Πανεπιστήμιο Ιωαννίνων
Τμήμα Χημείας
Τομέας Ανόργανης & Αναλυτικής Χημείας

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Παρίσης Νικόλαος
(Αριθμός Μητρώου**:** ***2029***)

ΔΙΑΚΡΙΣΗ ΜΕΛΑΝΙΩΝ ΑΠΟ ΣΤΥΛΟ ΔΙΑΡΚΕΙΑΣ ΜΕ ΦΑΣΜΑΤΟΦΩΤΟΜΕΤΡΙΚΗ ΜΕΘΟΔΟ ΚΑΙ ΧΗΜΕΙΟΜΕΤΡΙΑ ΣΤΗ ΔΙΚΑΝΙΚΗ ΕΞΕΤΑΣΗ ΕΓΓΡΑΦΩΝ

Υπεύθυνο μέλος ΔΕΠ:
Αθανάσιος Βλεσσίδης – Λέκτορας Παν/μίου Ιωαννίνων

ΙΩΑΝΝΙΝΑ – ΜΑΪΟΣ 2003

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Τα κύρια υλικά από τα οποία αποτελείται ένα έγγραφο είναι το υποστηρικτικό υλικό (κυρίως χαρτί, χαρτόνι ή πολυμερές υλικό) και το κείμενο το οποίο μπορεί να έχει γραφτεί με απόθεση μελάνης (χειρόγραφα ή με εκτύπωση), φωτοτυπία ή μολύβι. Τα δευτερεύοντα υλικά που μπορούν να βρίσκονται σε ένα έγγραφο περιλαμβάνουν διορθωτικά υλικά, υπολείμματα διαγραφών, κολλητικές ύλες, κηλίδες και δακτυλικά αποτυπώματα. Αναμφίβολα, η εξέταση των υλικών αυτών με επιστημονικές μεθόδους είναι μεγάλης σημασίας στην δικανική εξέταση αμφισβητούμενων εγγράφων.

Αρκετοί ερευνητές έχουν υποστηρίξει τη σημασία διαφόρων τεχνικών για την εξέταση εγγράφων και έχουν προτείνει πρωτόκολλα για τον έλεγχο των μελανιών που έχουν χρησιμοποιηθεί στη συγγραφή κειμένων. Οι μέθοδοι εξέτασης που έχουν προταθεί διακρίνονται σε καταστροφικές και μη καταστροφικές. Από τις καταστροφικές μεθόδους, η ανάλυση με χρωματογραφία λεπτής στοιβάδας (TLC) είναι η συνηθέστερα χρησιμοποιούμενη, ενώ η φασματοσκοπία ανάκλασης είναι η πιο διαδεδομένη μη καταστροφική μέθοδος. Αν και η αξία των μεθόδων αυτών είναι αδιαμφισβήτητη όσον αