόμησης σε γνωστούς φυσικές ή συνθετικές παραμέτρους για ένα σετ μάθησης παρέχουν μια εκτίμηση συσχέτισης της σύνθεσης κατηγορίας υδρογονανθράκων από διάφορα άγνωστα πολύπλοκα μίγματα.

Τα σημαντικά μόρια για τις μετρήσεις της NIR είναι πιο συχνά αυτά του νερού (όπως O-H δόνηση τάσης(O-H stretch)) και συνδεδεμένες διαλυμένες ουσίες, πρωτεΐνες, υδατάνθρακες, λίπη και έλαια, και οι τάξεις υδρογονανθράκων.

Η NIR συνεχίζει να παρέχει μια πολύτιμη τεχνική μέτρησης, ισχύει τόσο για φυσικά και συνθετικά υλικά, , για χρήση σε συνεχή παρακολούθηση ή σε πραγματικό χρόνο παρακολούθησης της διαδικασίας .

4.3 Η βενζίνη

Η βενζίνη είναι ένα διαφανές παράγωγο του πετρελαίου που χρησιμοποιείται κυρίως ως καύσιμο σε μηχανές εσωτερικής καύσης. Αποτελείται κυρίως από οργανικές ενώσεις που λαμβάνονται από την κλασματική απόσταξη του πετρελαίου, ενισχυμένη με μια ποικιλία προσθέτων.

Η βενζίνη παράγεται σε διυλιστήρια πετρελαίου. Το υλικό διαχωρίζεται από το αργό πετρέλαιο μέσω της απόσταξης, που ονομάζεται παρθένο ή βενζίνη απευθείας απόσταξης (virgin or straight-run gasoline), δεν πληρεί τις προδιαγραφές για σύγχρονους κινητήρες (ιδιαίτερα του βαθμού οκτανίων), αλλά αποτελεί τμήμα του μίγματος. Το μεγαλύτερο

μέρος μιας τυπικής βενζίνης συνίσταται από υδρογονάνθρακες με 4 έως και 12 άτομα άνθρακα ανά μόριο (που συνήθως αναφέρονται ως C4-C12).

Όπου «Υδρογονάνθρακες» ονομάζονται οι οργανικές ενώσεις που περιέχουν μόνο υδρογόνο (H) και άνθρακα (C) και έχουν γενικό χημικό τύπο C_xH_y .

Τα διάφορα ρεύματα διυλιστηρίου αναμειγνύονται για να κάνουν την βενζίνη να έχει διαφορετικά χαρακτηριστικά. Μερικά σημαντικά ρεύματα είναι:

- **Βενζίνη απευθείας απόσταξης(straight-run gasoline)** : αποστάζεται απευθείας από το αργό πετρέλαιο. Είναι χαμηλού βαθμού οκτανίων και απαιτεί πρόσθετα μολύβδου.

Έχει χαμηλή περιεκτικότητα σε αρωματικές ενώσεις (ανάλογα με την τάξη του αργού πετρελαίου), περιέχει κάποια κυκλοαλκάνια (ναφθένια) (cycloalkanes or naphthenes) και δεν περιέχει ολεφίνες (olefins) . Περίπου 0-20% της βενζίνης προέρχεται από αυτό το υλικό, εν μέρει επειδή η παροχή του κλάσματος αυτού είναι ανεπαρκής σε RON (Research Octane Number) . Δηλαδή αριθμού οκτανίων είναι πολύ χαμηλό.

Κυκλοαλκάνια ή Ναφθένια είναι οι τύποι των ενώσεων υδρογονάνθρακα που έχουν έναν ή περισσότερους δακτυλίους ατόμων άνθρακα στη χημική δομή των μορίων τους. Ο γενικός χημικός τύπος για τα κυκλοαλκάνια είναι $C_nH_{2(n+1-g)}$ όπου n = αριθμός των ατόμων C και g = ο αριθμός των δακτυλίων στο μόριο.

Η ολεφίνη είναι μία ακόρεστη χημική ένωση που περιέχει τουλάχιστον ένα άνθρακα-άνθρακα διπλό δεσμό. Ο γενικός τύπος είναι C_nH_{2n} .

- **Αναμορφωμένο ρεύμα (reformate):** παράγεται σε έναν καταλυτικό αναμορφωτή, έχει υψηλό βαθμό οκτανίων με υψηλή περιεκτικότητα σε αρωματικούς υδρογονάνθρακες και πολύ χαμηλές ολεφίνες (αλκένια). Το μεγαλύτερο μέρος του βενζολίου (benzene), τολουολίου (toluene), και ξυλολίου (xylene) είναι πιο πολύτιμο ως χημικές πρώτες ύλες και έτσι αφαιρούνται σε κάποιο βαθμό. Το BTX ακρωνύμιο αναφέρεται σε μίγματα από βενζόλιο, τολουόλιο και τα τρία ισομερή ξυλολίου.

- **Καταλυτική πυρολυμένη βενζίνη (cat cracked gasoline or cat cracked naphtha):**

παράγεται από μια μονάδα καταλυτικής πυρόλυσης, με έναν μέτριο βαθμό οκτανίων. Έχει υψηλή περιεκτικότητα σε ολεφίνες (αλκένιο) και μέτρια περιεκτικότητα σε αρωματικά.

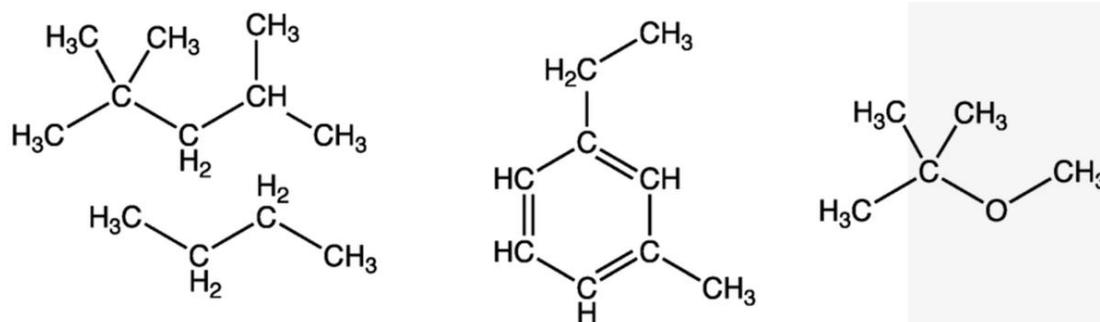
- **Ρεύμα υδρογονοπυρόλυσης (hydrocrackate):** (βαρύ, μεσαίο και ελαφρό) παράγεται από συσκευή υδρογονοπυρόλυσης, με μεσαίο έως χαμηλό βαθμό οκτανίων και μέτρια αρωματικά επίπεδα.

- **Ρεύμα αλκυλίωσης (alkylate):** παράγεται σε μία μονάδα αλκυλίωσης, που περιλαμβάνει την προσθήκη ισοβουτανίου (isobutene) σε αλκένια δίνοντας διακλαδισμένες αλυσίδες αλλά χαμηλά αρωματικά.

- **Ρεύμα ισομερισμού (isomerate):** λαμβάνεται με ισομερισμό Βενζίνης χαμηλών οκτανίων της απευθείας απόσταξης (straight-run gasoline) σε ισο-παραφίνες (iso-paraffins), όπως ισοοκτάνιο.

Οι όροι παραπάνω είναι η ορολογία που χρησιμοποιείται στη βιομηχανία πετρελαίου, αλλά αυτή η ορολογία μπορεί να ποικίλει.

Σχήμα 15, μερικά από τα κύρια συστατικά της βενζίνης: ισοοκτάνιο(isooctane) $((CH_3)_3CCH_2CH(CH_3)_2)$, βουτάνιο(butane) $(CH_3CH_2CH_2CH_3)$, 3-αιθυλοτολουόλιο(3-ethyltoluene) $(CH_3CH_2(C_6H_5)CH_3)$, και ο ενισχυτής οκτανίων MTBE (Methyl Tertiary Butyl Ether) (μεθύλιο τριτογενή βουτύλιο αιθέρας) $((CH_3)_3COCH_3)$. [22]



Μία τυπική βενζίνη είναι κατά κύριο λόγο ένα μίγμα από παραφίνες (αλκάνια), κυκλοαλκάνια (ναφθένια) και ολεφίνες (αλκένια). Η πραγματική αναλογία εξαρτάται από:

- Το διυλιστήριο πετρελαίου που κάνει την βενζίνη, καθώς όλα τα διυλιστήρια δεν έχουν το ίδιο σύνολο των μονάδων επεξεργασίας.
- Τη τροφοδοσία του αργού πετρελαίου που χρησιμοποιείται από το διυλιστήριο.
- Το βαθμό της βενζίνης, ιδίως, ο βαθμός οκτανίων.

Σήμερα, πολλές χώρες θέτουν όρια για τα αρωματικά βενζίνης, ειδικότερα για το βενζόλιο, και το περιεχόμενο ολεφινών (αλκένιο). Οι κανονισμοί αυτοί οδήγησαν στην αυξανόμενη προτίμηση για συστατικά καθαρής παραφίνης (αλκάνιο) υψηλού αριθμού οκτανίων, όπως το αλκυλικό (alkylate), και αναγκάζουν τα διυλιστήρια να προσθέσουν μονάδες επεξεργασίας για τη μείωση περιεκτικότητας βενζολίου.

Η βενζίνη μπορεί επίσης να περιέχει άλλες οργανικές ενώσεις, όπως οι οργανικοί αιθέρες (organic ethers) (σκόπιμα προστίθεται), συν μικρά επίπεδα προσμείξεων, ειδικότερα ενώσεις οργανοθείου (organosulfur) αλλά συνήθως αυτές αφαιρούνται στο διυλιστήριο.

4.4 Αρωματικοί υδρογονάνθρακες

Ένας αρωματικός υδρογονάνθρακας (aromatic hydrocarbon or arene or aryl hydrocarbon) είναι ένας υδρογονάνθρακας με εναλλασσόμενους διπλούς και μονούς δεσμούς μεταξύ των ατόμων άνθρακα που σχηματίζουν δακτυλίους .

Η διαμόρφωση των έξι ατόμων άνθρακα σε αρωματικές ενώσεις που είναι γνωστή ως ένα δακτύλιο βενζολίου, είναι ο απλούστερος αρωματικός υδρογονάνθρακας .

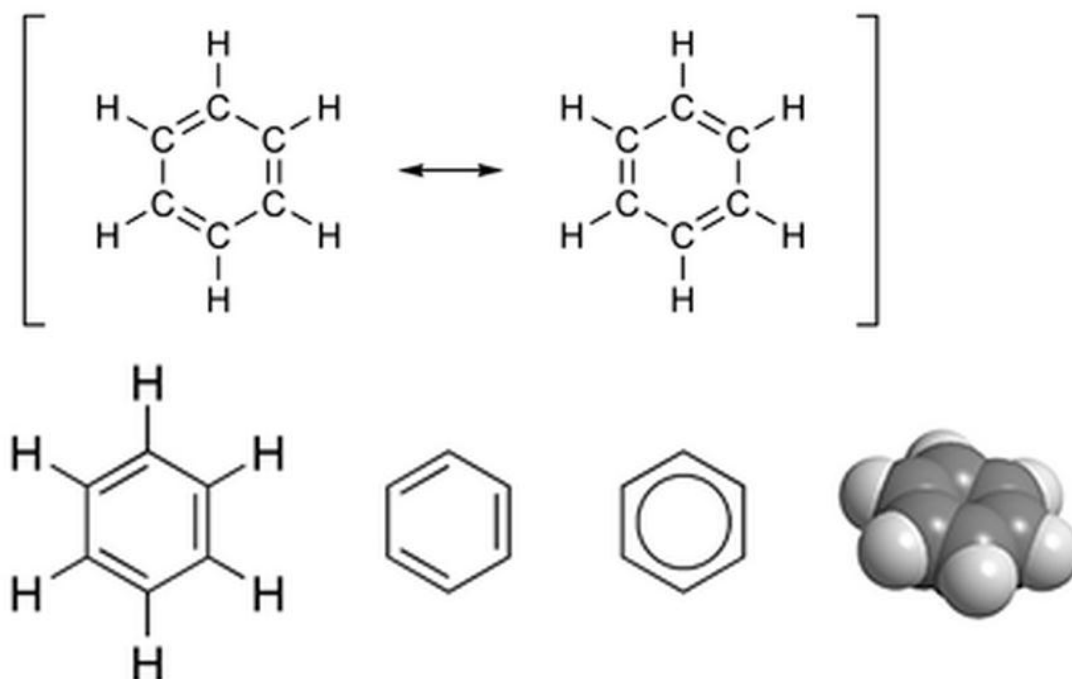
Οι αρωματικοί υδρογονάνθρακες μπορούν να είναι μονοκυκλικοί (MAH(Monocyclic Aromatic Hydrocarbons)) ή πολυκυκλικοί (PAH (Poly-aromatic hydrocarbons)).

Μερικές ενώσεις που δεν βασίζονται στο βενζόλιο ονομάζονται ετεροαρένια (heteroarenes) όμως ονομάζονται επίσης ως αρωματικές ενώσεις. Σε αυτές τις ενώσεις, τουλάχιστον ένα άτομο άνθρακα αντικαθίσταται από ένα από τα ετεροάτομα (ένα ετεροάτομο (heteroatom) είναι οποιοδήποτε άτομο που δεν είναι άνθρακας ή υδρογόνο.) Οξυγόνο, άζωτο, ή θείο. Παραδείγματα ενώσεων που δεν βασίζονται στο βενζόλιο με αρωματικές ιδιότητες είναι το φουράνιο(furan), μια ετεροκυκλική ένωση με πέντε μέλη στον δακτύλιο που περιλαμβάνει ένα άτομο οξυγόνου, και η πυριδίνη (pyridine), μια ετεροκυκλική ένωση με έξι μέλη στο δακτύλιο που περιέχει ένα άτομο αζώτου.

4.5 Δακτύλιο βενζολίου

Το βενζόλιο, C_6H_6 , είναι ο απλούστερος αρωματικός υδρογονάνθρακας, και ήταν ο πρώτος που αναγνωρίστηκε. Η φύση των δεσμών του αναγνωρίστηκε για πρώτη φορά από τον Αύγουστ Κέκουλε (August Kekulé) τον 19ο αιώνα. Κάθε άτομο άνθρακα στον εξαγωνικό κύκλο έχει τέσσερα ηλεκτρόνια για να μοιραστούν. Ένα ηλεκτρόνιο πηγαίνει στο άτομο του υδρογόνου, και δύο ηλεκτρόνια με τους δύο γειτονικούς άνθρακες. Αυτό αφήνει ένα ηλεκτρόνιο να μοιραστεί με ένα από τα δύο γειτονικά άτομα άνθρακα, το οποίο είναι ο λόγος που το μόριο βενζολίου σχεδιάζεται με εναλλασσόμενους διπλούς και μονούς δεσμούς γύρω από το εξάγωνο.

Σχήμα 16, διαφορετικοί τρόποι σχεδιασμού του βενζολίου.[24]



Ως πρόσθετο βενζίνης, το βενζόλιο αυξάνει τον βαθμό οκτανίων και μειώνει τα χτυπήματα. Όμως.....

Η ανθρώπινη έκθεση σε βενζόλιο είναι ένα παγκόσμιο πρόβλημα υγείας.

Το βενζόλιο στοχεύει το συκώτι, τον νεφρό, τον πνεύμονα, την καρδιά και τον εγκέφαλο και μπορεί να προκαλέσει διάσπαση της έλικας του DNA, χρωμοσωμική βλάβη, κλπ.

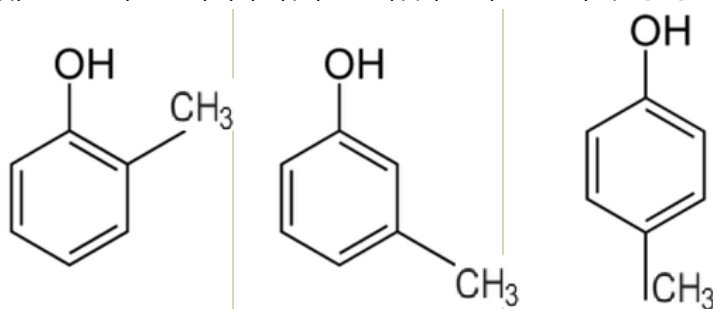
Το βενζόλιο προκαλεί καρκίνο σε ζώα συμπεριλαμβανομένων των ανθρώπων.
Το βενζόλιο έχει αποδειχθεί ότι προκαλεί καρκίνο και στα δύο φύλα των πολλαπλών ειδών των πειραματόζων, τα οποία εκτέθηκαν μέσω διαφόρων τρόπων.
Το βενζόλιο αυξάνει τον κίνδυνο του καρκίνου και άλλων ασθενειών.

4.6 Το βενζόλιο και τα παράγωγα βενζολίου

Τα παράγωγα βενζολίου έχουν από ένα έως έξι υποκατάστατα που συνδέονται στον κεντρικό πυρήνα βενζολίου. Παραδείγματα ενώσεων βενζολίου με ένα μόνο υποκατάστατο είναι η φαινόλη (phenol), η οποία έχει μία ομάδα υδροξυλίου (hydroxyl group), και το τολουόλιο (toluene) με μία ομάδα μεθυλίου (methyl group). Όταν υπάρχουν περισσότερα από ένα υποκατάστατα επί του δακτυλίου η χωρική σχέση τους γίνεται σημαντική για τα πρότυπα υποκατάστασης αρωματικού υδρογονάνθρακα και σχεδιάζονται σε διάταξη ορθό (ortho), μετά (meta) και παρά (para).

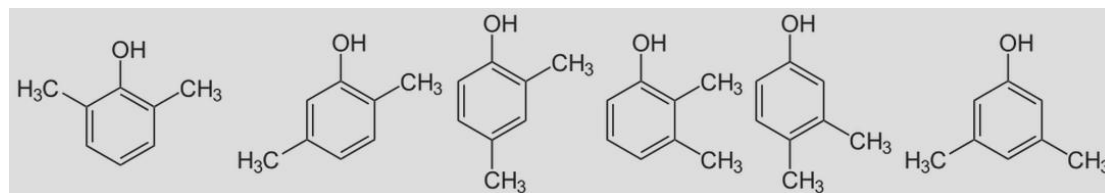
Για παράδειγμα, υπάρχουν τρία ισομερή για τη κρεσόλη (cresol), επειδή η ομάδα μεθυλίου και η ομάδα υδροξυλίου μπορεί να τοποθετηθεί η μία δίπλα στην άλλη (ορθό), ή με μία θέση κενή ανάμεσα τους (μετά), ή δύο θέσεις κενές (παρά).

Σχήμα 17, τα τρία ισομερή της κρεσόλης (ορθό, μετά, παρά) . [25]

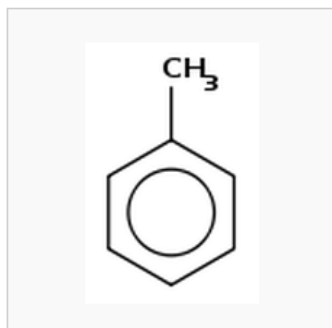


Η ξυλενόλη (Xylenol) έχει άλλες δύο ομάδες μεθυλίου εκτός από την ομάδα υδροξυλίου, και για αυτήν την δομή, υπάρχουν έξι ισομερή.

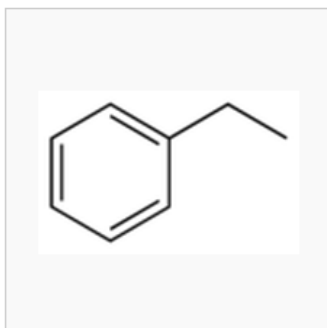
Σχήμα 18, τα έξι ισομερή της ξυλενόλης. [26]



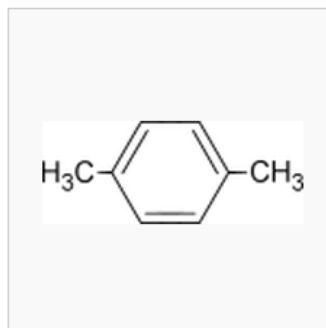
Σχήμα 19, αντιπροσωπευτικές ενώσεις αρωματικού υδρογονάνθρακα. Το τολουόλιο (Toluene), το αιθυλοβενζόλιο (Ethylbenzene), *p*- ξυλόλη (*p*- Xylene), *m*- ξυλόλη (*m*- Xylene), το μεσιτυλένιο (Mesitylene), το ντουρένιο (Durene), το 2-φενυλεξάνιο (2-Phenylhexane), το διφαινύλιο (Biphenyl), η φαινόλη (Phenol), η ανιλίνη (Aniline), το νιτροβενζόλιο (Nitrobenzene), το βενζοϊκό οξύ (Benzoic acid), η ασπιρίνη (Aspirin), η Παρακεταμόλη (Paracetamol), το πικρικό οξύ (Picric acid).[23]



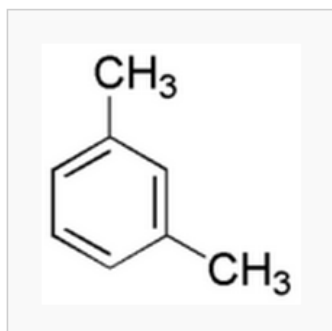
Toluene



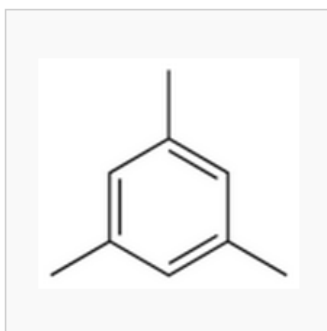
Ethylbenzene



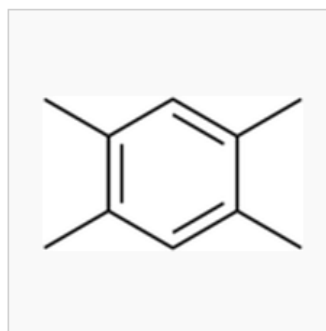
p-Xylene



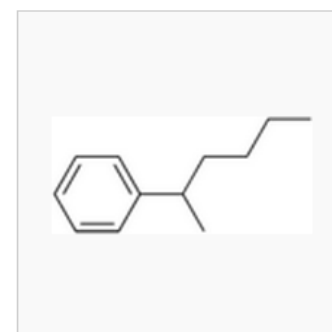
m-Xylene



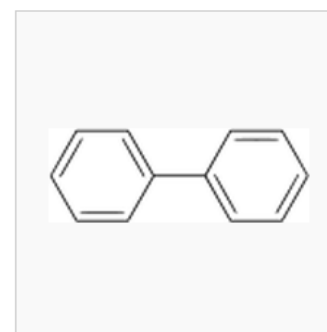
Mesitylene



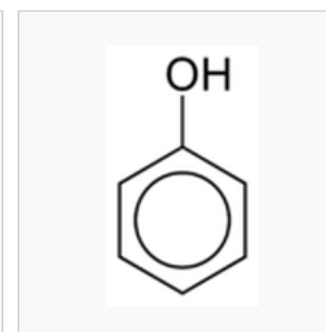
Durene



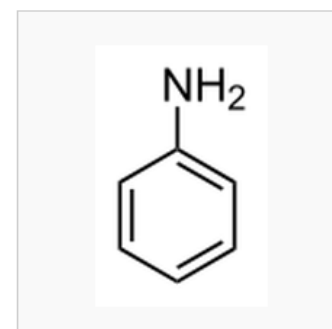
2-Phenylhexane



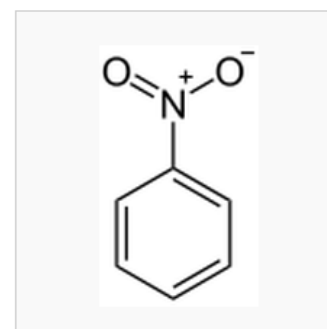
Biphenyl



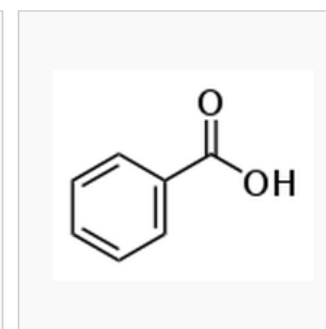
Phenol



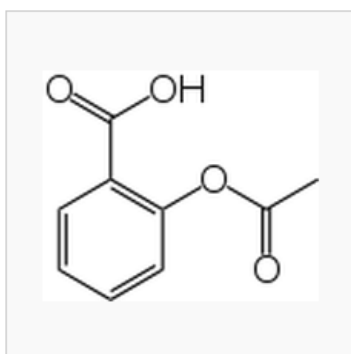
Aniline



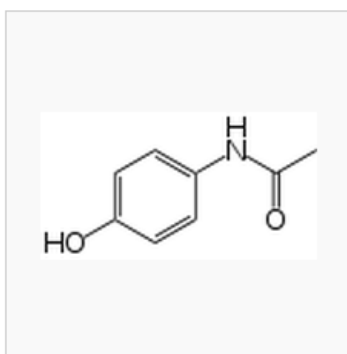
Nitrobenzene



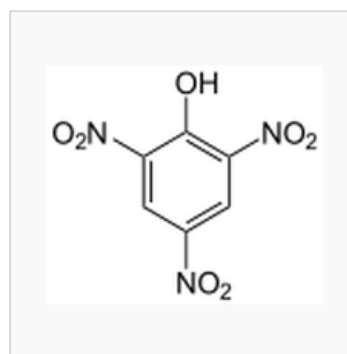
Benzoic acid



Aspirin



Paracetamol

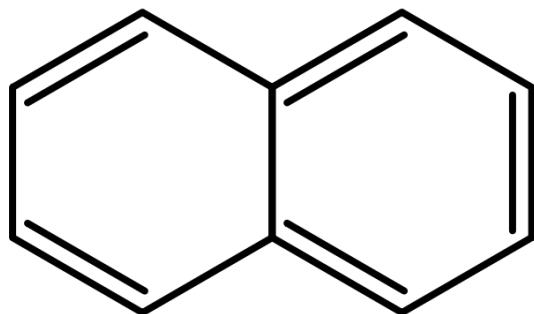


Picric acid

4.7 Πολυαρωματικοί υδρογονάνθρακες

Οι πολυαρωματικοί Υδρογονάνθρακες (Poly-aromatic hydrocarbons (PAHs) or aspolycyclic aromatic hydrocarbons or polynuclear aromatic hydrocarbons) Είναι ισχυροί ατμοσφαιρικοί ρύποι που αποτελούνται από ενωμένους αρωματικούς δακτυλίους και δεν περιέχουν ετεροάτομα ούτε φέρουν υποκαταστάτες. Η ναφθαλίνη (Naphthalene) ($C_{10}H_8$) είναι το απλούστερο παράδειγμα ενός PAH.

Σχήμα 20, η ναφθαλίνη. [23]



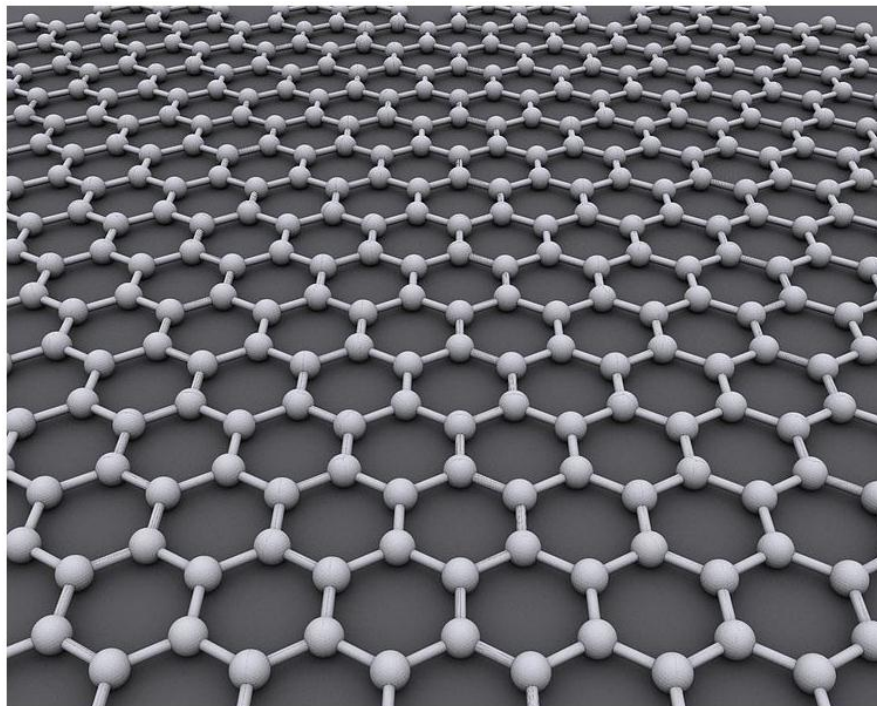
Οι PAH δημιουργούνται στο πετρέλαιο, στον άνθρακα, στη πίσσα, και παράγονται ως υποπροϊόντα της καύσης του καυσίμου (είτε ορυκτά καύσιμα ή βιομάζα).

Ως ρύποι προκαλούν ανησυχία διότι ορισμένες ενώσεις έχουν χαρακτηριστεί ως καρκινογόνες, μεταλλαξιογόνες και τερατογόνες. Οι PAH βρίσκονται επίσης σε μαγειρευμένα τρόφιμα. Μελέτες έχουν δείξει ότι τα υψηλά επίπεδα των πολυκυκλικών αρωματικών υδρογονανθράκων βρίσκονται, σε κρέας μαγειρευμένο σε υψηλές θερμοκρασίες, όπως το ψήσιμο στη σχάρα ή μπάρμπεκιου, και στο καπνιστό ψάρι.

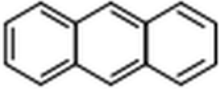
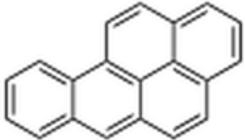
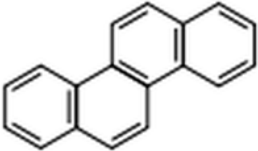
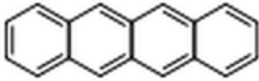
Οι PAH βρίσκονται επίσης σε διαστρικά μέσα, σε κομήτες και μετεωρίτες και είναι ένα υποψήφιο μόριο για να λειτουργήσει ως βάση για τις πρώτες μορφές ζωής.

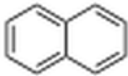
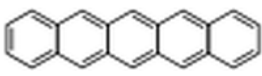
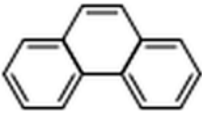
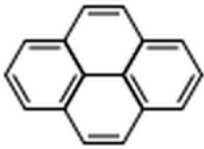
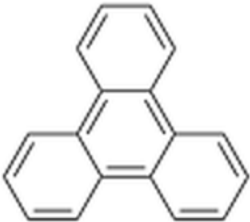
Σε γραφένιο, το μοτίβο PAH επεκτείνεται σε μεγάλα φύλλα 2D. Το γραφένιο θεωρείται το πιο λεπτό, ισχυρό και ηλεκτρικά αγωγίμο υλικό στον κόσμο.

Σχήμα 21, το γραφένιο. Το μοτίβο ΡΑΗ επεκτείνεται σε μεγάλα φύλλα 2D.[27]



Σχήμα 22, ενώσεις πολυαρωματικών υδρογονανθράκων. Το ανθρακένιο (Anthracene), το βενζο[α]πυρένιο (Benzo[a]pyrene), το χρυσένιο (chrysene), κορωνένιο (Coronene), Corannulene, Tetracene, η ναφθαλίνη (Naphthalene), το πεντακένιο (Pentacene), το φαινανθρένιο (Phenanthrene), το πυρένιο (Pyrene), το τριφαινυλένιο (Triphenylene), το οβαλένιο (Ovalene).[23]

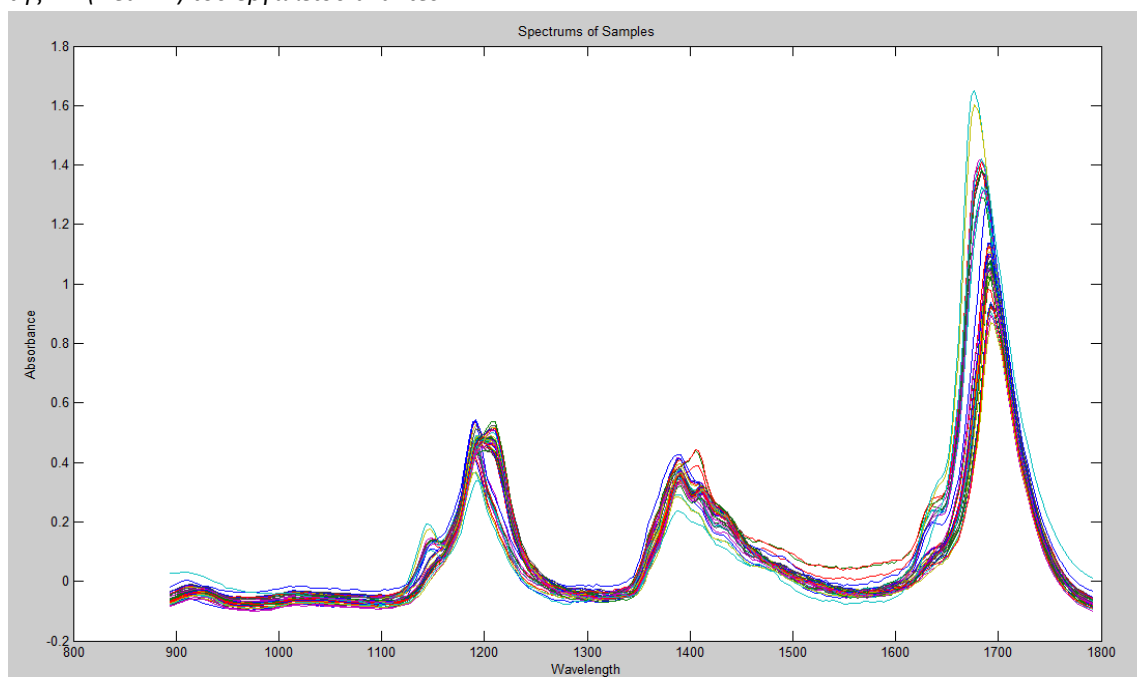
Chemical compound		Chemical compound	
Anthracene		Benzo[a]pyrene	
Chrysene		Coronene	
Corannulene		Tetracene	

Naphthalene		Pentacene	
Phenanthrene		Pyrene	
Triphenylene		Ovalene	

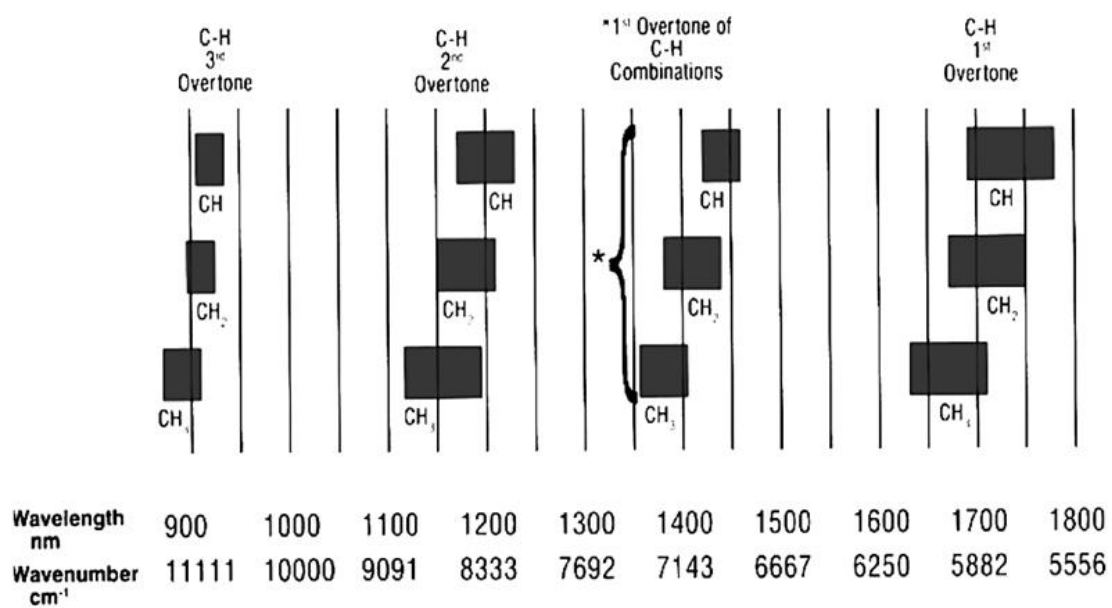
Τα φάσματα διαφόρων ειδών βενζίνης που έχουμε πάρει στο εργαστήριο με τη βοήθεια της NIR(Near-IR) του εργαλείου avantes φαίνονται στο Σχήμα 23.

Τα overtones που φαίνονται στα φάσματα αφείλονται στα διάφορα είδη δεσμών C-H που υπάρχουν στα διάφορα είδη βενζίνης, όπως το C-H του μεθυλίου (CH_3) (βλέπε σχήμα 19, το τολουόλιο (Toluene), p- ξυλόλη (p- Xylene), m- ξυλόλη (m- Xylene)), C-H του μεθυλενίου(CH_2) και C-H του αρωματικού. Όπως φαίνεται στα Σχήματα 23, 24 .

Σχήμα 23 , Τα φάσματα διάφορων ειδών βενζίνης που έχουμε πάρει στο εργαστήριο με τη βοήθεια της NIR(Near-IR) του εργαλείου avantes.



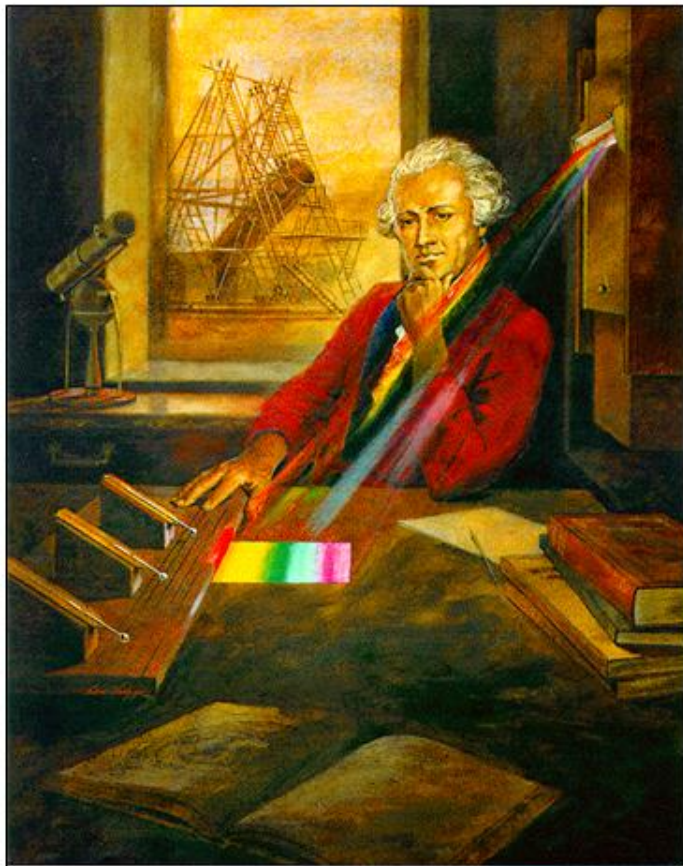
Σχήμα 24, το αντίστοιχο σχήμα των δονήσεων των δεσμών C-H (overtones) για Τα φάσματα διαφόρων ειδών βενζίνης που έχουμε πάρει στο εργαστήριο.[5]



4.8 Η ιστορία της NIR

Η ιστορία της near infrared (NIR) ξεκινά το 1800 με Frederick William Herschel,

Σχήμα 25, *Frederick William Herschel*. [4]



Προσπαθούσε με φίλτρα να παρατηρήσει κηλίδες από τον ήλιο και όταν χρησιμοποίησε ένα κόκκινο φίλτρο παρατήρησε ότι έχει παραχθεί μεγάλη ποσότητα θερμότητας, η οποία ήταν σε υψηλότερη θερμοκρασία από το ορατό φάσμα. Μετά από περαιτέρω μελέτη, κατέληξε στο συμπέρασμα ότι πρέπει να υπάρχει μια αόρατη μορφή του φωτός πέρα από το ορατό φάσμα.

Η φασματοσκοπία NIR ήταν παραμελημένη από φασματόσκοπος που για μεγάλο χρονικό διάστημα, δεν μπόρεσαν να βρουν οποιαδήποτε πρόσθετες ελκυστικές πληροφορίες σε αυτή τη φασματική περιοχή που είχε καταληφθεί από ευρεία, επάλληλα και αδύναμες ζώνες απορρόφησης.

Η πρώτη βιομηχανική εφαρμογή ξεκίνησε στη δεκαετία του 1950. Κατά τις πρώτες εφαρμογές, η NIRS χρησιμοποιήθηκε μόνο ως επιπρόσθετη μονάδα σε άλλες οπτικές συσκευές που χρησιμοποιούσαν άλλα μήκη κύματος όπως τα υπεριώδη (UV), τα ορατά (VIS), ή μέσο υπέρυθρα (MIR) φασματόμετρα. Στη δεκαετία του 1980, ως μια ενιαία μονάδα, ένα αυτόνομο σύστημα NIRS έγινε διαθέσιμο, αλλά η εφαρμογή των NIRS επικεντρώθηκε περισσότερο στην χημική ανάλυση. Με την εισαγωγή του οπτικών ινών στα μέσα της δεκαετίας του 1980 και τις εξελίξεις του Μονοχρωμάτορα-Ανιχνευτή (monochromator

detector) στις αρχές της δεκαετίας του 1990, η NIRS έγινε πιο ισχυρό εργαλείο για την επιστημονική έρευνα.

4.9 Πλεονεκτήματα της NIR

Αυτή η οπτική μέθοδος μπορεί να χρησιμοποιηθεί σε διάφορους τομείς της επιστήμης, συμπεριλαμβανομένης της φυσικής, της φυσιολογίας ή της ιατρικής. Μόνο κατά τις τελευταίες δεκαετίες η NIRS άρχισε να χρησιμοποιείται ως ένα ιατρικό εργαλείο παρακολούθησης για τους ασθενείς. Η φασματοσκοπία NIR είναι μια δονητική μέθοδος φασματοσκοπίας (vibrational spectroscopic method), ανήκει στο υπέρυθρο φάσμα φωτός που είναι πολύ κοντά στην ορατή περιοχή (από περίπου 750 nm έως 2500 nm ή 12820 cm^{-1} έως 4000 cm^{-1}), όπου οι περισσότερες οργανικές και επίσης μερικές ανόργανες ενώσεις έχουν εξαιρετική ανακλαστικότητα ή ιδιότητες διαπερατότητας, πράγμα που σημαίνει ότι είναι σε θέση να απορροφήσουν πολύ φως στην ακτινοβολία NIR. Τυπικές εφαρμογές περιλαμβάνουν φαρμακευτικές ιατρικές διαγνώσεις (συμπεριλαμβανομένης της ζάχαρης στο αίμα και παλμικής οξυμετρίας), έλεγχο ποιότητας τροφίμων και αγροχημικών, έρευνα καύσης (combustion research), καθώς και έρευνα σε λειτουργική νευροαπεικόνιση, αθλητιατρική και επιστήμη, ελίτ αθλητική κατάρτιση, εργονομία, αποκατάσταση, νεογνική έρευνα, επικοινωνία εγκεφάλου- υπολογιστή, ουρολογία (συστολή της ουροδόχου κύστης (bladder contraction)) και τη νευρολογία (νευραγγειακά σύζευξη (neurovascular coupling)).

Ένα πλεονέκτημα είναι ότι NIR μπορεί τυπικά να διεισδύσει πολύ μακρύτερα μέσα σε ένα δείγμα από την mid infrared ακτινοβολία. Εγγύς υπέρυθρη φασματοσκοπία (Near-infrared spectroscopy) ως εκ τούτου, δεν είναι μια ιδιαίτερα ευαίσθητη τεχνική, αλλά μπορεί να είναι πολύ χρήσιμη για διερεύνηση χύμα υλικού με λίγη ή καθόλου προετοιμασία του δείγματος κάτι που σημαίνει μείωση του χρόνου ανάλυσης, εξάλλου επειδή η NIR είναι μη καταστρεπτική μέθοδος τα απόβλητα και τα αντιδραστήρια ελαχιστοποιούνται.

Σχήμα 26, εξέταση μιας ουσίας κατά τη διάρκεια μιας διαδικασίας.[4]



Οι μοριακές κορυφές (The molecular overtone) και ταινίες συνδυασμού που φαίνονται στο εγγύς IR (NIRS) είναι τυπικά πολύ ευρεία, που οδηγεί σε πολύπλοκα φάσματα, μπορεί να

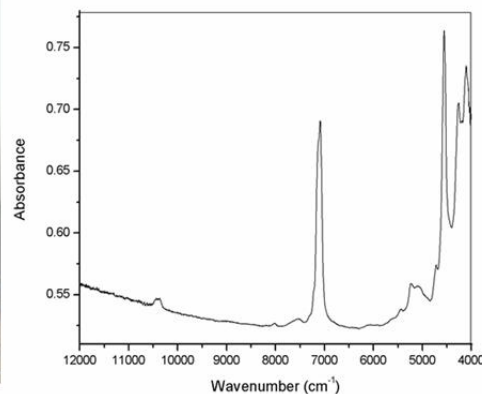
είναι δύσκολο να εκχωρήσει ειδικά χαρακτηριστικά σε συγκεκριμένα χημικά συστατικά. Οι πολυμεταβλητές (Multivariate) τεχνικές βαθμονόμησης (calibration techniques π.χ., principal components analysis, partial least squares, or artificial neural networks) χρησιμοποιούνται συχνά για να εξαχθούν οι επιθυμητές χημικές πληροφορίες, η NIR μας επιτρέπει να δημιουργήσουμε μοντέλα βαθμονόμησης για την πρόβλεψη των συγκεντρώσεων της φαρμακευτικής βιομηχανίας κατά τη διάρκεια της διαδικασίας παρασκευής.

Η NIR είναι εξαιρετική μέθοδος ανάλυσης για την μελέτη των στερεών, για παράδειγμα, στην ανάλυση των ορυκτών.

Σχήμα 27, εξέταση του βράχου Lepidolite.[4]



Lepidolite rock



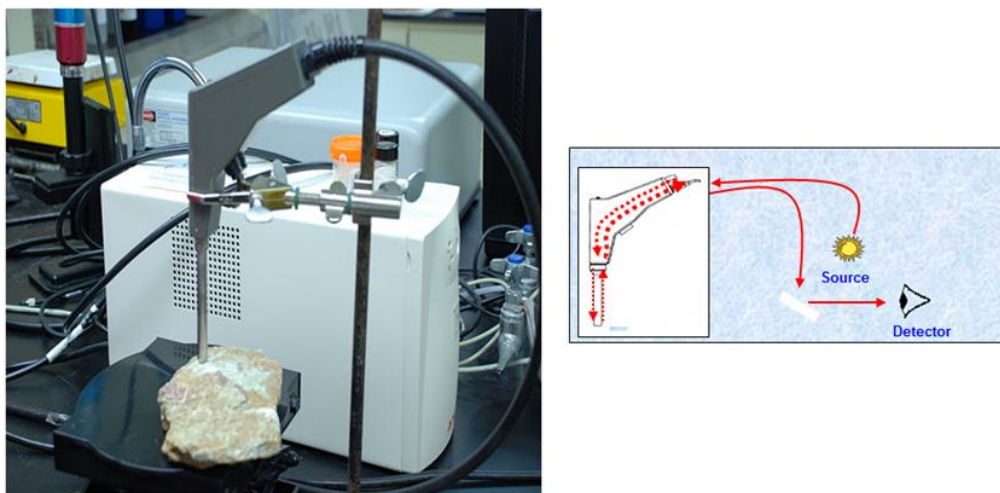
Το φάσμα λαμβάνεται με μη επεμβατικό τρόπο.

Σχήμα 28, εντελώς μη επεμβατική ανάλυση της γλυκόζης του αίματος με την NIR.[4]



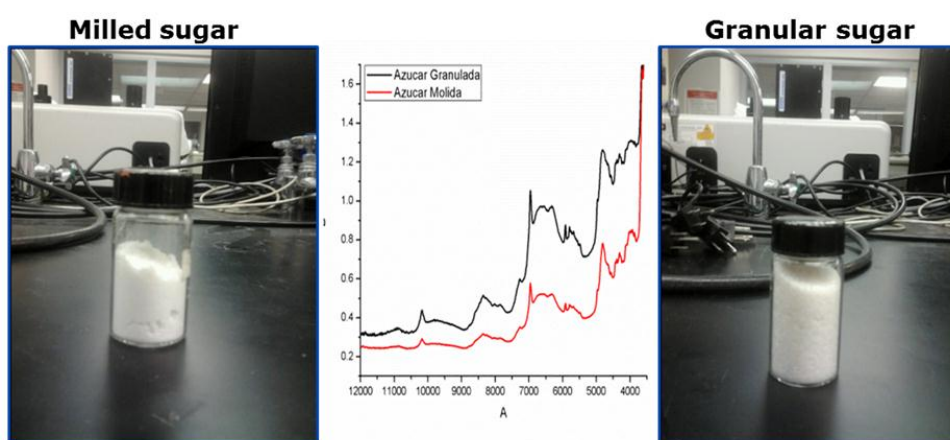
Με την NIR η απομακρυσμένη δειγματοληψία είναι δυνατή, καλή για επικίνδυνα υλικά.

Σχήμα 29, εξέταση επικίνδυνων υλικών με την NIR.[4]



Δυνατότητα χρησιμοποίησης της NIR σε ένα ευρύ φάσμα εφαρμογών (φυσικών και χημικών), και της προβολής των σχέσεων που είναι δύσκολο να παρατηρηθούν με άλλα μέσα.

Σχήμα 30, σύγκριση φάσματος αλεσμένης ζάχαρης με το φάσμα της κοκκώδους ζάχαρης.[4]

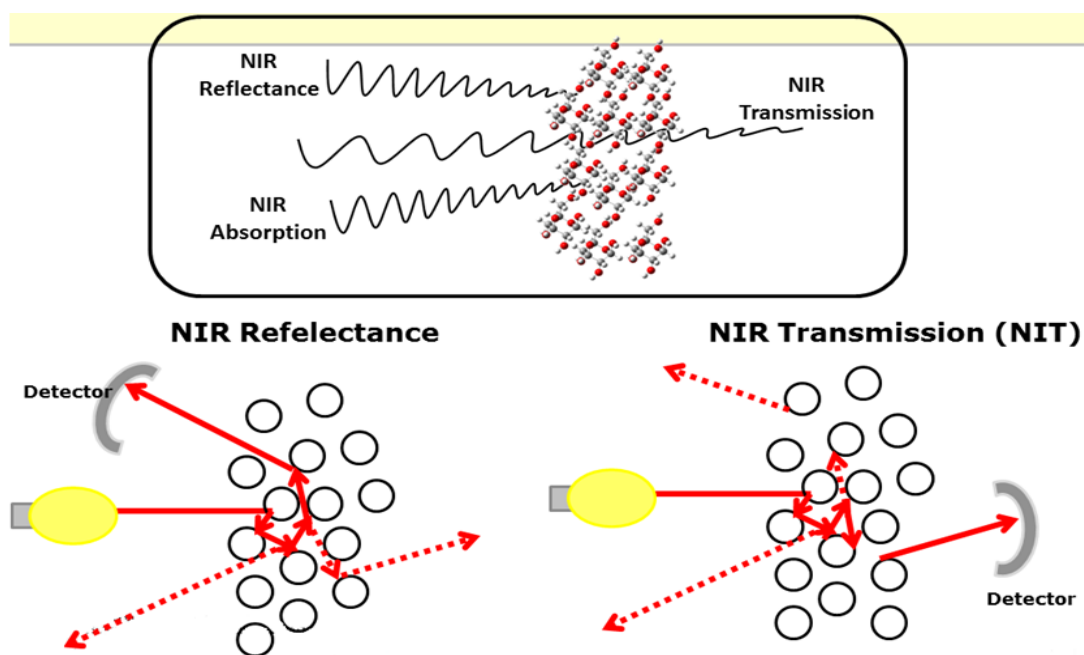


4.10 Τα όργανα

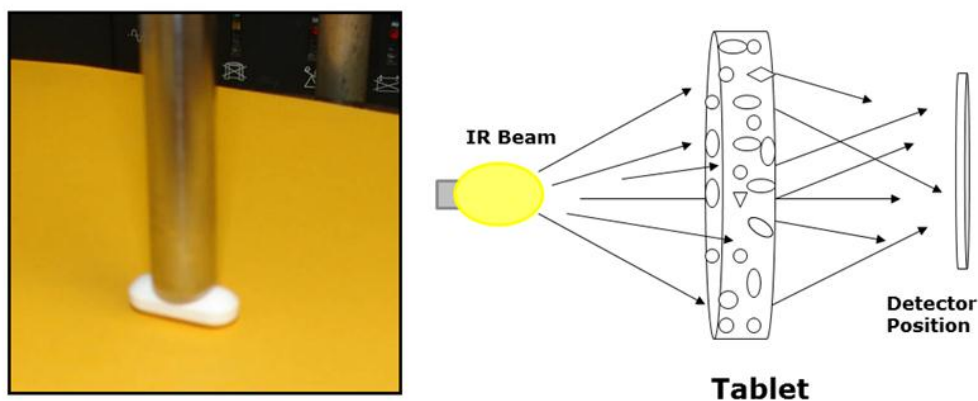
Τα όργανα για τη εγγύς IR (NIR) φασματοσκοπία είναι παρόμοια με τα όργανα για τις UV-ορατού και μεσαίου IR περιοχές(UV-visible and mid-IR ranges). Υπάρχει μια πηγή, ένας ανιχνευτής, και ένα στοιχείο διασποράς για να επιτρέψει η ένταση των διαφορετικών μήκων κύματος να καταγράφονται. Όργανα NIR με Μετασχηματισμό Fourier που χρησιμοποιούν συμβολόμετρο είναι επίσης κοινά, ειδικά για τα μήκη κύματος πάνω από ~ 1000 nm. Ανάλογα με το δείγμα, το φάσμα μπορεί να μετρηθεί είτε σε ανάκλαση (reflection) ή μετάδοση (transmission). Στην πράξη, το δείγμα προς ανάλυση NIR βομβαρδίζεται με ακτίνες διαφορετικών μηκών κύματος. Για κάθε μήκος κύματος,

ορισμένες από τις ακτίνες απορροφώνται από συγκεκριμένους χημικούς δεσμούς. Ταυτόχρονα άλλες ακτίνες είναι διάσπαρτες και αντανακλώνται από άλλους χημικούς δεσμούς. Αυτή η διαδικασία περιγράφεται συνήθως ως ανάκλαση NIR (NIR reflectance). Αντίθετα ορισμένες από τις ακτίνες περνούν μέσα από το δείγμα, το οποίο ονομάζεται μετάδοση NIR (Transmission NIR).

Σχήμα 31, ανάκλαση NIR (*reflectance NIR*), μετάδοση NIR (*Transmission NIR*), απορρόφηση NIR (*absorption NIR*).[4]

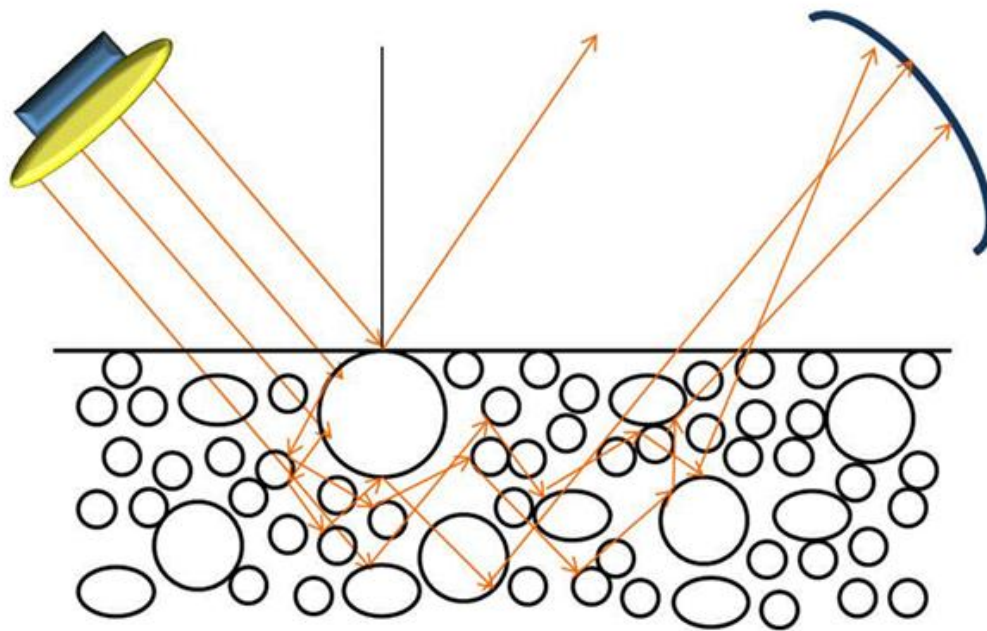


Σχήμα 32, οπτικοποίηση αλληλεπίδρασης ακτινοβολίας με σωματίδια.[4]



Η ανάκλαση ονομάζεται διάχυτη (Diffuse Reflectance) όταν η γωνία του ανακλώμενου φωτός είναι ανεξάρτητη από τη γωνία προσπτώσεως.

Σχήμα 33, διάχυτη ανάκλαση.[4]



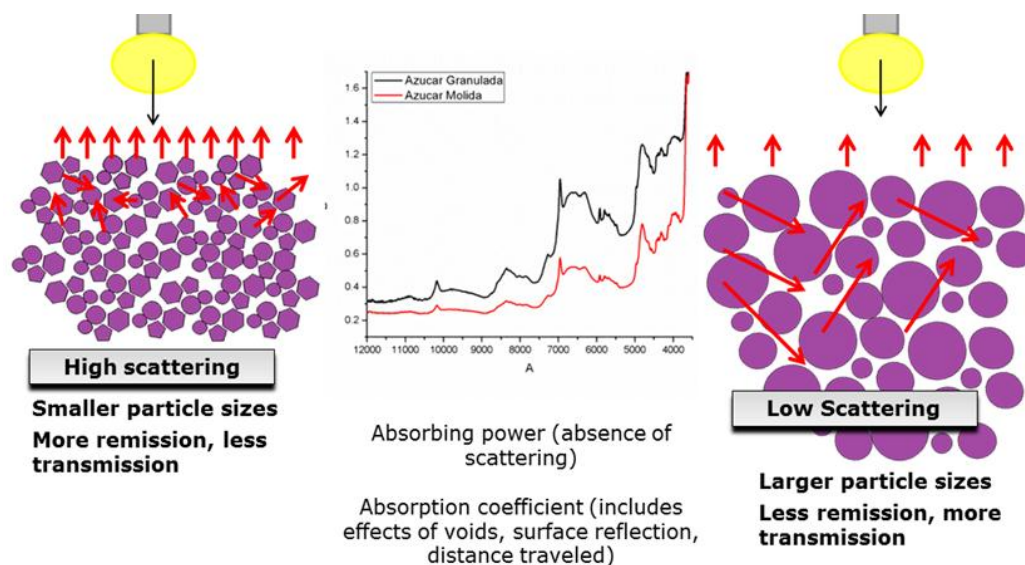
Τα φάσματα Επηρεάζονται από:

- Το μέγεθος των σωματιδίων του δείγματος.
- Πυκνότητα συσκευασίας του δείγματος, και τη πίεση επί του δείγματος.
- Δείκτης διάθλασης του δείγματος.
- Κρυσταλλική μορφή του δείγματος.
- Συντελεστές απορρόφησης του δείγματος.
- Χαρακτηριστικά της επιφάνειας του δείγματος.

Συνελεστής απορρόφησης περιλαμβάνει τις επιπτώσεις των κενών, επιφανειακή ανάκλαση, διανυθείσα απόσταση.

Η ισχύς απορρόφησης εξαρτάται από τη απουσία διασκόρπισης(scattering), μικρό μέγεθος σωματιδίων κάνει Υψηλή διασκόρπιση και κακή απορρόφηση, μεγάλο μέγεθος σωματιδίων σημαίνει χαμηλή διασκόρπιση και καλή απορρόφηση.

Σχήμα 34, το μέγεθος σωματιδίων και η διασκόρπιση, με την σύγκριση φάσματος αλεσμένης ζάχαρης με το φάσμα της κοκκώδους ζάχαρης.[4]



Κοινοί λαμπτήρες πυρακτώσεως ή χαλαζιακοί λαμπτήρες αλογόνου (Common incandescent or quartz halogen light bulbs) είναι πιο συχνά χρησιμοποιούνται ως ευρυζωνικές πηγές εγγύς υπέρυθρης ακτινοβολίας (near-infrared radiation) για αναλυτικές εφαρμογές. Επίσης χρησιμοποιούνται οι δίοδοι εκπομπής φωτός (LED) οι οποίοι προσφέρουν μεγαλύτερη διάρκεια ζωής και φασματική σταθερότητα και μειωμένες απαιτήσεις ισχύος.

Ο τύπος του ανιχνευτή που χρησιμοποιείται εξαρτάται κυρίως από την έκταση του μήκους κύματος που πρόκειται να μετρηθεί, CCDs (charge-coupled device). Βασίζονται στο πυρίτιο είναι κατάλληλα για το βραχύτερο άκρο του φάσματος NIR, αλλά δεν είναι αρκετά ευαίσθητα στο μεγαλύτερο μέρος της περιοχής (πάνω από 1000 nm), InGaAs (Indium gallium arsenide) και PBS (Lead(II) sulfide (Lead 'Pb', Sulfur 'S')) συσκευές είναι πιο κατάλληλες αν και λιγότερο ευαίσθητες από CCDs. Σε ορισμένα όργανα συστοιχίας διόδων (DA (diode array)) NIRS, και με βάση το πυρίτιο και InGaAs ανιχνευτές χρησιμοποιούνται στο ίδιο όργανο. Τέτοια όργανα μπορούν να καταγράψουν τόσο UV-ορατό και NIR φάσματα «ταυτόχρονα».

Τα όργανα για την χημική απεικόνιση (chemical imaging) στον NIR μπορεί να χρησιμοποιήσει ένα 2D ανιχνευτή συστοιχίας με οπτικοακουστικό συντονιζόμενο φίλτρο (an acousto-optic tunable filter), Πολλαπλές εικόνες μπορούν να καταγράφονται διαδοχικά σε διαφορετικές στενές ζώνες μήκους κύματος.

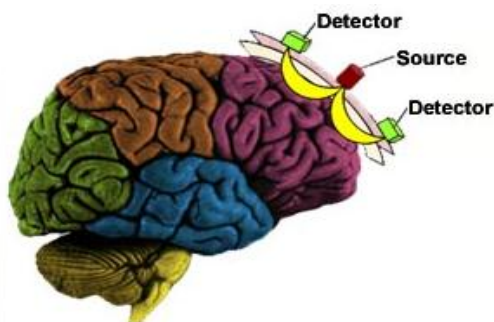
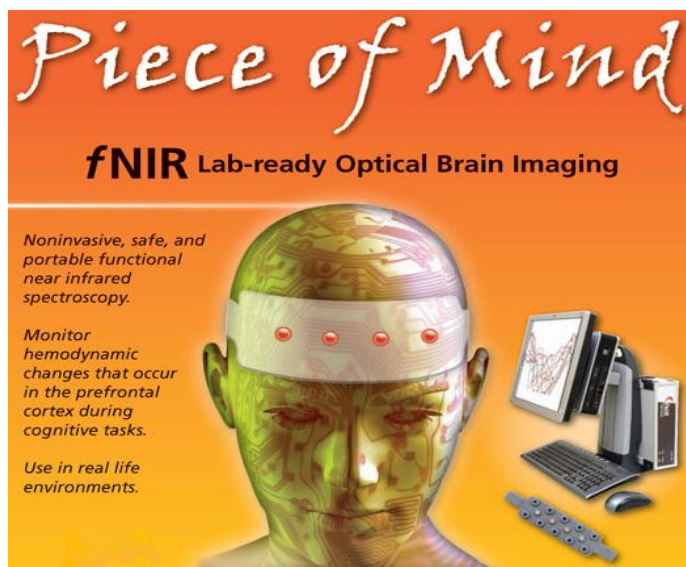
Πολλά εμπορικά όργανα για UV / VIS φασματοσκοπία είναι ικανά για εγγραφή φασμάτων στο εύρος NIR (ίσως μέχρι ~ 900 nm). Κατά τον ίδιο τρόπο, το εύρος μερικών οργάνων mid-IR μπορεί να εκτείνεται μέσα στο NIR, Σε αυτά τα όργανα, ο ανιχνευτής που χρησιμοποιείται για τα μήκη κύματος NIR συχνά είναι ο ίδιος ανιχνευτής που χρησιμοποιείται για την "κύρια" περιοχή ενδιαφέροντος του οργάνου.

5. Εφαρμογές της NIR

5.1 Εισαγωγή

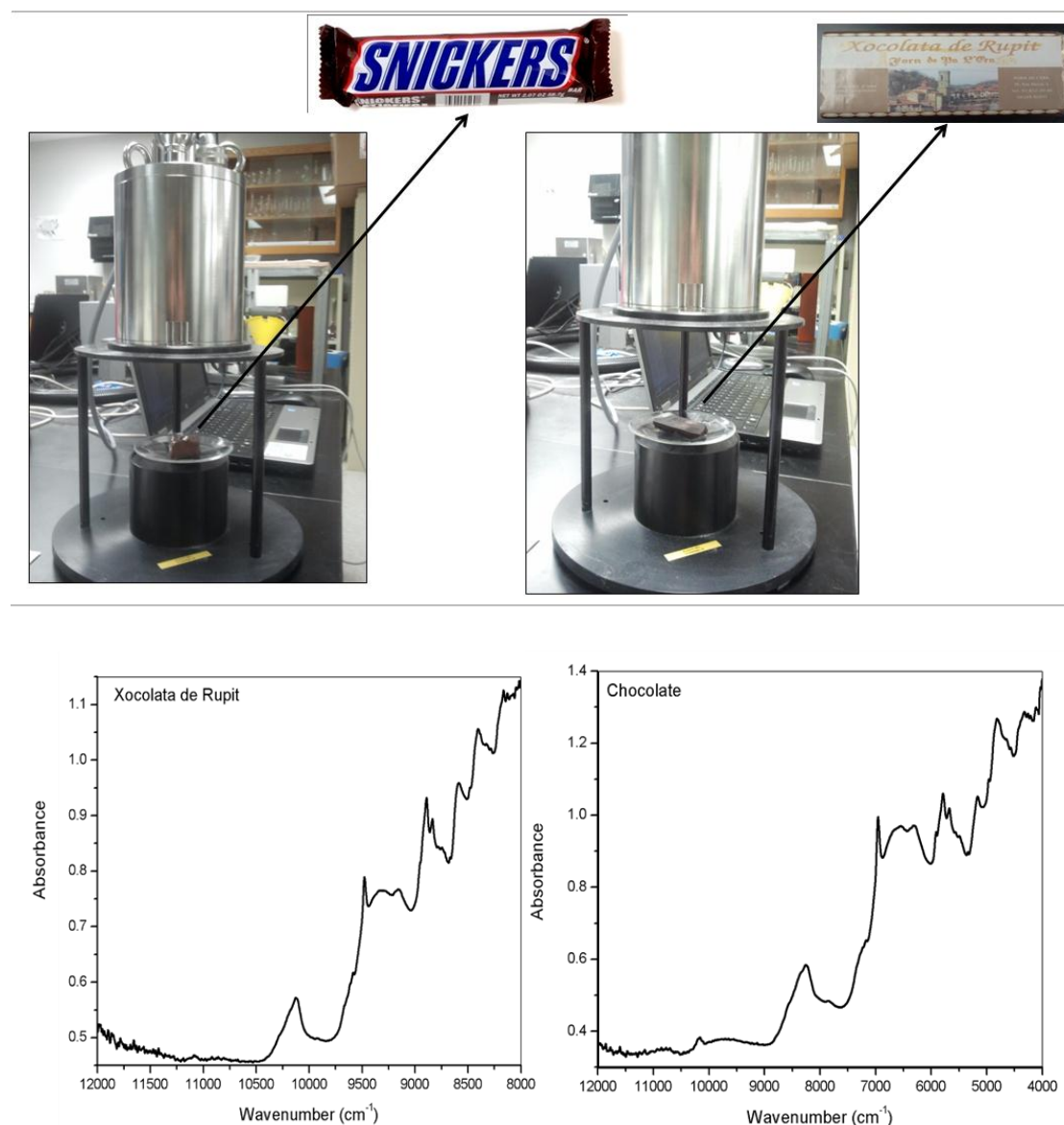
Η κύρια εφαρμογή της NIRS στο ανθρώπινο σώμα χρησιμοποιεί το γεγονός ότι η μετάδοση και η απορρόφηση του φωτός της NIR στους ανθρώπινους ιστούς του σώματος περιέχει πληροφορίες σχετικά με τις μεταβολές της συγκέντρωσης αιμοσφαιρίνης. Όταν μια συγκεκριμένη περιοχή του εγκεφάλου είναι ενεργοποιημένη, ο εντοπισμένος όγκος του αίματος στην περιοχή αυτή αλλάζει γρήγορα. Οπτική απεικόνιση μπορεί να μετρήσει τη θέση και τη δραστηριότητα των συγκεκριμένων περιοχών του εγκεφάλου από τη συνεχή παρακολούθηση των επιπέδων της αιμοσφαιρίνης του αίματος μέσω του προσδιορισμού των οπτικών συντελεστών απορρόφησης.

Σχήμα 35, *fNIR(functional near infrared) οπτικά συστήματα απεικόνισης της λειτουργικής NIR, Μετρά τις μεταβολές της στάθμης του οξυγόνου στον προμετωπιαίο φλοιό του ανθρώπινου υποκείμενου.*[13]



Τυπικές εφαρμογές της φασματοσκοπίας NIR περιλαμβάνουν την ανάλυση των τροφίμων, των φαρμακευτικών προϊόντων, των προϊόντων καύσης και ένα σημαντικό κλάδο της αστρονομικής φασματοσκοπίας.

Σχήμα 36, ανάλυση δυο ειδών σοκολάτας με τα δικά τους φάσματα NIR που προκύπτουν μετά την ανάλυση.[4]



5.2 Αστρονομική φασματοσκοπία

Η εγγύς υπέρυθρη φασματοσκοπία (Near-infrared spectroscopy) χρησιμοποιείται στην αστρονομία για τη μελέτη των ατμοσφαιρών δροσερών αστερών, όπου τα μόρια μπορούν να σχηματιστούν. Οι δονητικές και περιστροφικές υπογραφές των μορίων όπως το οξείδιο του τιτανίου, κυάνιο, και μονοξειδίου του άνθρακα μπορούν να φανούν σε αυτή την περιοχή μήκους κύματος και μπορούν να δώσουν μια ένδειξη για τον φασματικό τύπο του άστρου, επίσης χρησιμοποιείται για τη μελέτη μορίων σε άλλα αστρονομικά περιβάλλοντα, όπως σε μοριακά νέφη όπου σχηματίζονται νέα άστρα. Το αστρονομικό φαινόμενο γνωστό ως (reddening) σημαίνει ότι τα μήκη κύματος της εγγύς υπέρυθρης (near-infrared) επηρεάζονται λιγότερο από τη σκόνη στο διαστρικό μέσο, έτσι ώστε οι περιοχές απρόσιτες από την οπτική φασματοσκοπία μπορούν να μελετηθούν στην εγγύς υπέρυθρη. Δεδομένου

ότι η σκόνη και το αέριο συνδέονται στενά, οι σκονισμένες περιοχές είναι ακριβώς αυτές όπου υπέρυθρη φασματοσκοπία είναι η πιο χρήσιμη, Το φάσμα της εγγύς υπέρυθρης (near-infrared) των πολύ νεαρών άστρων παρέχουν σημαντικές πληροφορίες σχετικά με τις ηλικίες και οι μάζες τους, πράγμα που είναι σημαντικό για την κατανόηση του σχηματισμού των αστεριών σε γενικές γραμμές.

5.3 Γεωργία

Εγγύς υπέρυθρη φασματοσκοπία εφαρμόζεται ευρέως στη γεωργία για τον προσδιορισμό ποιότητας των ζωοτροφών, των σπόρων και των δημητριακών προϊόντων, των ελαιούχων σπόρων, του καφέ, του τσαγιού, των μπαχαρικών, των φρούτων, των λαχανικών, του ζαχαροκάλαμου, των ποτών, των λιπών, των ελαίων, των γαλακτοκομικών προϊόντων, των αυγών, του κρέατος και άλλων γεωργικών προϊόντων. Η Εγγύς υπέρυθρη φασματοσκοπία NIR χρησιμοποιείται ευρέως για την ποσοτικοποίηση της σύνθεσης των γεωργικών προϊόντων, διότι ικανοποιεί τα κριτήρια να είναι ακριβής, αξιόπιστη, ταχεία, μη καταστρεπτική, και ανέξοδη.

5.4 Απομακρυσμένη παρακολούθηση

Τεχνικές έχουν αναπτυχθεί για φασματοσκοπική απεικόνιση της NIR. Υπερφασματική απεικόνιση (Hyperspectral imaging) έχει εφαρμοστεί για ένα ευρύ φάσμα χρήσεων, συμπεριλαμβανομένης της απομακρυσμένης έρευνας των φυτών και των εδαφών. Τα δεδομένα μπορούν να συλλέγονται από όργανα σε αεροπλάνα ή δορυφόρους για να εκτιμηθούν τα μέτρα κάλυψης του εδάφους και τη χημεία του εδάφους.

5.5 Επιστήμη των υλικών

Έχουν αναπτυχθεί τεχνικές για τη φασματοσκοπία NIR των μικροσκοπικών περιοχών δείγματος για μετρήσεις πάχους μεμβράνης, η έρευνα στα οπτικά χαρακτηριστικά των νανοσωματιδίων και οπτικές επικαλύψεις για τη βιομηχανία των τηλεπικοινωνιών.

5.6 Ιατρικές χρήσεις

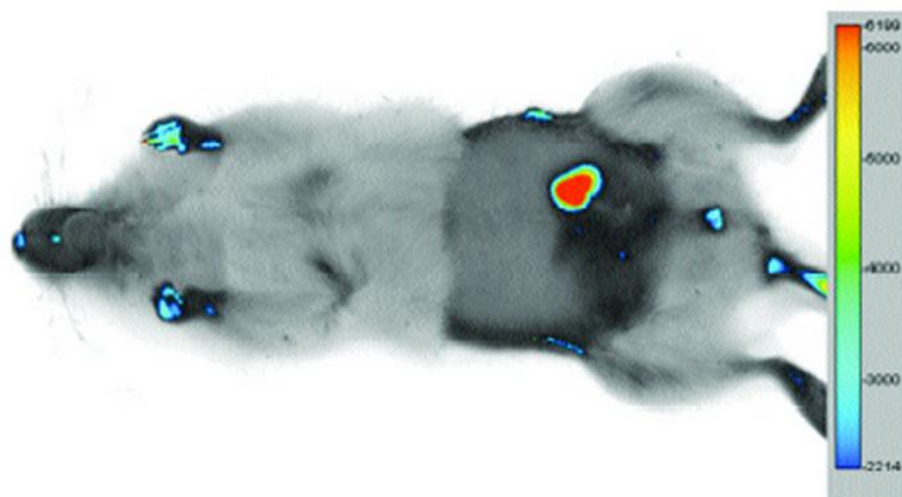
Ιατρικές εφαρμογές του κέντρου της NIRS για την μη επεμβατική μέτρηση (non-invasive measurement) της ποσότητας και του περιεχομένου οξυγόνου της αιμοσφαιρίνης, καθώς και τη χρήση των οπτικών εξωγενούς ιχνηθετών σε συνδυασμό με κινητική ροή.

NIRS μπορεί να χρησιμοποιηθεί για μη επεμβατική εκτίμηση της λειτουργίας του εγκεφάλου μέσω του άθικτου κρανίου σε ανθρώπινα υποκείμενα με ανίχνευση των αλλαγών στις συγκεντρώσεις αιμοσφαιρίνης του αίματος που σχετίζονται με την νευρική δραστηριότητα, Π.χ., στους κλάδους της γνωστικής ψυχολογίας (Cognitive psychology) ως μερική αντικατάσταση fMRI (Functional magnetic resonance imaging) τεχνικές. Η NIRS μπορεί να χρησιμοποιηθεί σε βρέφη και NIRS είναι πολύ πιο φορητή από fMRI μηχανές, ακόμα και ασύρματα όργανα είναι διαθέσιμα, η οποία επιτρέπουν τις έρευνες σε ελεύθερα κινούμενα θέματα. Ωστόσο, η NIRS δεν μπορεί αντικαταστήσει πλήρως την fMRI, επειδή το μόνο για το οποίο μπορεί να χρησιμοποιηθεί είναι για να ανιχνεύσει φλοιώδη ιστό, εκεί όπου η fMRI μπορεί να χρησιμοποιηθεί για τη μέτρηση της ενεργοποίησης σε όλο τον

εγκέφαλο. Στατιστικές εργαλειοθήκες για την ανάλυση αυτόνομων και συνδυασμένων NIRS / MRI μετρήσεων έχουν αναπτυχθεί.

Η εφαρμογή σε λειτουργική χαρτογράφηση του ανθρώπινου φλοιού ονομάζεται διάχυτη οπτική τομογραφία DOT (diffuse optical tomography), εγγύς υπέρυθρη απεικόνιση NIRI (near infrared imaging) ή λειτουργική NIRS (functional NIRS « fNIR »), ο όρος διάχυτη οπτική τομογραφία χρησιμοποιείται για τα τρισδιάστατα NIRS. Οι όροι NIRS, NIRI και DOT συχνά χρησιμοποιούνται αδιακρίτως αλλά έχουν κάποιες διαφορές, η πιο σημαντική διαφορά μεταξύ NIRS και DOT / NIRI είναι ότι DOT / NIRI χρησιμοποιούνται κυρίως για την ανίχνευση μεταβολών στις οπτικές ιδιότητες του ιστού ταυτόχρονα από πολλαπλά σημεία μέτρησης και εμφανίζει τα αποτελέσματα με τη μορφή ενός χάρτη ή εικόνα σε μια συγκεκριμένη περιοχή. Στο επόμενο σχήμα φαίνεται το αποτέλεσμα όταν χρησιμοποιείται η NIR για την ανίχνευση όγκου προστάτη του ποντικού μετά την έγχυση του ποντικού με τη βαφή της NIR (IRDye 800CW-EGF).

Σχήμα 37, απεικόνιση όγκου του προστάτη του ποντικού το οποίο εγχύθηκε με βαφή της NIR (IRDye 800CW-EGF).[14]



Ενώ NIRS παρέχει ποσοτικά δεδομένα σε απόλυτες τιμές μέχρι μερικά συγκεκριμένα σημεία, η τελευταία χρησιμοποιείται επίσης για τη διερεύνηση άλλους ιστούς όπως, π.χ., των μυών, του μαστού και όγκου, NIRS μπορεί να χρησιμοποιηθεί για την ποσοτικοποίηση της ροής του αίματος, του όγκου του αίματος, κατανάλωση οξυγόνου, τα ποσοστά επανοξυγόνωσης και μυϊκή χρόνο αποκατάστασης στους μύες.

Χρησιμοποιώντας αρκετά μήκη κύματος και time resolved (συχνότητα ή πεδίο του χρόνου) και/ή χωρικές αποφασισμένες μέθοδοι ροή του αίματος (spatially resolved methods blood flow), ο όγκος και ο απόλυτος κορεσμός ιστού (volume and absolute tissue saturation) (StO_2 or Tissue Saturation Index (TSI)) μπορούν να ποσοτικοποιηθούν. Εφαρμογές της οξυμετρίας με NIRS μεθόδους περιλαμβάνουν νευροεπιστήμη, την εργονομία, την αποκατάσταση, επικοινωνία εγκεφάλου-υπολογιστή, ουρολογία, τη διάγνωση των ασθενειών που επηρεάζουν την κυκλοφορία του αίματος (π.χ., περιφερική αγγειακή ασθένεια), την ανίχνευση και την εκτίμηση των όγκων του μαστού, καθώς και τη βελτιστοποίηση της κατάρτισης στην αθλητιατρική.

Η χρήση της NIRS σε συνδυασμό με έναν βλωμό ένεσης πράσινου ινδοκυανίνης (indocyanine green) (ICG) έχει χρησιμοποιηθεί για τη μέτρηση της εγκεφαλικής ροής του αίματος και εγκεφαλικού μεταβολικού ρυθμού κατανάλωσης οξυγόνου, έχει επίσηςδειχθεί ότι CMRO₂ (cerebral metabolic rate for O₂) μπορεί να υπολογιστεί σε συνδυασμένες NIRS / MRI μετρήσεις.

NIRS έχει αρχίσει να χρησιμοποιείται στην παιδιατρική εντατική θεραπεία, να συμβάλει στην αντιμετώπιση με το χρόνο αμέσως μετά από μια καρδιοχειρουργική επέμβαση. Πράγματι, NIRS είναι σε θέση να μετρήσει φλεβικό κορεσμό οξυγόνου (SVO₂) ο οποίος καθορίζεται από την καρδιακή παροχή, καθώς και άλλους παραμέτρους (FiO₂, αιμοσφαιρίνη, η πρόσληψη οξυγόνου). Επομένως η NIRS δίνει στους κρίσιμους γιατρούς μια αντίληψη της καρδιακής εξόδου. NIRS είναι αρεστή από τους ασθενείς, επειδή είναι μη επεμβατική (non-invasive), είναι ανώδυνη, και χρησιμοποιεί μη ιοντίζουσες ακτινοβολίες (non-ionizing radiation).

Οπτική τομογραφία συνοχής (Optical Coherence Tomography (OCT)) είναι μια άλλη ιατρική τεχνική απεικόνισης ικανή για 3D απεικόνισης με υψηλή ανάλυση στο ίδιο επίπεδο με χαμηλής ισχύος μικροσκόπιο. Χρησιμοποιώντας οπτική συνοχή για τη μέτρηση μήκους διαδρομής φωτονίων επιτρέπει OCT κατασκευάσει τις εικόνες των ζωντανών ιστών και σαφή εξετάσεις της μορφολογίας του ιστού. Η OCT περιορίζεται στην απεικόνιση 1-2 mm κάτω επιφάνειες του ιστού αλλά παρά τον περιορισμό αυτό OCT έχει γίνει μια καθιερωμένη ιατρική τεχνική απεικόνισης ειδικά για την απεικόνιση του αμφιβληστροειδούς και πρόσθιο τμήμα του οφθαλμού.

Σχήμα 38, το σύστημα Topcon 3D OCT-2000 είναι το πρώτο φασματικό σύστημα τομέα OCT ενσωματώνει μια κάμερα υψηλής ανάλυσης του βυθού και έγχρωμη οθόνη αφής φιλική προς το χρήστη σε μια συμπαγή σχεδίαση για εξοικονόμηση χώρου.[15]



Η οργανική ανάπτυξη των NIRS / NIRI / DOT / OCT έχει προχωρήσει σημαντικά κατά τη διάρκεια των τελευταίων ετών και ιδίως όσον αφορά την ποσοτικοποίηση, απεικόνιση και σμίκρυνση.

5.7 Μέτρηση σωματιδίων

NIR χρησιμοποιείται συχνά στο μέτρημα μέγεθους σωματιδίου σε μια σειρά από διάφορους τομείς, συμπεριλαμβανομένων μελέτης φαρμακευτικών και γεωργικών σκόνες.

5.8 Βιομηχανικές χρήσεις

Σε αντίθεση με τη NIRS που χρησιμοποιείται στην οπτική τοπογραφία, η NIRS που χρησιμοποιείται για τον προσδιορισμό χημικών δεν παρέχει απεικόνιση με χαρτογράφηση. Για παράδειγμα, ένας κλινικός αναλυτής διοξειδίου του άνθρακα CO₂ απαιτεί τεχνικές αναφοράς και ρουτίνες βαθμονόμησης για να είναι σε θέση να πάρει ακριβή μεταβολή του περιεχομένου CO₂. Σε αυτή την περίπτωση, βαθμονόμηση εκτελείται με ρύθμιση του μηδενικού ελέγχου του δείγματος που δοκιμάζεται μετά από σκόπιμη παροχή 0% CO₂ ή άλλη γνωστή ποσότητα του CO₂ στο δείγμα. Κανονικό συμπιεσμένο αέριο από τους διανομείς περιέχει περίπου 95% O₂ και 5% CO₂, το οποίο μπορεί επίσης να χρησιμοποιηθεί για την προσαρμογή % CO₂ του μετρητή ανάγνωσης να είναι ακριβώς στο 5% της αρχικής βαθμονόμησης.

6. PARTIAL LEAST-SQUARES REGRESSION

6.1 Η σημασία της PLS μεθόδου

Partial least squares (PLS) είναι μια δημοφιλής μέθοδος για μαλακή Μοντελοποίηση (soft modelling) σε βιομηχανικές εφαρμογές. Η PLS είναι μια γενικευμένη μέθοδος παλινδρόμησης (generalized regression method) που χρησιμοποιείται για την κατασκευή ενός μοντέλου πρόβλεψης μεταξύ των δύο μπλοκ μεταβλητών: το Χ-μπλοκ των μεταβλητών πρόβλεψης (predictor variables) και το Υ-μπλοκ των μεταβλητών απόκρισης (response variables).

Η Partial least squares (PLS) αποκτά μεγαλύτερη σημασία σε πολλούς τομείς της χημείας, της αναλυτικής, φυσικής, κλινικής χημείας και της βιομηχανικής. Αρθρα δείχνουν ότι PLS είναι μια καλή εναλλακτική λύση για τις περισσότερες κλασσικές πολλαπλές γραμμικές μεθόδους παλινδρόμησης (multiple linear regression methods) και μεθόδους κύριων συνιστωσών παλινδρόμησης (principal component regression methods) επειδή είναι πιο ισχυρή. Αυτό σημαίνει ότι οι παράμετροι του μοντέλου δεν αλλάζουν πάρα πολύ όταν τα νέα δείγματα βαθμονόμησης (new calibration samples) λαμβάνονται από το σύνολο πληθυσμού.

Η έρευνα στον τομέα της επιστήμης και της μηχανικής συχνά περιλαμβάνει την χρήση ελεγχόμενων ή/και εύκολων στη μέτρηση μεταβλητών (παράγοντες (factors)) για να εξηγήσει, να ρυθμίζει, ή να προβλέψει τη συμπεριφορά των άλλων μεταβλητών (αποκρίσεις).

Όταν οι παράγοντες είναι λίγοι σε αριθμό, δεν είναι σημαντικά συγγραμμικοί, και έχουν μια καλά κατανοητή σχέση με τις αποκρίσεις, τότε η πολλαπλή γραμμική παλινδρόμηση (multiple linear regression (MLR)) μπορεί να είναι ένας καλός τρόπος για να μετατραπούν τα δεδομένα σε πληροφορίες. Ωστόσο, εάν οποιαδήποτε από αυτές τις τρεις προϋποθέσεις καταρρέει, η MLR μπορεί να είναι αναποτελεσματική ή ακατάλληλη.

Σε τέτοιες λεγόμενες μαλακές επιστημονικές εφαρμογές (soft science applications), ο ερευνητής βρίσκεται αντιμέτωπος με πολλές μεταβλητές και ακατανόητες σχέσεις και το αντικείμενο μπορεί απλώς να κατασκευάσει ένα καλό προγνωστικό μοντέλο. Για παράδειγμα, οι φασματογράφοι χρησιμοποιούνται συχνά για να εκτιμηθεί η ποσότητα των διαφόρων ενώσεων σε ένα χημικό δείγμα, στην περίπτωση αυτή, οι παράγοντες είναι οι μετρήσεις που αποτελούν το φάσμα, μπορούν να ανέρχονται σε εκατοντάδες, αλλά είναι πιθανό να είναι άκρως συγγραμμικοί (highly collinear). Οι αποκρίσεις είναι τα ποσά συνιστώσας τα οποία ο ερευνητής επιθυμεί να προβλέψει στα δείγματα του μέλλοντος. Η Partial least squares (PLS) είναι μια μέθοδος για την κατασκευή μοντέλων πρόβλεψης όταν οι παράγοντες είναι πολλοί και άκρως συγγραμμικοί. Να σημειωθεί ότι η έμφαση δίνεται στην πρόβλεψη των αποκρίσεων και όχι κατ' ανάγκη στην προσπάθεια να κατανοήσουμε την υποκείμενη σχέση μεταξύ των μεταβλητών. Για παράδειγμα, PLS συνήθως δεν είναι κατάλληλη για τη διαλογή των παραγόντων που έχουν αμελητέα επίδραση στην απόκριση. Ωστόσο, όταν η πρόβλεψη είναι ο στόχος και δεν υπάρχει καμία πρακτική ανάγκη να

περιοριστεί ο αριθμός των μετρούμενων παραγόντων, η PLS μπορεί να είναι ένα χρήσιμο εργαλείο. Η PLS αναπτύχθηκε στη δεκαετία του 1960 από τον Herman Wold ως οικονομετρική τεχνική, αλλά μερικοί από τους πιο μακρινούς υποστηρικτές της (συμπεριλαμβανομένου Wold's son Svante) είναι χημικοί μηχανικοί. Εκτός από την φασματομετρική βαθμονόμησης όπως συζητήθηκε παραπάνω, η PLS έχει εφαρμοστεί για την παρακολούθηση και τον έλεγχο βιομηχανικών διεργασιών, μια μεγάλη διαδικασία μπορεί εύκολα να έχει εκατοντάδες ελεγχόμενων μεταβλητών και δεκάδες εξόδους. Οι επόμενες ενότητες παρέχουν μια επισκόπηση του τρόπου με τον οποίο η PLS λειτουργεί, σχετίζοντας την με άλλες τεχνικές πολλαπλών μεταβλητών, όπως principal components regression analysis, multiple linear regression και maximum redundancy analysis.

6.2 Η PLS μέθοδος, ξεκινώντας από άλλες τεχνικές πολλαπλών μεταβλητών

6.2.1 Πολλαπλή γραμμική παλινδρόμηση (MULTIPLE LINEAR REGRESSION (MLR))

Στην πολλαπλή γραμμική παλινδρόμηση (MLR), το πρόβλημα μπορεί να διατυπωθεί ως εξής :

Τα χαρακτηριστικά μετρούνται για m μεταβλητές x_j ($j = 1 \dots m$) και για μια μεταβλητή y , με στόχο να δημιουργηθεί μια γραμμική (ή πρώτης τάξης) σχέση μεταξύ τους, αυτό μπορεί να παρασταθεί μαθηματικά ως:

$$y = b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots + b_m x_m + e \quad (1a)$$

$$y = \sum_{j=1}^m b_j x_j + e \quad (1b)$$

$$y = x'b + e \quad (1c)$$

Στην εξίσωση (1a) οι x_j ονομάζονται ανεξάρτητες μεταβλητές και το y είναι η εξαρτημένη μεταβλητή, και οι b_j είναι ευαισθησίες (sensitivities) και e είναι το σφάλμα ή κατάλοιπο (residual). Στην εξίσωση (1c) η y είναι βαθμωτό (scalar), η b είναι ένα διάνυσμα στήλης, και το x' είναι ένα διάνυσμα γραμμής.

Η εξίσωση 1 περιγράφει πολυγραμμικές εξαρτήσεις (multilinear dependencies) για ένα μόνο δείγμα. Εάν κάποιος παίρνει n δείγματα, y_i ($i = 1 \dots n$) μπορεί να γραφτεί ως y διάνυσμα στήλης, b παραμένει το ίδιο, και τα διανύσματα, x_i' , σχηματίζουν τις γραμμές ενός πίνακα X :

$$y = Xb + e \quad (2)$$

Για μια καλύτερη κατανόηση αυτής της εξίσωσης, δίνεται επίσης σε γραφική αναπαράσταση:

The diagram illustrates the matrix equation $y = Xb + e$. On the left, a vertical vector y is shown with a dimension of n at the bottom and a dimension of 1 at the top. This is followed by an equals sign. To the right of the equals sign is a matrix X with a dimension of n at the bottom and a dimension of m at the top. This is followed by a vertical vector b with a dimension of m at the bottom and a dimension of 1 at the top. To the right of b is a plus sign, followed by a vertical vector e with a dimension of n at the bottom and a dimension of 1 at the top.

Στην περίπτωση αυτή, n είναι ο αριθμός των δειγμάτων και m ο αριθμός των ανεξάρτητων μεταβλητών.

Τώρα είναι δυνατό να γίνει διάκριση ανάμεσα σε τρεις περιπτώσεις.

(1) $m > n$: υπάρχουν περισσότερες μεταβλητές από τα δείγματα.

Σε αυτή την περίπτωση, υπάρχουν άπειρες λύσεις για το b , οι οποίες όλες ταιριάζουν στην εξίσωση, δεν είναι αυτό που θέλουμε.

(2) $m = n$: ο αριθμός των δειγμάτων και των μεταβλητών είναι ίσος, η κατάσταση αυτή μπορεί να μην αντιμετωπίζεται συχνά σε πρακτικές περιπτώσεις, ωστόσο, δίνει μια μοναδική λύση για το b υπό την προϋπόθεση ότι το X έχει πλήρη βαθμό (full rank). Αυτό μας επιτρέπει να γράψουμε.

$$e = y - Xb = 0 \quad (3)$$

e ονομάζεται διάνυσμα υπολειμματικού (residual vector). Σε αυτή την περίπτωση, είναι το μηδενικό διάνυσμα: 0

(3) $m < n$: υπάρχουν περισσότερα δείγματα από τις μεταβλητές, αυτό δεν επιτρέπει μια ακριβή λύση για το b , αλλά μπορεί κανείς να πάρει μια λύση με την ελαχιστοποίηση του μήκους του διανύσματος υπολειμματικού (residual vector) e στην ακόλουθη εξίσωση:

$$e = y - Xb \quad (4)$$

Η πιο δημοφιλής μέθοδος για να γίνει αυτό ονομάζεται «μέθοδος ελαχίστων τετραγώνων (least-squares method)» η λύση των ελαχίστων τετραγώνων είναι

$$b = (X'X)^{-1} X' y \quad (5)$$

Η εξίσωση 5 δίνει μια ιδέα για το πιο συχνό πρόβλημα στην MLR: το αντίστροφο του $X'X$ μπορεί να μην υπάρχει. Συγγραμμικότητα (Collinearity), μηδενική ορίζουσα (zero determinant) και μοναδικότητα (singularity) είναι όλα ονόματα για το ίδιο πρόβλημα.

Σε αυτό το σημείο, μπορεί να φαίνεται ότι πάντα πρέπει να είναι τουλάχιστον τόσα πολλά δείγματα όσες μεταβλητές, αλλά υπάρχουν και άλλοι τρόποι για τη διαμόρφωση αυτού του προβλήματος κι ένας από αυτούς είναι να διαγράψουμε κάποιες μεταβλητές στην περίπτωση $m > n$.

6.2.2 Πολλαπλή γραμμική παλινδρόμηση (MULTIPLE LINEAR REGRESSION (MLR)) με περισσότερες από μία εξαρτημένη μεταβλητή.

Είναι εύκολο να επεκταθεί η MLR για περισσότερες εξαρτημένες μεταβλητές. Το παράδειγμα που δίνεται εδώ είναι για δύο μεταβλητές, αλλά η επέκταση σε περισσότερες από δύο είναι απλή.

Ας υποθέσουμε ότι υπάρχουν δύο εξαρτημένες μεταβλητές, y_1 και y_2 , στην περίπτωση αυτή, μπορεί κανείς να γράψει απλώς δύο MLR και βρίσκει δύο διανύσματα των ευαισθησιών, b_1 και b_2 :

$$y_1 = Xb_1 + e_1 \quad ; \quad y_2 = Xb_2 + e_2 \quad (6)$$

Αλλά μπορεί κανείς να θέσει στη συνέχεια y_1 και y_2 πλάι-πλάι σε έναν πίνακα $n \times 2$ και κάνουμε το ίδιο για b_1 και b_2 και e_1 και e_2 .

$$Y = XB + E \quad (7)$$

Όπου $Y = (y_1 \ y_2)$, $B = (b_1 \ b_2)$ και $E = (e_1 \ e_2)$

Μια γραφική αναπαράσταση για 2 - p εξαρτημένες μεταβλητές είναι

$$\begin{array}{c} 2-p \\ \boxed{Y} \\ n \end{array} = \begin{array}{c} m \\ \boxed{X} \\ n \end{array} \begin{array}{c} 2-p \\ \boxed{B} \\ m \end{array} + \begin{array}{c} 2-p \\ \boxed{E} \\ n \end{array}$$

Αυτή είναι η γενική περίπτωση που θα αναφέρεται στο περαιτέρω κείμενο.

6.2.3 Ανάλυση Κύριων Συνιστωσών (PRINCIPAL COMPONENT ANALYSIS (PCA)): ΜΕΘΟΔΟΣ NIPALS

Ανάλυση Κύριων Συνιστωσών (PCA) είναι μια μέθοδος για τη σύνταξη ενός πίνακα X βαθμού r (rank r) ως άθροισμα r πινάκων βαθμού 1:

$$X = M_1 + M_2 + M_3 + \dots + M_r \quad (8)$$

Ή στη γραφική αναπαράσταση

$$\boxed{X} = \boxed{M_1} + \boxed{M_2} + \dots + \boxed{M_r}$$

(ο βαθμός πίνακα (The matrix rank) προσδιορίζεται από τον αριθμό των ανεξάρτητων γραμμών ή στηλών που υπάρχουν σε αυτό)

Αυτοί οι πίνακες M_h με βαθμό 1 μπορούν να γραφτούν ως εξωτερικά γινόμενα δύο διανυσμάτων, score t_h και loading p'_h :

$$X = t_1 p'_1 + t_2 p'_2 + \dots + t_a p'_a \quad (9)$$

ή το ισοδύναμο $X = TP'$ (P' αποτελείται από p' ως γραμμές και T από t ως στήλες)

ή γραφικά

$$\begin{array}{c} m \\ \boxed{X} \\ n \end{array} = \begin{array}{c} 1 \\ \boxed{t_1} \\ n \end{array} \begin{array}{c} m \\ \boxed{p'_1} \\ 1 \end{array} + \begin{array}{c} 1 \\ \boxed{t_2} \\ n \end{array} \begin{array}{c} m \\ \boxed{p'_2} \\ 1 \end{array} + \dots + \begin{array}{c} 1 \\ \boxed{t_r} \\ n \end{array} \begin{array}{c} m \\ \boxed{p'_r} \\ 1 \end{array}$$

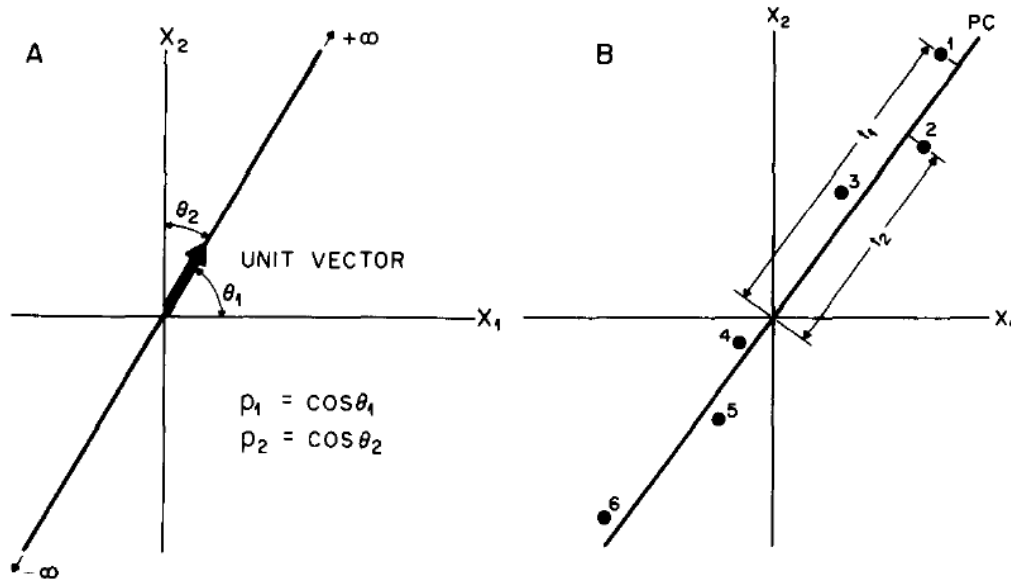
$$= \begin{array}{c} a \\ \boxed{T} \\ n \end{array} \begin{array}{c} m \\ \boxed{P'} \\ a \end{array}$$

Για να δειχθεί τι σημαίνει t_h και p'_h , ένα παράδειγμα για δύο μεταβλητές, στο δισδιάστατο επίπεδο, παρουσιάζεται το παρακάτω σχήμα.

Η επέκταση σε περισσότερες διαστάσεις είναι εύκολο, αλλά είναι δύσκολο να το δείξουμε στο χαρτί. Για παράδειγμα, στο Σχήμα 39 A, η κύρια συνιστώσα (the principal component) είναι η γραμμή της καλύτερης προσαρμογής (best fit) για τα σημεία δεδομένων που εμφανίζονται στο Σχήμα 39 B. Καλύτερη προσαρμογή σημαίνει ότι το άθροισμα των τετραγώνων των x_1 και x_2 των καταλοίπων ελαχιστοποιείται (the sum of squares of x_1 and x_2 residuals is minimized), αυτό είναι επίσης ο μέσος όρος των δύο γραμμών παλινδρόμησης (regression lines), Πηγαίνει από $-\infty$ έως $+\infty$, το p'_h 1×2 διάνυσμα

γραμμής. Τα στοιχεία του p_1 και p_2 είναι τα συνημίτονα κατεύθυνσης (direction cosines) ή οι προβολές ενός μοναδιαίου διανύσματος (unit vector) κατά μήκος της κύριας συνιστώσας στους άξονες του οικοπέδου.

Σχήμα 39, η κύρια συνιστώσα στην περίπτωση των δύο μεταβλητών: (A) loadings είναι τα συνημίτονα γωνίας του διανύσματος κατεύθυνσης, (B) scores είναι Οι προβολές των σημείων του δείγματος (1-6) στην διεύθυνση της κύριας συνιστώσας. Σημειώστε ότι τα δεδομένα είναι μέσο-επικεντρωμένα (mean-centered).[1]



Το διάνυσμα scores t_h είναι $n \times 1$ διάνυσμα στήλης, στοιχεία του είναι οι συντεταγμένες των αντίστοιχων σημείων στην γραμμή κύριας συνιστώσας Σχήμα 39 B, για αυτό το παράδειγμα, μπορεί εύκολα να γίνει κατανοητό γιατί κάποιος θέλει το μήκος

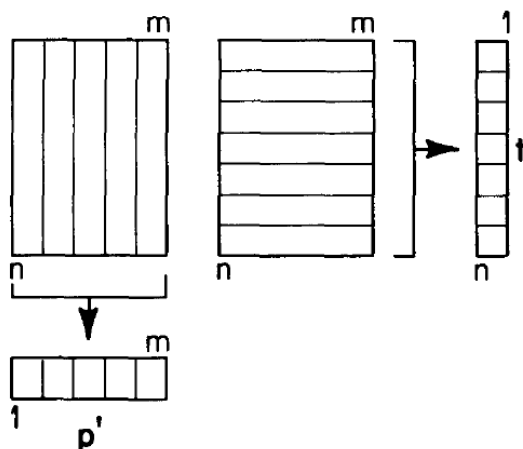
του p'_h να είναι ένα $[\cos(\theta_1)^2 + \cos(\theta_2)^2 = \cos(\theta_1)^2 + \sin(\theta_1)^2 = 1]$

παρόμοιοι κανόνες υπάρχουν για περισσότερες από δύο διαστάσεις.

Γενικά, αυτό που θέλουμε είναι ένας φορέας που προβάλλει τις στήλες του X σε μία μόνο διάσταση και ένας φορέας που προβάλλει τις γραμμές του X σε μία μόνο διάσταση (βλέπε Σχήμα 40). Στην πρώτη περίπτωση, κάθε στήλη του X εκπροσωπείται από ένα βαθμωτό, στη δεύτερη περίπτωση, κάθε γραμμή του X εκπροσωπείται από ένα βαθμωτό. Στο υπόλοιπο αυτής της ενότητας θα αποδειχθεί ότι οι εν λόγω φορείς είναι από έναν πολύ απλό χαρακτήρα.

Η Nonlinear iterative partial least squares (NIPALS) δεν υπολογίζει όλες τις κύριες συνιστώσες ταυτόχρονα. Υπολογίζει t_1 και p'_1 από τον πίνακα X , στη συνέχεια το εξωτερικό γινόμενο $t_1 p'_1$ αφαιρείται από το X και το κατάλοιπο E_1 (residual E_1) υπολογίζεται. Αυτό το κατάλοιπο E_1 μπορεί να χρησιμοποιηθεί για τον υπολογισμό t_2 και p'_2 :

Σχήμα 40, Scores και loadings λαμβάνονται με την προβολή X σε διανύσματα. *loadings* Κάθε στήλη του X προβάλλεται σε ένα στοιχείο του διανύσματος p' . *Scores* κάθε γραμμή του X προβάλλεται σε ένα στοιχείο του διανύσματος t . [1]



$$E_1 = X - t_1 p'_1 \quad E_2 = E_1 - t_2 p'_2 \quad \dots$$

$$E_h = E_{h-1} - t_h p'_h \quad \dots \quad E_{\tanh(X)} = 0 = E_{\tanh(X)-1} - t_{\tanh(X)} p'_{\tanh(X)} \quad (10)$$

Ο αλγόριθμος NIPALS έχει ως εξής:

$$1) \text{ λάβε ένα διάνυσμα } x_j \text{ από το } X \text{ και να το ονομάσουμε } t_h: t_h = x_j \quad (11)$$

$$2) \text{ υπολόγισε } p'_h: p'_h = t'_h X / t'_h t_h \quad (12)$$

$$3) \text{ ομαλοποίησε } p'_h \text{ σε μήκος 1: } p'_{h \text{ new}} = p'_{h \text{ old}} / \|p'_{h \text{ old}}\| \quad (13)$$

$$4) \text{ υπολόγισε } t_h: t_h = X p_h / p'_h p_h \quad (14)$$

5) κάνετε σύγκριση για το t_h που χρησιμοποιήθηκε στο βήμα 2 με εκείνο που λαμβάνεται στο βήμα 4, Αν είναι ίδια, να σταματήσει (η επανάληψη έχει συγκλίνει), Αν εξακολουθούν να διαφέρουν, προχωρήστε στο βήμα 2.

(Σημειώστε ότι μετά τον υπολογισμό της πρώτης συνιστώσας (first component), X στα βήματα 2 και 4 πρέπει να αντικατασταθεί από το κατάλοιπο του (its residual))

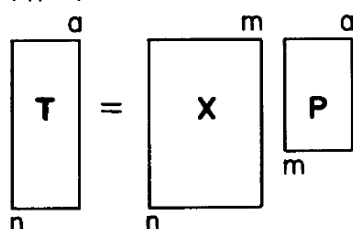
Μια εξήγηση για το πώς λειτουργεί NIPALS μπορεί να δει κανείς όταν κάποιος συνειδητοποιεί ότι τα $t'_h t_h$ της Εξ. 12, $\|p'_h\|$ της Εξ. 13 και $p'_h p_h$ της Εξ. 14 είναι βαθμωτά. Αυτές οι σταθερές βαθμωτών καλύτερα συνδυάζονται σε μία γενική σταθερά C . Στη συνέχεια, μπορεί κανείς να υποκαταστήσει την Εξ. 12 σε Εξ. 14: $t_h = X p_h$ και $p'_h = t'_h X$ δίνουν $C p'_h = (X p_h)' X$, ή $C p'_h = p'_h X' X$, ή $(C I_m - X' X) p_h = 0$, ή μπορεί κανείς να υποκαταστήσει Εξ. 14 σε Εξ. 12 και να έχει $(C' I_n - X X') t_h = 0$. Αυτές είναι οι εξισώσεις ιδιοτιμών / ιδιοδιανυσμάτων (eigenvalue/eigenvector) για $X' X$ και $X X'$ όπως χρησιμοποιούνται στον κλασικό υπολογισμό. I_n είναι μοναδιαίος πίνακας του μεγέθους $n \times n$, I_m είναι μοναδιαίος πίνακας του μεγέθους $m \times m$. Έχει αποδειχθεί ότι στη σύγκλιση, η λύση του NIPALS είναι η ίδια με εκείνη που υπολογίζεται από τον τύπο ιδιοδιανύσματος (eigenvector). Η μέθοδος NIPALS είναι βολική για μικροϋπολογιστές, είναι επίσης απαραίτητη για την καλή κατανόηση της PLS.

6.2.4 Παλινδρόμηση Κύριων Συνιστωσών

Τα αποτελέσματα από το τμήμα της PCA μπορούν να χρησιμοποιηθούν για να εξηγήσουμε τον μετασχηματισμό κύριας συνιστώσας ενός πίνακα δεδομένων X . Αυτή είναι μια αναπαράσταση του X ως scores του πίνακα T (όπου αποκλείονται οι διαστάσεις που έχουν μικρές ιδιοτιμές (eigenvalues)), Ο μετασχηματισμός είναι :

$$T = XP (= TP'P = TI_n) \quad (15)$$

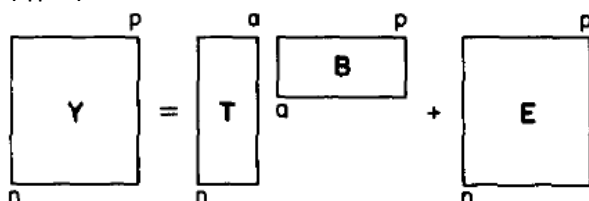
ή γραφικά:



Έτσι τώρα ο τύπος MLR μπορεί να γραφτεί ως

$$Y = TB + E \text{ (η λύση: } B = (T'T)^{-1}T'Y \text{)} \quad (16)$$

ή γραφικά:



Οι μεταβλητές του X αντικαθίστανται από νέες που έχουν καλύτερες ιδιότητες (ορθογωνιότητα) και εκτείνονται στο πολυδιάστατο χώρο της X , η αντιστροφή της $T'T$ δεν πρέπει να δώσει κανένα πρόβλημα, λόγω της αμοιβαίας ορθογωνιότητας των scores, διανύσματα των scores που αντιστοιχούν σε μικρές ιδιοτιμές μπορούν να αφεθούν έξω, προκειμένου να αποφευχθούν προβλήματα συγγραμμικότητας από το να επηρεάζουν τη λύση. Η PCR λύνει το πρόβλημα συγγραμμικότητας (εξασφαλίζοντας έναν αντιστρέψιμο πίνακα στον υπολογισμό του B) και η ικανότητα για την εξάλειψη των λιγότερων κύριων συνιστωσών επιτρέπει κάποια μείωση θορύβου (τυχαίο σφάλμα). Ωστόσο η PCR είναι μια μέθοδος δύο βημάτων και έτσι έχει τον κίνδυνο οι χρήσιμες (προγνωστικές) πληροφορίες να καταλήξουν στις κύριες συνιστώσες που απορρίφθηκαν και ότι κάποιος θόρυβος θα παραμείνει στις συνιστώσες που χρησιμοποιούνται για την παλινδρόμηση.

6.2.5 Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων (PARTIAL LEAST-SQUARES REGRESSION)

Το μοντέλο PLS είναι χτισμένο στις ιδιότητες του αλγορίθμου NIPALS. Όπως αναφέρθηκε στο τμήμα PCR, είναι δυνατόν να αφήσουμε τον πίνακα των scores T να αντιπροσωπεύει τον πίνακα δεδομένων. Ένα απλοποιημένο μοντέλο θα αποτελείται από μία παλινδρόμηση μεταξύ των scores για το X και Y μπλοκ. Το μοντέλο PLS μπορεί να θεωρηθεί ότι αποτελείται από εξωτερικές σχέσεις (X και Y τμήματα χωριστά) και μια εσωτερική σχέση (που συνδέει τα δύο τμήματα).

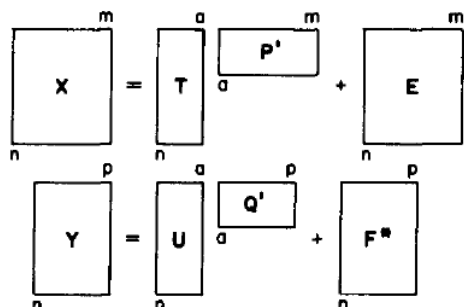
Η εξωτερική σχέση για το μπλοκ X (βλ. ενότητα PCA)

$$X = TP' + E = \sum t_h p'_h + E \quad (17)$$

Κάποιος μπορεί να κατασκευάσει την εξωτερική σχέση για το μπλοκ Y κατά τον ίδιο τρόπο:

$$Y = UQ' + F^* = \sum u_h q'_h + F^* \quad (18)$$

Γραφικά, Εξισώσεις 17 και 18 μπορούν να εμφανίζονται ως

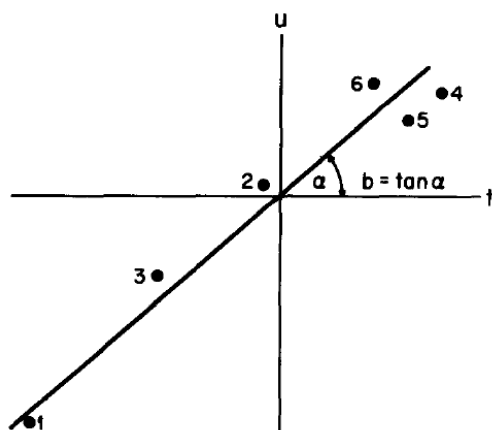


Τα αθροίσματα είναι από 1 έως α. Κάποιος μπορεί να περιγράψει όλες τις συνιστώσες και έτσι δίνει $E = F^* = 0$ ή όχι. Πώς και γιατί γίνεται αυτό συζητείται παρακάτω. Η πρόθεση είναι να περιγραφεί η Y όσο το δυνατό καλύτερα και ως εκ τούτου να γίνει το $\|F^*\|$ όσο το δυνατόν μικρότερο και την ίδια στιγμή να πάρουμε μια χρήσιμη σχέση μεταξύ X και Y. Η εσωτερική σχέση μπορεί να γίνει κοιτάζοντας το γράφημα score u του μπλοκ Y ενάντια το score t του μπλοκ X για κάθε συνιστώσα (Σχ. 41). Το απλούστερο μοντέλο για αυτή τη σχέση είναι μία γραμμική σχέση :

$$\tilde{u}_h = b_h t_h \quad (19)$$

Όπου $\tilde{u}_h$ είναι ο εκτιμητής του u_h

Σχήμα 41 η εσωτερική σχέση. Μια γραμμική παλινδρόμηση του u ενάντια t. Σημειώστε ότι τα δεδομένα είναι μέσο-επικεντρωμένα (mean-centered).[1]



Όπου $b_h = u'_h t_h / t'_h t_h$. Το b_h παίζει το ρόλο των συντελεστών παλινδρόμησης b, στα μοντέλα MLR και PCR.

Αυτό το μοντέλο, ωστόσο, δεν είναι το καλύτερο δυνατό. Ο λόγος είναι ότι οι κύριες συνιστώσες υπολογίζονται για τα δύο μπλοκ ξεχωριστά με αποτέλεσμα να έχουν μια αδύναμη σχέση μεταξύ τους. Θα ήταν καλύτερα να τους δώσουμε πληροφορίες για καθ'

ένα από αυτά έτσι ώστε να περιστραφεί ελαφρώς το αποτέλεσμα συνιστώσων που βρίσκονται πιο κοντά στη γραμμή παλινδρόμησης του Σχ. 41.

Το υπερ- απλοποιημένο μοντέλο(Over-simplified model): 2xPCA. Ένα υπερ-απλουστευμένο μοντέλο μπορεί να γραφτεί στην αλγοριθμική μορφή όπως στην ενότητα NIPALS,

Για το μπλοκ X:

$$(1) \text{παρε } t_{\text{start}} = x_j,$$

$$(2) p' = t'X/t't (= u'X/u'u),$$

$$(3) p'_{\text{new}} = p'_{\text{old}} / \|p'_{\text{old}}\|,$$

$$(4) t = Xp/p'p,$$

(5) κάνετε σύγκριση για το t που χρησιμοποιήθηκε στο βήμα 2 με εκείνο που λαμβάνεται στο βήμα 4, Αν είναι ίδια, να σταματήσει, Αν εξακολουθούν να διαφέρουν, προχωρήστε στο βήμα 2.

Για το μπλοκ Y:

$$(1) \text{παρε } u_{\text{start}} = y_j,$$

$$(2) q' = u'Y/u'u (= t'Y/t't),$$

$$(3) q'_{\text{new}} = q'_{\text{old}} / \|q'_{\text{old}}\|,$$

$$(4) u = Yq/q'q,$$

(5) κάνετε σύγκριση για το u που χρησιμοποιήθηκε στο βήμα 2 με εκείνο που λαμβάνεται στο βήμα 4, Αν είναι ίδια, να σταματήσει, Αν εξακολουθούν να διαφέρουν, προχωρήστε στο βήμα 2.

Βελτίωση της εσωτερικής σχέσεως: ανταλλαγή των scores. Οι παραπάνω σχέσεις γράφονται ως εντελώς ξεχωριστά. Ο τρόπος που το καθένα μπορεί να πάρει πληροφορίες για το άλλο είναι να αφήσουμε t και u να αλλάξουν θέση στο βήμα 2. (βλέπε τα τμήματα στις παρενθέσεις σε αυτό το βήμα). Έτσι, οι δύο αλγόριθμοι μπορεί να γραφτούν σε

αλληλουχία:

$$(1) \text{παρε } u_{\text{start}} = y_j,$$

$$(2) p' = u'X/u'u (w' = u'X/u'u),$$

$$(3) p'_{\text{new}} = p'_{\text{old}} / \|p'_{\text{old}}\| (w'_{\text{new}} = w'_{\text{old}} / \|w'_{\text{old}}\|),$$

$$(4) t = Xp/p'p (t = Xw/w'w),$$

$$(5) q' = t'Y/t't,$$

$$(6) q'_{\text{new}} = q'_{\text{old}} / \|q'_{\text{old}}\|,$$

$$(7) u = Yq/q'q,$$

(8) Συγκρίνετε το t στο βήμα 4 με το t στο προηγούμενο στάδιο επανάληψης. Αν είναι ίσα (εντός ορισμένου σφάλματος στρογγυλοποίησης) να σταματήσει, αλλιώς πάμε σε βήμα 2, (Στην περίπτωση για την οποία το μπλοκ Y έχει μόνο μία μεταβλητή, τα βήματα 5-8 μπορούν να παραλειφθούν με την τοποθέτηση $q = 1$.) . Αυτός ο αλγόριθμος συνήθως συγκλίνει πολύ γρήγορα για να δώσει περιστρεφόμενες συνιστώσες για X και Y μπλοκ.

Η απόκτηση ορθογώνιων scores μπλοκ X : Υπάρχει ακόμα ένα πρόβλημα, ο αλγόριθμος δεν δίνει ορθογώνιες t τιμές, ο λόγος είναι ότι η σειρά των υπολογισμών που είχε χρησιμοποιηθεί για την PCA έχει αλλάξει. Επομένως, τα p' αντικαθίστανται από τα w' (βλέπε τύπους σε παρένθεση στην προηγούμενη υποενότητα). Ένας επιπλέον βρόχος μπορεί να συμπεριληφθεί μετά από τη σύγκλιση για να πάρουμε ορθογώνιες τιμές του t :

$$p' = t'X/t't \quad (20)$$

με $p'_{\text{new}} = p'_{\text{old}} / \|p'_{\text{old}}\|$. Καθίσταται δυνατόν να υπολογιστεί η νέα t : $t = Xp/p'p$, αλλά αυτό αποδεικνύεται ότι είναι απλώς ένας πολλαπλασιασμός βαθμωτού με το νόρμα του p' στην Εξ. 20: $t_{\text{new}} = t_{\text{old}} \|p'_{\text{old}}\|$. Οι ορθογώνιες τιμές t δεν είναι απολύτως απαραίτητες, αλλά κάνουν τη σύγκριση με PCR ευκολότερη. Κάποιος πρέπει να δώσει την ίδια διαδικασία επανακλιμακώσεως με τα βάρη W' . Αν η πρόβλεψη είναι να γίνει χωρίς σφάλμα: $w'_{\text{new}} = w'_{\text{old}} \|p'_{\text{old}}\|$, τότε t μπορεί να χρησιμοποιηθεί για την εσωτερική σχέση όπως στην Εξίσωση. 19, και τα κατάλοιπα(residuals) μπορούν να υπολογιστούν από $E_1 = X - t_1 p'_{11}$ και $F_1^* = Y - u_1 q'_{11}$,

Γενικά ισχύει:

$$E_h = E_{h-1} - t_h p'_{h-1} \quad ; \quad X = E_0 \quad (21)$$

$$F_h^* = F_{h-1}^* - u_h q'_{h-1} \quad ; \quad Y = F_0 \quad (22)$$

Αλλά στην εξωτερική σχέση για το μπλοκ Y , u_h αντικαθίσταται από τον εκτιμητή του,

$\tilde{u}_h = b_h t_h$, και μία μικτή σχέση επιτυγχάνεται:

$$F_h = F_{h-1} - b_h t_h q'_{h-1} \quad (23)$$

Υπενθυμίζεται ότι ο στόχος είναι να κάνουμε το $\|F_h\|$ μικρό, αυτή η μικτή σχέση εξασφαλίζει την ικανότητα να χρησιμοποιεί τις παραμέτρους του μοντέλου για την πρόβλεψη από ένα σύνολο δοκιμής. Εξάλλου, επειδή ο βαθμός του Y (Y rank) δεν μειώνεται κατά 1 για κάθε συνιστώσα, κάποιος μπορεί να συνεχιστεί μέχρι ο βαθμός του μπλοκ X να έχει εξαντληθεί.

Περίληψη: PLS

Υπάρχουν εξωτερικές σχέσεις της μορφής $X = TP' + E$ και $Y = UQ' + F^*$.

Υπάρχει μια εσωτερική σχέση $\tilde{u}_h = b_h t_h$.

Η μικτή σχέση είναι $Y = TBQ' + F$ όπου $\|F\|$ είναι προς ελαχιστοποίηση.

Στον επαναληπτικό αλγόριθμο, τα μπλοκ ανταλλάσσουν τα scores, αυτό δίνει μια καλύτερη εσωτερική σχέση.

Για να αποκτήσουμε ορθογώνια scores του X , όπως στο PCA, είναι απαραίτητο να εισαχθούν βάρη.

Ιδιότητες των παραγόντων PLS

Για τον χρήστη PLS, είναι προφανώς ενδιαφέρον να γνωρίζουμε τι είδους ιδιότητες να περιμένουμε από αυτήν. Οι βασικές ιδιότητες μπορούν να συνοψιστούν ως εξής:

Οι ποσότητες p'_h και q'_h έχουν μήκος μονάδας για κάθε h : $\|p'_h\| = \|q'_h\| = 1$.

Οι δύο t_h και u_h είναι επικεντρωμένα γύρω από το μηδέν για κάθε h : $\sum t_{hj} = 0$ και $\sum u_{hj} = 0$ για $j = 1 \dots n$.

Το w' είναι ορθογώνιο: $w'_i w'_j = \delta_{ij} \|w'_i\|^2$ όπου δ_{ij} είναι το Kronecker delta, $\delta_{ij} = 1$ αν $i = j$ και $\delta_{ij} = 0$ αν $i \neq j$.

Το t_h είναι ορθογώνιο: $t'_i t'_j = \delta_{ij} \|t_i\|^2$.

6.2.6 Η πρόβλεψη

Το σημαντικό μέρος της κάθε παλινδρόμησης είναι η χρήση της στην πρόβλεψη του εξαρτημένου μπλοκ από το ανεξάρτητο μπλοκ. Αυτό γίνεται με την αποσύνθεση του μπλοκ X και την κατασκευή του μπλοκ Y . Για το σκοπό αυτό, p' , q' , w' και b από το τμήμα

βαθμονόμησης αποθηκεύονται για κάθε παράγοντα PLS. Θα πρέπει να σημειωθεί ότι το νέο μπλοκ X έχει r δείγματα αντί n , αποσύνθεση του μπλοκ X και κατασκευή του μπλοκ Y , για το μπλοκ X , t εκτιμάται πολλαπλασιάζοντας X με το w , όπως στο τμήμα κατασκευής του μοντέλου.

$$\hat{t}_h = E_h - 1 w_h \quad (24)$$

$$E_h = E_{h-1} - \hat{t}_h p'_h \quad (25)$$

Για το μπλοκ Y :

$$Y = F_h = \sum b_h \hat{t}_h q'_h \quad (26)$$

όπου το άθροισμα είναι πάνω στο h για όλους τους παράγοντες (α) κάποιος θέλει να συμπεριλάβει και $X = E_0$, $Y = F_\alpha$. Ο πλήρης αλγόριθμος PLS δίνεται στο Σχήμα 44.

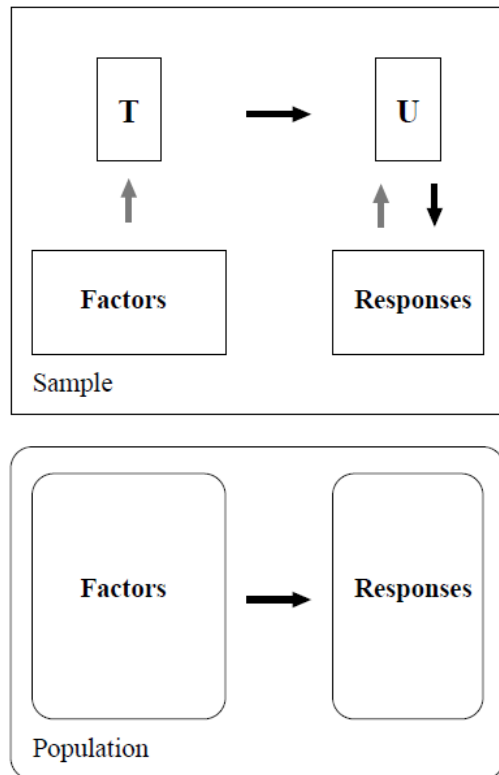
6.3 Πώς λειτουργεί η PLS :

Κατ' αρχήν, η MLR μπορεί να χρησιμοποιηθεί με πάρα πολλούς παράγοντες. Ωστόσο, εάν ο αριθμός των παραγόντων γίνει πάρα πολύ μεγάλος (για παράδειγμα, μεγαλύτερος από τον αριθμό των δειγμάτων), είναι πιθανό να έχει ένα μοντέλο που ταιριάζει στα δεδομένα δειγμάτων τέλεια αλλά θα αποτύχει να προβλέψει νέα δεδομένα καλά. Αυτό το φαινόμενο ονομάζεται over-fitting.

Σε τέτοιες περιπτώσεις, αν και υπάρχουν πολλοί παράγοντες πρόδηλοι, μπορεί να υπάρχουν λίγοι μόνο υποκείμενοι ή λανθάνουσι παράγοντες (latent factors) που αντιπροσωπεύουν το μεγαλύτερο μέρος της διακυμάνσεως στην απόκριση (of the variation in the response). Η γενική ιδέα της PLS είναι να προσπαθήσουμε όσο το δυνατόν να εξαγάγουμε αυτούς λανθάνοντες παράγοντες που ευθύνονται για το μεγαλύτερο μέρος της διακύμανσης πρόδηλου παράγοντα (of the manifest factor variation) καθώς μοντελοποιούμε καλά τις αποκρίσεις.

Για το λόγο αυτό, το αρκτικόλεξο PLS έχει ληφθεί επίσης για να σημάνει « `` projection to latent structure `` προβολή στη δομή λανθάνουσας ». Θα πρέπει να σημειωθεί, ωστόσο, ότι ο όρος « λανθάνουσα » δεν έχει την ίδια τεχνική έννοια στο πλαίσιο της PLS όπως κάνει για άλλες πολυμεταβλητές τεχνικές. Ειδικότερα, PLS δεν αποφέρει συνεπείς εκτιμήσεις των λεγόμενων « λανθάνουσες μεταβλητές » στα επίσημα δομικά μοντέλα εξισώσεων. Το επόμενο σχήμα δίνει μία σχηματική περιγραφή της μεθόδου. Ο γενικός στόχος (που δείχνεται στο κάτω πλαίσιο) είναι η χρήση των παραγόντων για την πρόβλεψη των αποκρίσεων στον πληθυσμό.

Σχήμα 42, σχηματική περιγραφή της μεθόδου PLS.[10]



Indirect modeling

Αυτό επιτυγχάνεται έμμεσα με εκχύλιση λανθανουσών μεταβλητών T και U από παράγοντες και αποκρίσεις του δείγματος αντίστοιχα. Οι εξαγόμενοι παράγοντες T (που αναφέρεται επίσης ως X -scores) χρησιμοποιούνται για την πρόβλεψη των U (Y -scores) και κατόπιν οι προβλεπόμενοι Y -scores χρησιμοποιούνται για την κατασκευή προβλέψεων για τις αποκρίσεις. Η διαδικασία αυτή στην πραγματικότητα καλύπτει διάφορες τεχνικές ανάλογα με την ποια πηγή της παραλλαγής (of variation) θεωρείται η πιο κρίσιμη.

Principal Components Regression (PCR):

Οι X -scores επιλέγονται να εξηγήσουν όσο το δυνατόν την διακύμανση παράγοντα (factor variation). Αυτή η προσέγγιση αποδίδει πληροφοριακές κατευθύνσεις στο χώρο παράγοντα, αλλά δεν μπορούν να σχετίζονται με το σχήμα της προβλεπόμενης επιφάνειας.

Maximum Redundancy Analysis (MRA) (van den Wollenberg 1977):

Οι Y -score επιλέγονται να εξηγήσουν όσο το δυνατόν τη διακύμανση προβλεπόμενου Y , η προσέγγιση αυτή επιδιώκει κατευθύνσεις στο χώρο παράγοντα που σχετίζονται με την περισσότερη διακύμανση των αποκρίσεων (with the most variation in the responses), αλλά οι προβλέψεις δεν μπορούν να είναι πολύ ακριβείς.

Partial Least Squares:

Το X -και Y -scores επιλέγονται έτσι ώστε η σχέση μεταξύ των διαδοχικών ζευγών των scores είναι όσο το δυνατόν ισχυρή. Αυτό είναι σαν μια ισχυρή μορφή ανάλυσης πλεονασμού, αναζητώντας τις κατευθύνσεις στο χώρο παράγοντα που σχετίζονται με την υψηλή

διακύμανση των αποκρίσεων, αλλά προδιαθέτοντας τους προς τις κατευθύνσεις που έχουν προβλεφθεί με ακρίβεια.

Εάν ο αριθμός των εξαγόμενων παράγοντων είναι μεγαλύτερο από ή ίσο με τη τάξη μεγέθους του χώρου παράγοντα του δείγματος (greater than or equal to the rank of the sample factor space), τότε PLS είναι ισοδύναμη με MLR. Ένα σημαντικό χαρακτηριστικό της μεθόδου είναι ότι συνήθως απαιτούνται πολύ λιγότεροι παράγοντες.

Ο ακριβής αριθμός των εξαγόμενων παράγοντων συνήθως επιλέγεται από κάποια ευρετική τεχνική με βάση την ποσότητα της μεταβλητότητας καταλοίπου (the amount of residual variation). Μια άλλη προσέγγιση είναι να κατασκευαστεί το μοντέλο PLS για ένα δεδομένο αριθμό παραγόντων σε ένα σύνολο δεδομένων και στη συνέχεια να δοκιμαστεί σε ένα άλλο, επιλέγοντας τον αριθμό των εξαγόμενων παράγοντων για τον οποίο το συνολικό σφάλμα πρόβλεψης ελαχιστοποιείται. Εναλλακτικά, ο van der Voet (1994) προτείνει την επιλογή του λιγότερο αριθμού εξαγόμενων παράγοντων των οποίων τα κατάλοιπα (residuals) δεν είναι σημαντικά μεγαλύτερα από εκείνα του μοντέλου με ελάχιστο σφάλμα. Εάν δεν είναι διαθέσιμο ένα βολικό σετ δοκιμής (test set) τότε κάθε δείγμα μπορεί να χρησιμοποιηθεί με τη σειρά ως μια δοκιμή (test set). Αυτό είναι γνωστό ως cross-validation.

6.4 Βήματα Βαθμονόμησης (κατάρτισης) και πρόβλεψης (δοκιμής) (Calibration (training) and prediction (test) steps)

Χημική ανάλυση συνήθως αποτελείται από δύο βήματα. Πρώτον, τα χαρακτηριστικά μιας μεθόδου ή οργάνου διερευνώνται και γίνεται προσπάθεια να βρούμε ένα μοντέλο για τη συμπεριφορά του (ένα μοντέλο είναι μια σχέση $y = f(X)$ μεταξύ των δύο ομάδων μεταβλητών, που ονομάζονται συχνά εξαρτώμενες Y και ανεξάρτητες X). Αυτό είναι το βήμα της βαθμονόμησης ή της κατάρτισης. Τα δεδομένα που χρησιμοποιούνται για αυτό το βήμα ονομάζονται calibration or training set, οι παράμετροι του μοντέλου ονομάζονται συντελεστές παλινδρόμησης ή ευαισθησίες (regression coefficients or sensitivities). Το δεύτερο βήμα είναι το ένα στο οποίο οι ανεξάρτητες μεταβλητές λαμβάνονται για ένα ή περισσότερα δείγματα, αυτά χρησιμοποιούνται μαζί με τις ευαισθησίες (sensitivities) για να προβλεφθούν οι τιμές για τις εξαρτημένες μεταβλητές, αυτό είναι το βήμα δοκιμής ή πρόβλεψης, τα δεδομένα που χρησιμοποιούνται σε αυτό το βήμα είναι the prediction or test set.

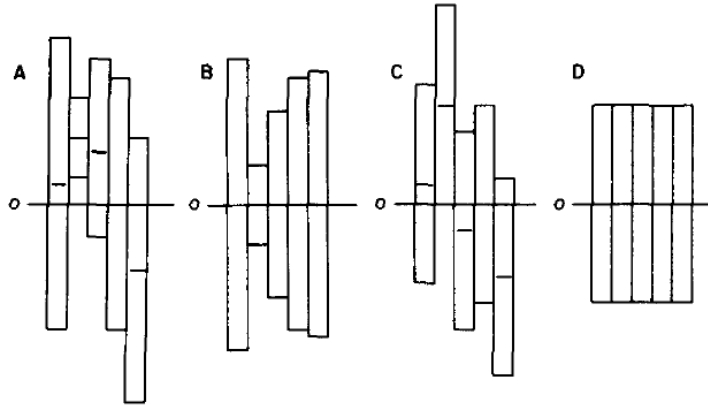
6.5 Κεντράρισμα και η κλιμάκωση των μεταβλητών (Mean-centering and scaling of variables)

Πριν αναπτυχθεί το μοντέλο είναι βολικό να προσαρμόσουμε τα δεδομένα στο σύνολο βαθμονόμησης (calibration set) ώστε να γίνουν ευκολότεροι οι υπολογισμοί. Για την ευκολία της εξήγησης, οι τιμές για κάθε μεταβλητή χρησιμοποιούνται στη mean-centered μορφή, η μέση τιμή για κάθε μεταβλητή υπολογίζεται από το σύνολο βαθμονόμησης και στη συνέχεια αφαιρείται από κάθε αντίστοιχη μεταβλητή. Η διαδικασία εφαρμόζεται και στο test set.

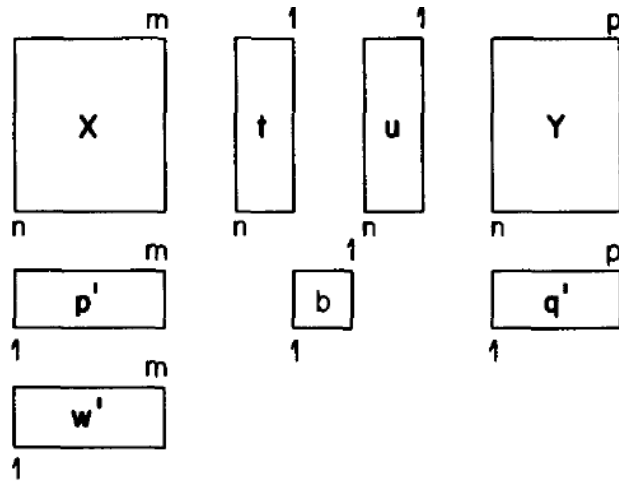
Η κλιμάκωση επιτυγχάνεται διαιρώντας όλες τις τιμές για μια ορισμένη μεταβλητή από την τυπική απόκλιση (standard deviation) για την εν λόγω μεταβλητή, έτσι ώστε η διακύμανση (variance) για κάθε μεταβλητή είναι ενότητα (unity).

η επίδραση των mean-centering and scaling of variables στα δεδομένα φαίνεται στο επόμενο σχήμα.

Σχήμα 43, Προεπεξεργασία δεδομένων. Τα δεδομένα για κάθε μεταβλητή αντιπροσωπεύονται από μια μπάρα διακύμανσης και το κέντρο της. (A) Τα περισσότερα ανεπεξέργαστα δεδομένα μοιάζουν με αυτό. (B) Το αποτέλεσμα μετά από κεντράρισμα μέσης (mean-centering) μόνο. (C) Το αποτέλεσμα μετά από κλιμάκωση διακύμανσης (variance-scaling) μόνο. (D) Το αποτέλεσμα μετά από κεντράρισμα μέσης και την κλιμάκωση διακύμανσης.[1]



Σχήμα 44, ο αλγόριθμος της PLS με την γραφική αναπαράσταση των πινάκων και διανυσμάτων που χρησιμοποιούνται σε αυτόν.[1]



The PLS algorithm

It is assumed that X and Y are mean-centered and scaled:

For each component: (1) take $u_{\text{start}} = \text{some } y_j$.

In the X block: (2) $w' = u'X/u'u$

$$(3) w'_{\text{new}} = w'_{\text{old}} / \|w'_{\text{old}}\| \text{ (normalization)}$$

$$(4) t = Xw/w'w$$

In the Y block: (5) $q' = t'Y/t't$

$$(6) q'_{\text{new}} = q'_{\text{old}} / \|q'_{\text{old}}\| \text{ (normalization)}$$

$$(7) u = Yq/q'q$$

Check convergence: (8) compare the t in step 4 with the one from the preceding iteration. If they are equal (within a certain rounding error) go to step 9, else go to step 2. (If the Y block has only one variable, steps 5–8 can be omitted by putting $q = 1$, and no more iteration is necessary.)

Calculate the X loadings and rescale the scores and weights accordingly:

$$(9) p' = t'X/t't$$

$$(10) p'_{\text{new}} = p'_{\text{old}} / \|p'_{\text{old}}\| \text{ (normalization)}$$

$$(11) t_{\text{new}} = t_{\text{old}} \|p'_{\text{old}}\|$$

$$(12) w'_{\text{new}} = w'_{\text{old}} \|p'_{\text{old}}\|$$

(p' , q' and w' should be saved for prediction; t and u can be saved for diagnostic and/or classification purposes).

Find the regression coefficient b for the inner relation:

$$(13) b = u't/t't$$

Calculation of the residuals. The general outer relation for the X block (for component h) is

$$E_h = E_{h-1} - t_h p'_h; X = E_0$$

The mixed relation for the Y block (for component h) is

$$F_h = F_{h-1} - b_h t_h q'_h; Y = F_0$$

From here, one goes to Step 1 to implement the procedure for the next component. (Note: After the first component, X in steps 2, 4 and 9 and Y in steps 5 and 7 are replaced by their corresponding residual matrices E_h and F_h .)

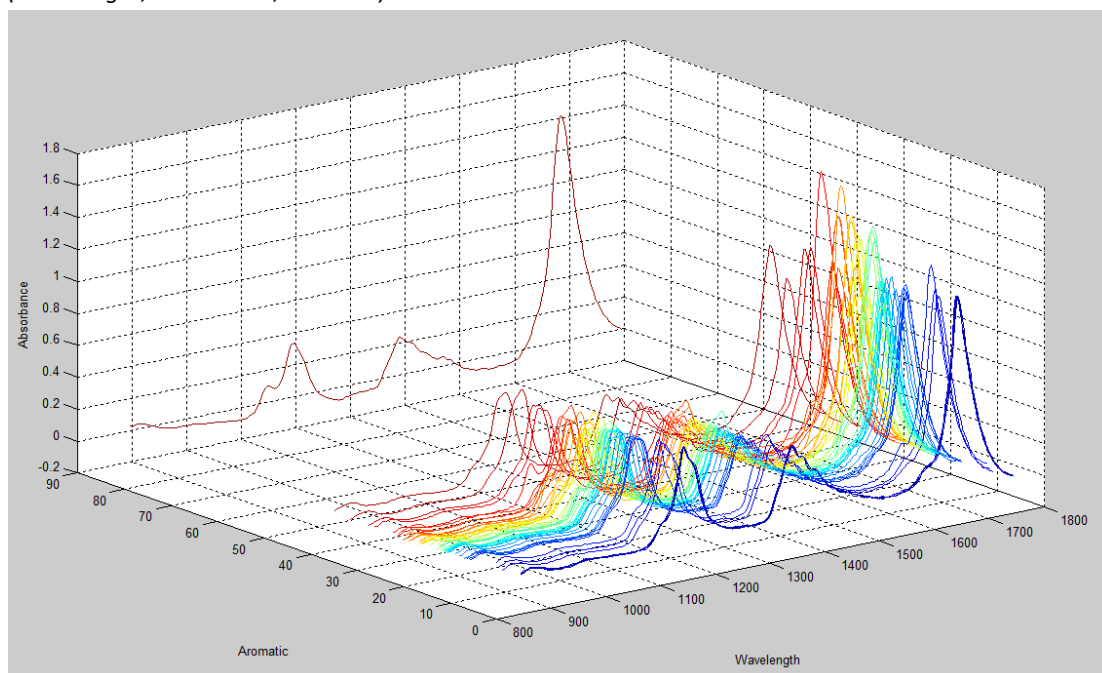
7. Ανάπτυξη του προτεινόμενου μοντέλου

7.1 Εισαγωγή

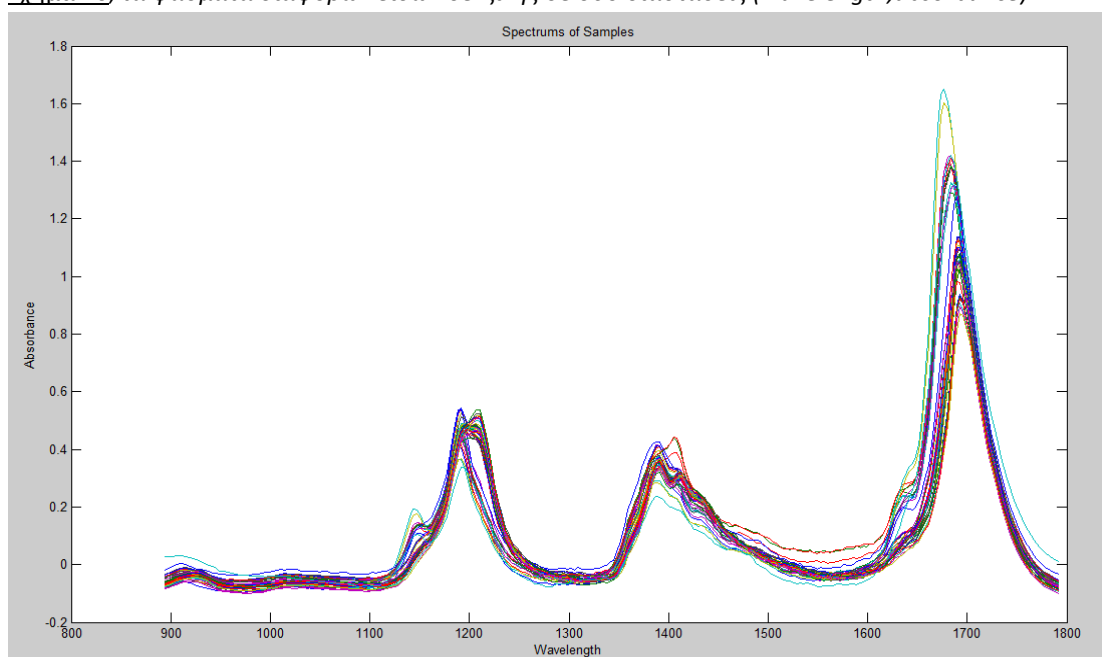
το προτεινόμενο μοντέλο πρόβλεψης ποσοτήτων του αρωματικού περιεχομένου (Aromatic Content) σε διάφορα είδη δειγμάτων βενζίνης βασίζεται στις μεθόδους Cross-Validation, Interactive Variable Selection for PLS, Partial least squares (PLS) χρησιμοποιώντας τα 47 φάσματα βενζίνης που έχουμε πάρει στο εργαστήριο.

Τα φάσματα διάφορων ειδών βενζίνης με τις τιμές Aromatics τους σε αύξουσα σειρά φαίνονται στο παρακάτω σχήμα.

Σχήμα 45, τα φάσματα διαφόρων ειδών βενζίνης με τις τιμές Aromatics τους σε τρεις διαστάσεις (wavelength, absorbance, aromatic).



Σχήμα 46, τα φάσματα διάφορων ειδών βενζίνης σε δύο διαστάσεις (wavelength, absorbance).



7.2 Ξεκίνημα της διαδικασίας ανάπτυξης και εξέταση διαφόρων περιπτώσεων

Με τη βοήθεια τον παραπάνω PLS αλγόριθμο υλοποίησα μια συνάρτηση plsDip η οποία περιέχει αλγόριθμο PLS όταν το Y (dependent block) έχει μόνο μια μεταβλητή (στη περίπτωση μας είναι η τιμή του Aromatic) η συνάρτηση δέχεται σαν είσοδο το calibration or training set το οποίο αποτελείται από δύο πίνακες ,τον πίνακα X των φασμάτων βενζίνης και τον πίνακα Y με τις αντίστοιχες τιμές Aromatics τών φασμάτων βενζίνης (πρίν λάβει η συνάρτηση τη συγκεκριμένη είσοδο εφαρμόζεται Mean-centering and scaling of variables), η συνάρτηση επιστρέφει σαν έξοδο τον πίνακα w τον πίνακα p και την παράμετρο b. Τα στοιχεία της εξόδο της συνάρτησης (w,b) ή (p,b) ή (w) ή (p) μαζί με ένα φάσμα βενζίνης(ανεξάρτητες μεταβλητές) χρησιμοποιούνται για να μαντέψουμε μια τιμή Aromatic (εξαρτημένη μεταβλητή) για το αντίστοιχο φάσμα test set, για την ανάπτυξη του μοντέλου και την εξέταση διαφόρων περιπτώσεων έγραψα έναν αλγόριθμο σε ένα Matlab αρχείο με όνομα seeAll . Αυτός ο αλγόριθμος με τη βοήθεια του παραπάνω κώδικα της συνάρτησης plsDip δημιουργούν δύο μεγάλους πίνακες που χρησιμεύουν στη ανάπτυξη του μοντέλου. Ο αλγόριθμος δουλεύει με τον εξής τρόπο :

Αφαιρεί ένα δείγμα από τα 47 δείγματα για να είναι το test set και τα υπόλοιπα 46 δείγματα μένουν σαν training set, ο κώδικας της συνάρτησης plsDip τρέχει 500 φορές (500 components), για κάθε component συλλέγονται πληροφορίες για τα errors των προβλέψεων και τα norm(t-t1) .Ο αλγόριθμος τερματίζει όταν καθ'ένα από τα 47 δείγματα μπαίνει στο test set μία φορά .

Χρησιμοποιώντας τον παραπάνω αλγόριθμο θα μάθουμε ποίο στοιχείο από αυτά της εξόδο της συνάρτησης plsDip (w,b) ή (p,b) ή (w) ή (p) είναι περισσότερο κατάλληλο για τη πρόβλεψη.

Ο αλγόριθμος δημιουργεί τον πίνακα ooo2 που έχει διαστάσεις 47 x 500 και είναι γεμάτος με τα errors τών προβλέψεων για κάθε component και δείγμα .

Μαζική εξέταση ανά component

Εξετάζουμε πρώτα το (w,b)

*****matlab

$Sss = [\sum_{i=1}^{47} ooo2(i,j)]$ για j =1.....500

ο πίνακας Sss έχει διαστάσεις 1 x 500

theMinimum = min(Sss)

theMinimum = 343.1660

Το αποτέλεσμα στο matlab μας λέει ότι το καλύτερο σύνολο λαθών είναι στη θέση 4 (component 4) με τιμή 343.1660

Ας εξετάσουμε τώρα το (p,b)

*****matlab

$Sss = [\sum_{i=1}^{47} ooo2(i,j)]$ για $j=1.....500$

ο πίνακας **Sss** έχει διαστάσεις 1 x 500

theMinimum = min(**Sss**)

theMinimum = 351.1575

Το αποτέλεσμα στο matlab μας λέει ότι το καλύτερο σύνολο λαθών είναι στη θέση 4 (component 4) με τιμή 351.1575

Ας εξετάσουμε τώρα το (w)

*****matlab

$Sss = [\sum_{i=1}^{47} ooo2(i,j)]$ για $j=1.....500$

ο πίνακας **Sss** έχει διαστάσεις 1 x 500

theMinimum = min(**Sss**)

theMinimum = 366.2169

Το αποτέλεσμα στο matlab μας λέει ότι το καλύτερο σύνολο λαθών είναι στη θέση 15 (component 15) με τιμή 366.2169

Ας εξετάσουμε τώρα το (p)

*****matlab

$Sss = [\sum_{i=1}^{47} ooo2(i,j)]$ για $j=1.....500$

ο πίνακας **Sss** έχει διαστάσεις 1 x 500

theMinimum = min(**Sss**)

theMinimum = 376.5664

Το αποτέλεσμα στο matlab μας λέει ότι το καλύτερο σύνολο λαθών είναι στη θέση 15 (component 15) με τιμή 376.5664

Από τα παραπάνω προκύπτει ότι το καλύτερο για πρόβλεψη στη μαζική εξέταση ανά component είναι το (w,b)

Εξέταση ανά δείγμα

Στην εξέταση ανά δείγμα ψάχνουμε για το λιγότερο λάθος πρόβλεψης για κάθε δείγμα ανάμεσα στα 500 components και μετά αθροίζουμε αυτά τα λάθη για όλα τα δείγματα .

Εξετάζουμε πρώτα το (w,b)

```
*****matlab
Sss =  $\sum_{i=1}^{47} (\min[\text{ooo2}(i,j)] \text{ για } i = 1 \dots 47)$ 
```

Sss = 167.9109

Το αποτέλεσμα στο matlab μας λέει ότι το καλύτερο σύνολο λαθών είναι 167.9109

Ας εξετάσουμε τώρα το (p,b)

```
*****matlab
Sss =  $\sum_{i=1}^{47} (\min[\text{ooo2}(i,j)] \text{ για } i = 1 \dots 47)$ 
```

Sss = 155.5494

Το αποτέλεσμα στο matlab μας λέει ότι το καλύτερο σύνολο λαθών είναι 155.5494

Ας εξετάσουμε τώρα το (w)

```
*****matlab
Sss =  $\sum_{i=1}^{47} (\min[\text{ooo2}(i,j)] \text{ για } i = 1 \dots 47)$ 
```

Sss = 5.5686

Το αποτέλεσμα στο matlab μας λέει ότι το καλύτερο σύνολο λαθών είναι 5.5686

Ας εξετάσουμε τώρα το (p)

```
*****matlab
Sss =  $\sum_{i=1}^{47} (\min[\text{ooo2}(i,j)] \text{ για } i = 1 \dots 47)$ 
```

Sss = 7.9222

Το αποτέλεσμα στο matlab μας λέει ότι το καλύτερο σύνολο λαθών είναι 7.9222

Από τα παραπάνω προκύπτει ότι το καλύτερο για πρόβλεψη στην εξέταση ανά δείγμα είναι το (w)

Από τα παραπάνω αποτελέσματα αποφασίζω να χρησιμοποιήσω το (w) στο τελικό μοντέλο ενώ το (w,b) θα το χρησιμοποιήσω με τη μέθοδο βελτίωσης προβλέψης και σβήσιμο μεταβλητών INTERACTIVE VARIABLE SELECTION.

7.3 INTERACTIVE VARIABLE SELECTION (IVS) FOR PLS

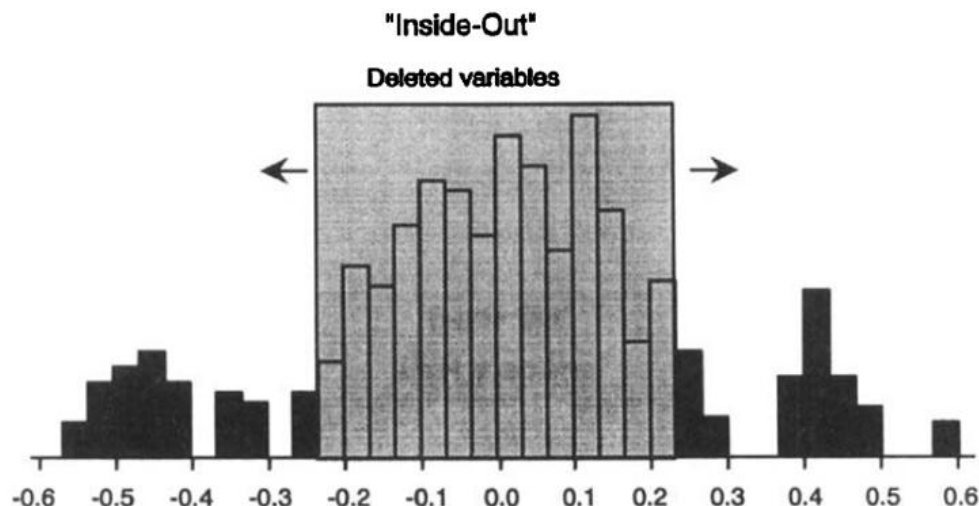
Η IVS τροποποιεί τον πίνακα w σβήνοντας μεταβλητές και χρησιμοποιεί τη μέθοδο cross-validation για να μάθουμε ποια είναι η καλύτερη τροποποίηση, υπάρχουν δύο τρόποι για να διαλέξουμε ποιές μεταβλητές να σβήσουμε. Οι δύο τρόποι αυτοί είναι οι Inside-Out και Outside-In

Inside-Out

Δίνεται το όριο ($0 < \alpha < 1$), το α μετακινείται από 0 στο 1 κάθε φορά κατά 0.01 και για κάθε τιμή του α θα φτιάχνουμε έναν καινούριο πίνακα w σβήνοντας τις τιμές του w οι οποίες είναι στην απόλυτη τιμή μικρότερη από α . Δηλαδή:

αν $\text{abs}(w(i)) < \alpha \Rightarrow w(i)=0$ όπου $i=1,2,\dots,K$ και K το σύνολο των μεταβλητών του w.

Σχήμα 47, διαγραφή των μεταβλητών με τον τρόπο Inside-Out .[2]



Για κάθε τροποποίηση του w υπολογίζουμε μία καινούρια τιμή για το CV (cross-validation) όπου $CV = PRESS/RSS$, $PRESS = \text{total prediction error sum of squares} = \sum (y_{pp} - y)^2$, $RSS = \text{the residual sum of squares} = \sum (y^2 - y_{pp}^2)$, y η αληθινή τιμή aromatic του δείγματος (original value), y_{pp} η τιμή που προβλέψαμε (predicted value) και το άθροισμα $\sum$ είναι

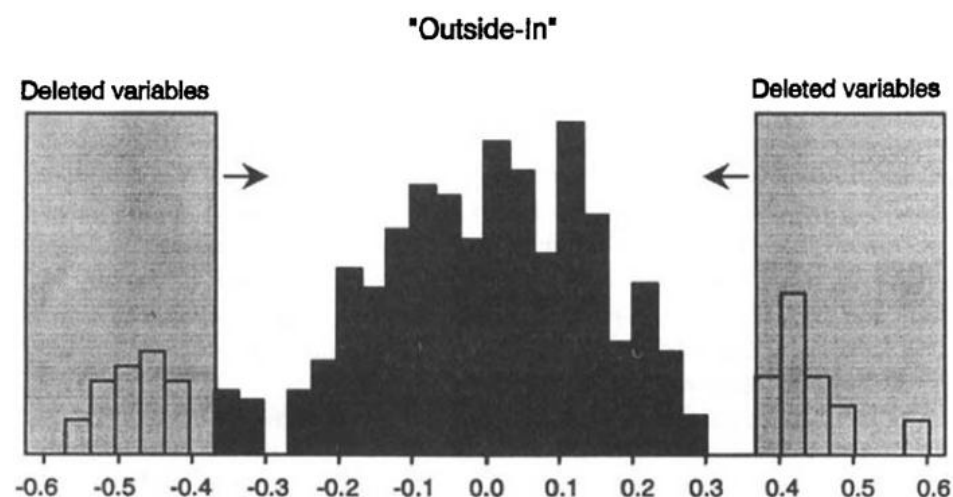
για όλα τα tests που αφορούν μια τιμή α , όπου σε κάθε test αφαιρούμε ένα δείγμα από το σύνολο δειγμάτων για να είναι το test set και αφήνουμε τα υπόλοιπα δείγματα σαν training set η διαδικασία συνεχίζεται μέχρι όλα τα δείγματα να έρθουν στη θέση του test set μία φορά, έτσι στο τέλος θα έχουμε για κάθε τιμή α μία τιμή CV. Η καλύτερη τροποποίηση του w αντιστοιχεί στη μικρότερη τιμή του CV.

Outside-In

Στο outside-in το α μετακινείται από 1 στο 0 κάθε φορά κατά 0.01, και έχουμε την ίδια διαδικασία απλώς αντί να σβήσουμε τις τιμές του w που είναι, στην απόλυτη τιμή, μικρότερες από το α σβήνουμε τις τιμές του w που είναι, στην απόλυτη τιμή, μεγαλύτερες από α . Δηλαδή:

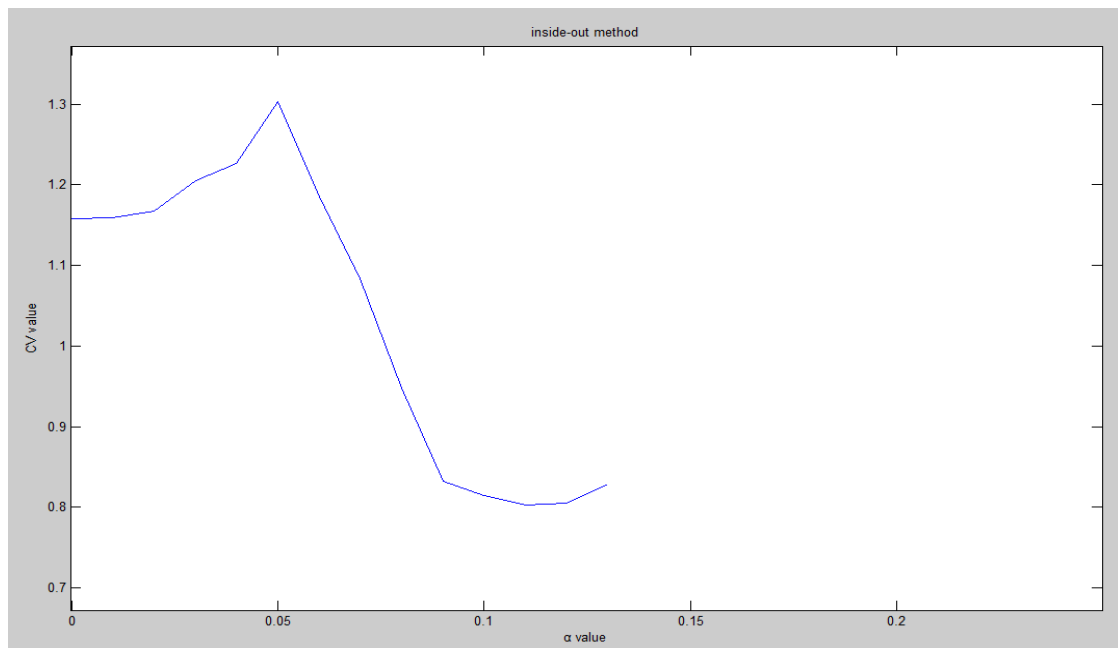
Αν $\text{abs}(w(i)) > \alpha \Rightarrow w(i)=0$ όπου $i=1,2,\dots,K$ και K το σύνολο των μεταβλητών του w

Σχήμα 48, διαγραφή των μεταβλητών με τον τρόπο Outside-In.[2]

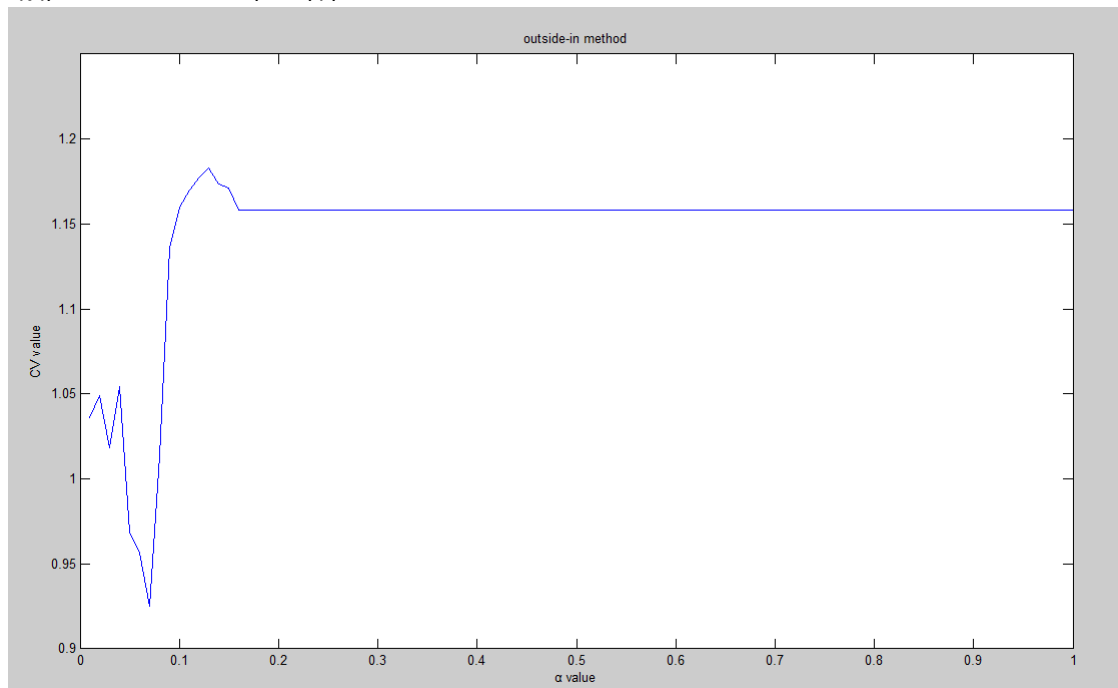


Ακολουθούν τα δύο plots των δύο μεθόδων που αφορούν τα δεδομένα μας και το ιστόγραμμα του πίνακα w .

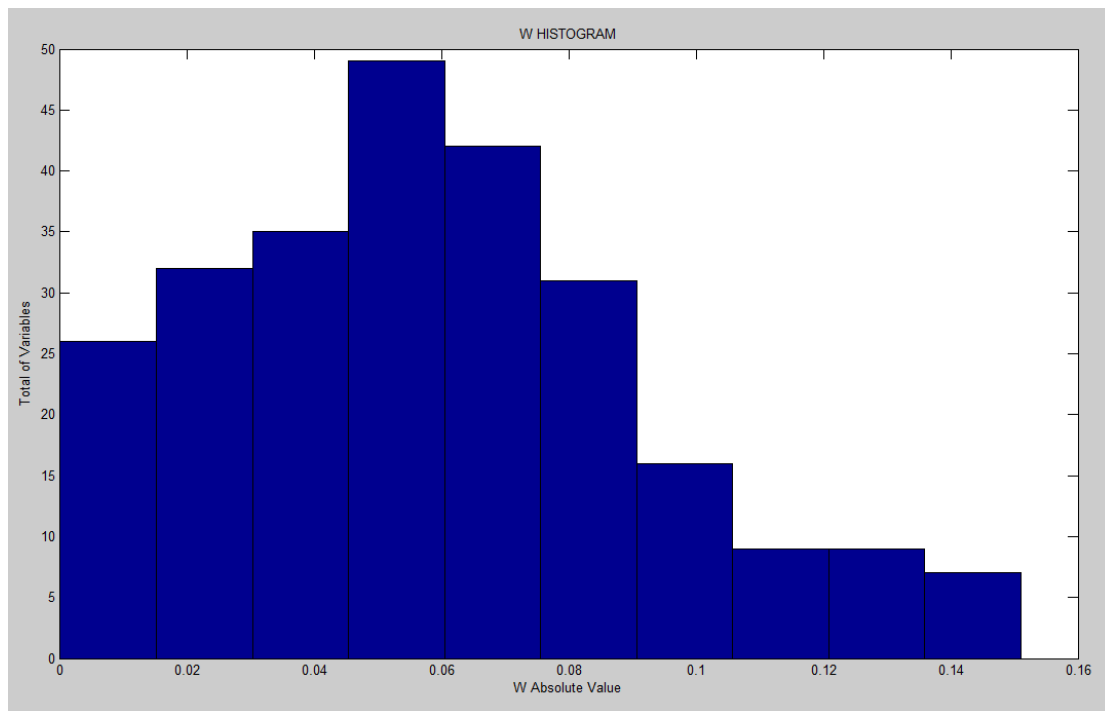
Σχήμα 49, το αποτέλεσμα της μεθόδου *Inside-Out*.



Σχήμα 50, το αποτέλεσμα της μεθόδου *Outside-In*.



Σχήμα 51, το ιστόγραμμα του πίνακα w.



Όπως φαίνεται στο ιστόγραμμα δεν υπάρχουν τιμές στο w που να είναι μεγαλύτερες από 0.16, για αυτό τον λόγο στον κώδικα το α παίρνει τιμές από 0 μέχρι 0.5 για να τρέξει ο κώδικας πιο γρήγορα .

Για τα δεδομένα που έχουμε πέτυχε η μέθοδος Inside-Out στη τιμή $\alpha=0.12$ και βελτιώθηκε η ικανότητα προβλέψης του μοντέλου μετά το σβήσιμο πολλών μεταβλητών

οι οποίες έπαιξαν αρνητικό ρόλο στο μοντέλο. Ο αριθμός των μεταβλητών που έμειναν μετά το σβήσιμο είναι $16 = \sqrt{256}$, ο οποίος είναι ο λιγότερος αριθμός μεταβλητών που επιτρέπεται για το συγκεκριμένο σύνολο μεταβλητών.

Για να αποδείξω τη βελτίωση της ικανότητας προβλέψης επιστρέφουμε στο matlab αρχείο με το όνομα seeAll, και τον πίνακα ooo2 που έχει διαστάσεις 47 x 500 και είναι γεμάτος με τα errors των προβλέψεων για κάθε component και δείγμα. Στην αρχή υπολογίζουμε το καλύτερο σύνολο λαθών για όλα τα δείγματα πριν την εφαρμογή της INTERACTIVE VARIABLE SELECTION method με maximum components = 45 = rank(X) (όπου X ο πίνακας που περιέχει τα φάσματα βενζίνης στο training set), και υπολογίζουμε το ίδιο σύνολο μετά την εφαρμογή της μεθόδου.

πρώτα πριν τη εφαρμογή της μεθόδου χρησιμοποιώντας μόνο το (w) στη πρόβλεψη
 *****matlab
 σε αυτή την περίπτωση ooo2 έχει διαστάσεις 47 x [45 = rank(X)]

$$Sss = \sum_{i=1}^{47} (\min[ooo2(i,j)] \text{ για } i = 1 \dots 47)$$

Sss = 71.8681

Άρα πριν την εφαρμογή της μεθόδου το καλύτερο σύνολο λαθών με μέγξιμουμ components = 45 = rank(X) είναι 71.8681

Τώρα μετά την εφαρμογή της μεθόδου χρησιμοποιώντας μόνο το (w) στη πρόβλεψη
 *****matlab
 σε αυτή την περίπτωση ooo2 έχει διαστάσεις 47 x [45 = rank(X)]

$$Sss = \sum_{i=1}^{47} (\min[ooo2(i,j)] \text{ για } i = 1 \dots 47)$$

Sss = 32.4480

μετά την εφαρμογή της μεθόδου το καλύτερο σύνολο λαθών με μέγξιμουμ components = 45 = rank(X) είναι 32.4480

Αρα έχει γίνει σημαντική βελτίωση στη πρόβλεψη μετά την εφαρμογή της μεθόδου (IVS)

7.4 Ενσωμάτωση τών περιοχών συγκλίσεων τών καλύτερων προβλέψεων στη συνάρτηση plsDip και ολοκλήρωση του μοντέλου

Πριν την ενσωμάτωση πρέπει να αναφέρουμε ότι γενικά ισχύει το εξής: πολλά components σημαίνει καλύτερη πρόβλεψη και λιγότερη σταθερότητα για το μοντέλο και το αντίστροφο ισχύει για τα λίγα components. Για αυτό το λόγω θα έχουμε δύο συναρτήσεις plsDip, η πρώτη συνάρτηση περιέχει τις περιοχές συγκλίσεων τών καλύτερων προβλέψεων οι οποίες βρίσκονται στα components που είναι λιγότερα ή ίσα με 45 = rank(X). Αυτή η συνάρτηση αντιπροσωπεύει τη συνάρτηση με τη καλύτερη σταθερότητα για το μοντέλο.

Πρώτη συνάρτηση plsDip

```
function [w,p,b] = plsDip(X,Y) //η συνάρτηση plsDip λαμβάνει τους πίνακες (X,Y) τα φάσματα
                                βενζίνης και οι αντίστοιχες τιμές αρωματικών και επιστρέφει [w,p,b]
[IX,cX] = size(X);           //παίρνουμε τις διαστάσεις του πίνακα X
t1 = zeros(IX,1);            //δημιουργούμε ένα διάνυσμα t1 με διαστάσεις (IX x 1) όπου IX
                                είναι ο αριθμός γραμμών του πίνακα X.
[v,idx] = max(sum(X.*X));    // δημιουργούμε ένα αρχικό score t παίρνοντας από τον πίνακα X
t = X(:,idx);                τη στήλη με μέγιστο μήκος
u = Y;
for i=1:rank(X)
    t1=t;
    w = X'*u/(u'*u);          //υπολογισμός του w
    w = w/norm(w);            // ομαλοποίηση (normalization)
    t = X*w/(w'*w);           //υπολογισμός του t
    q=1;
    p=X'*t/(t'*t);            //υπολογισμός του p
    pnorm=norm(p);
    p=p/pnorm;                // ομαλοποίηση (normalization)
    t=t*pnorm;                // επανακλιμάκωση (rescale) t
    w=w*pnorm;                // επανακλιμάκωση (rescale) w

% regression and residuals
```

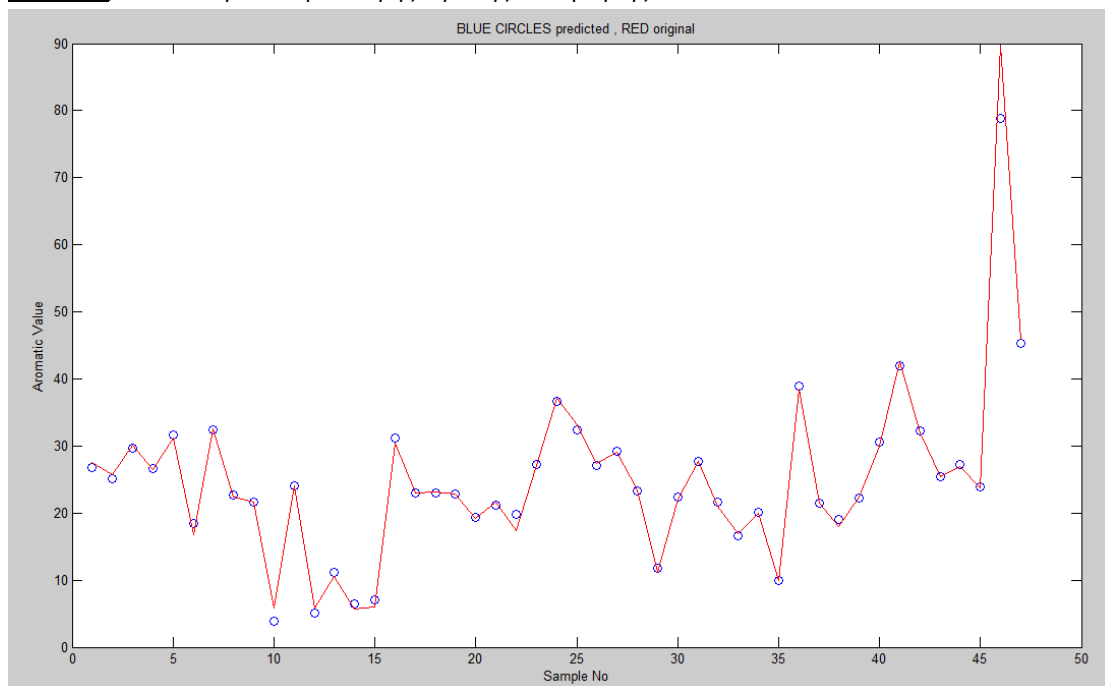
```

b = u'*t/(t'*t); // εύρεση συντελεστή παλινδρόμησης b για την
                    εσωτερική σχέση
X = X - t*p'; // Η εξωτερική σχέση για το μπλόκ X
Y = Y - b*t*q'; // Η μικτή σχέση για το μπλόκ Y
If ((((*)))) // έλεγχος για την επιστροφή, Όπου ((((*)))) μία τιμή
               norm(t-t1) αντιστοιχεί σε μία από τις περιοχές σύγκλισης των
               καλύτερων προβλέψεων με έναν αριθμό components
               λιγότερο ή ίσο με το rank(X), όπου X είναι ο πίνακας
               που περιέχει τα φάσματα βενζίνης στο training set
               componentNO=i // εκτύπωση του αριθμού component πριν την επιστροφή
return
end
end
*****

```

Με τη βοήθεια του Matlab αρείου με το όνομα `Dip` θα περάσουμε το training set στη συνάρτηση `plsDip` η οποία θα μας επιστρέψει τον πίνακα `w` και αφήνουμε στο `w` μόνο τις μεταβλητές που βρήκαμε με τη μέθοδο (IVS). Μετά το σβήσιμο των μεταβλητών στο `w` δεν ξεχνάμε να εφαρμόσουμε την σχέση ($w=w/\text{norm}(w)$) για να κρατήσουμε το μήκος 1 στο `w`. Στη συνέχεια χρησιμοποιούμε το `w` με το φάσμα βενζίνης το οποίο θέλουμε να προβλέψουμε τη τιμή `aromatic` του (αφού του εφαρμόσαμε Mean-centering and scaling of variables) για να βγάλουμε την προβλεπόμενη τιμή για το συγκεκριμένο φάσμα κι έπειτα, κάνουμε την αντιστροφή διαδικασία του Mean-centering and scaling of variables στη προβλεπόμενη τιμή έτσι ώστε να γυρίσουμε στη σωστή κλίμακα τιμών. Η προηγούμενη διαδικασία εφαρμόζεται για κάθε test, όπου σε κάθε test αφαιρείται ένα δείγμα από τα 47 για να είναι το test set και τα υπόλοιπα 46 δείγματα μένουν σαν training set, ο αλγόριθμος τερματίζει όταν κάθε δείγμα από τα 47 μπαίνει στο test set μία φορά .

Σχήμα 52, αποτελέσματα πρόβλεψης πρώτης συνάρτησης.



Το σύνολο λαθών με τη πρώτη συνάρτηση είναι (32.4480).

Το σύνολο λαθών με τη πρώτη συνάρτηση χωρίς το αποτέλεσμα του δείγματος στη θέση 46 με aromatic τιμή ίση με 89.8 είναι (21.5534).

Η ακρίβεια του μοντέλου είναι ίση με το σύνολο τών λαθών των προβλέψεων δια τον αριθμό των tests.

Η ακρίβεια του μοντέλου με την πρώτη συνάρτηση ίσο με (0.690383).

Η ακρίβεια του μοντέλου με την πρώτη συνάρτηση χωρίς το αποτέλεσμα του δείγματος στη θέση 46 με aromatic τιμή ίση με 89.8 είναι (0.468552).

-Στη δεύτερη συνάρτηση δίνουμε πιο πολύ σημασία στη πρόβλεψη αλλά ταυτόχρονα προσπαθούμε όσο μπορούμε να έχουμε λιγότερα components, για αυτό το λόγω έχω θεωρήσει ότι το λάθος προβλέψης είναι δεκτό όταν είναι μικρότερο από 0.6 κι επιπλέον έχω θεωρήσει ότι δεν χρειάζεται η συνάρτηση να κάνει περισσότερα από 500 components, αφού μετά από αυτόν τον αριθμό δεν υπάρχει σημαντική βελτίωση στην πρόβλεψη, δεδομένου ότι ο μεγαλύτερος αριθμός components που χρειάζεται είναι 437 και είναι για το δείγμα στη θέση 46, το οποίο φαίνεται ότι το μοντέλο το προβλέπει πιο δύσκολα από τα άλλα δείγματα.

Δεύτερη συνάρτηση plsDip

```
function [w,p,b] = plsDip(X,Y)    //η συνάρτηση plsDip λαμβάνει τους πίνακες (X,Y) τα φάσματα
                                   βενζίνης και οι αντίστοιχες τιμές αρωματικών και επιστρέφει [w,p,b]

[IX,cX] = size(X);               //παίρνουμε τις διαστάσεις του πίνακα X
t1 = zeros(IX,1);               //δημιουργούμε ένα διάνυσμα t1 με διαστάσεις (IX x 1) όπου IX
                                   είναι ο αριθμός γραμμών του πίνακα X.

[v,idx] = max(sum(X.*X));        // δημιουργούμε ένα αρχικό score t παίρνοντας από τον πίνακα X
t = X(:,idx);                   τη στήλη με μέγιστο μήκος

u = Y;
for i=1:500
    t1=t;
    w = X'*u/(u'*u);              //υπολογισμός του w

    w = w/norm(w);                // ομαλοποίηση (normalization)

    t = X*w/(w'*w);              //υπολογισμός του t
    q=1;
    p=X'*t/(t'*t);              //υπολογισμός του p
    pnorm=norm(p);
    p=p/pnorm;                 // ομαλοποίηση (normalization)

    t=t*pnorm;                  // επανακλιμάκωση (rescale) t
    w=w*pnorm;                  // επανακλιμάκωση (rescale) w

% regression and residuals
```

```

b = u'*t/(t'*t); // εύρεση συντελεστή παλινδρόμησης b για την
// εσωτερική σχέση
X = X - t*p'; // Η εξωτερική σχέση για το μπλόκ X
Y = Y - b*t*q'; // Η μικτή σχέση για το μπλοκ Y

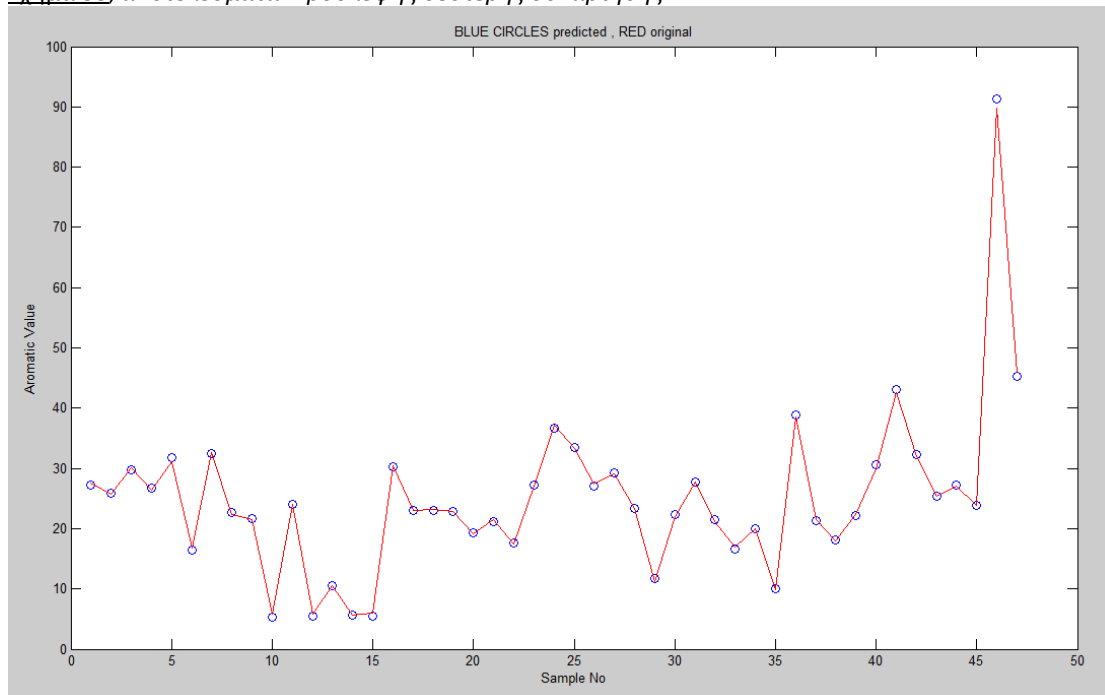
if ((((*)))) // έλεγχος για την επιστροφή, Όπου ((((*)))) μία τιμή norm(t-t1)
// αντιστοιχεί σε μία από τις περιοχές σύγκλισης των
// καλύτερων προβλέψεων με έναν αριθμό components μπορεί
// να είναι μεγαλύτερος του rank(X)
componentNO=i // εκτύπωση του αριθμού component πριν την επιστροφή
return
end

end
*****

```

Πάλι με τη ίδια διαδικασία του Matlab αρχείου Dip θα προβλέψουμε τα δείγματα .

Σχήμα 53, αποτελέσματα πρόβλεψης δεύτερης συνάρτησης.



Το σύνολο λαθών με τη δεύτερη συνάρτηση είναι (13.1817).

Το σύνολο λαθών με τη δεύτερη συνάρτηση χωρίς το αποτέλεσμα του δείγματος στη θέση 46 με aromatic τιμή ίση με 89.8 είναι (11.6224).

Η ακρίβεια του μοντέλου ισούται με το σύνολο των λαθών των προβλέψεων δια τον αριθμό των tests.

Η ακρίβεια του μοντέλου με την δεύτερη συνάρτηση ισούται με (0.280462).

Η ακρίβεια του μοντέλου με την δεύτερη συνάρτηση χωρίς το αποτέλεσμα του δείγματος στη θέση 46 με aromatic τιμή ίση με 89.8 είναι (0.252661).

Συμπεράσματα

Σε αυτή την διπλωματική εργασία αφού μέτρησα κάποια δείγματα βενζίνης στο εργαστήριο με τη βοήθεια της NIR(Near-IR) του εργαλείου Avantes και προχώρησα για να αναλύσω τα φάσματα των μετρήσεων για να δω το περιεχόμενο τους, κατάφερα με μαθηματική προγραμματιστική διαδικασία να αναλύσω τα φάσματα των NIR μετρήσεων των διαφόρων ειδών δειγμάτων βενζίνης σε συνιστώσες για να μάθω περιεχόμενο τους. Αφού τα ανέλυσα, μπόρεσα να διαλέξω τις συνιστώσες οι οποίες εκφράζουν καλύτερα τις πραγματικές τιμές του αρωματικού περιεχόμενου διαφόρων ειδών δειγμάτων βενζίνης. Αυτές οι συνιστώσες μπορούν να χρησιμοποιηθούν για πρόβλεψη του αρωματικού περιεχόμενου καινούριων δειγμάτων βενζίνης με άγνωστες τιμές αρωματικού περιεχόμενου, με αποτέλεσμα να διευκολυνθεί πάρα πολύ η διαδικασία ανάλυσης και μέτρησης διαφόρων ειδών δειγμάτων βενζίνης όσο αφορά το αρωματικό περιεχόμενο. Αξίζει να σημειωθεί ότι ο αριθμός των δειγμάτων και ποιότητα τους στο μοντέλο, παίζει βασικό ρόλο στην επιτυχία του μοντέλου.

REFERENCES

1. P. Geladi and B. Kowalski, *Analytica Chimica Acta*, 186 (1986) 1-17.
2. F. Lindgren, P. Geladi, S. Rannar and S. Wold, *Journal of Chemometrics*, vol. 8, 349-363 (1994).
3. Jerome (Jerry) Workman, Jr., an ASTM Fellow, *An Introduction to Near Infrared Spectroscopy*, Channels: Raman / Infrared Spectroscopy, Published: Sep 12, 2005.
<http://www.spectroscopynow.com/details/education/sepspec1881education/An-Introduction-to-Near-Infrared-Spectroscopy.html>
4. University of Puerto Rico - Mayagüez, Nataly Galan Freyle, Jenny Vargas Irizarry, Carlos Ortega Zuñiga, *NIRS An Advanced Alternative*, QUIM 8995 – Special Topics in Solid State Vibrational Spectroscopy.
<https://www.google.gr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CDMQFjAA&url=http%3A%2F%2Fwww.idrc-chambersburg.org%2Fcnirs%2Fpresentations%2FNIRS%2520An%2520Advanced%2520Alternative.ppt&ei=-MI3Ure6A-Xm4QTdj4CYCg&usq=AFQjCNHfk6xTGxS9ghofvly0plpm4ydf6w&bvm=bv.55819444,d.bGE&ad=rja>
5. Department of Analytical Chemistry, ICT Prague, Pavel Matejka, *NIR spectrometry*.
<http://www.vscht.cz/anl/vibspec/NIR%20spectrometry.pdf>
6. Nicolet Antaris FT-NIR Analyzer, *Theory and Utility of FT-NIR Spectroscopy*.
<http://www.cpatc.embrapa.br/eventos/seminariodequimica/8%BA%20Minicurso%20Infravermelho%20por%20Transformada%20de%20Fourier/nirclaris.pdf>
7. Wikipedia, *Near-infrared spectroscopy*.
http://en.wikipedia.org/wiki/Near-infrared_spectroscopy
8. Physics at High Point University.
<http://physics.highpoint.edu/~mdewitt/phy1050/?page=week2>
9. Μιχαήλ Ι. Κονσολάκης, *Γενική Χημεία Θεωρία & Εφαρμογές*, Η Υπέρυθρη φασματοσκοπία, 2008.
10. Randall D. Tobias, SAS Institute Inc., Cary, NC, *An Introduction to Partial Least Squares Regression*, 1995.
<http://www.ats.ucla.edu/stat/sas/library/pls.pdf>
11. Wikipedia, *molecular vibration*.
http://en.wikipedia.org/wiki/Molecular_vibration
12. Academic Dictionaries and Encyclopedias.
<http://en.academic.ru/dic.nsf/enwiki/267725>
13. BIOPAC Systems, Inc., *fNIR functional near infrared optical brain imaging*.
<http://www.biopac.com/fnir-optical-brain-imaging-hemodynamic-response>

- 14.** D. Michael Olive, Ph.D. Molecular Biology Department, Near Infrared Technology and Optical Agents For Molecular Imaging, 2007.
<http://www.biocompare.com/Application-Notes/43200-Near-Infrared-Technology-And-Optical-Agents-For-Molecular-Imaging/>
- 15.** Medical Expo, The Online Medical Devices Exhibition.
<http://www.medicalexpo.com/prod/topcon-europe-medical/oct-ophthalmoscopes-77876-471530.html>
- 16.** Topcon connecting vision (Medical), 3D OCT-2000 Spectral Domain OCT.
<http://www.topconmedical.com/products/3doct2000.htm>
- 17.** Wikipedia, Absorbance.
<http://en.wikipedia.org/wiki/Absorbance>
- 18.** Wikipedia, Spectroscopy.
<http://ar.wikipedia.org/wiki/%D9%85%D8%B7%D9%8A%D8%A7%D9%81%D9%8A%D8%A9>
- 19.** Susan J. Foulk and Bruce E. DeSimas, Near-infrared (NIR) determination of key parameters for diesel fuels, Process control and quality, 2 (1992) 69-72
- 20.** Xin Bao, Liankui Dai, Partial least squares with outlier detection in spectral analysis: A tool to predict gasoline properties, Fuel 88 (2009) 1216–1222
- 21.** M.R. Maleki; A.M. Mouazen; H. Ramon; J. De Baerdemaeker, Multiplicative Scatter Correction during On-line Measurement with Near Infrared Spectroscopy, Biosystems Engineering (2007) 96(3), 427–433
- 22.** Wikipedia, Gasoline.
<http://en.wikipedia.org/wiki/Gasoline>
- 23.** Wikipedia, Aromatic Hydrocarbon.
http://en.wikipedia.org/wiki/Aromatic_hydrocarbon
- 24.** Wikipedia, Benzene.
[http://ar.wikipedia.org/wiki/%D8%A8%D9%86%D8%B2%D9%8A%D9%86_\(%D8%AD%D9%84%D9%82%D8%A9\)](http://ar.wikipedia.org/wiki/%D8%A8%D9%86%D8%B2%D9%8A%D9%86_(%D8%AD%D9%84%D9%82%D8%A9))
- 25.** Wikipedia, Cresol.
<http://en.wikipedia.org/wiki/Cresol>
- 26.** Wikipedia, Xylenol.
<http://en.wikipedia.org/wiki/Xylenol>
- 27.** Wikipedia, Graphene.
<http://en.wikipedia.org/wiki/Graphene>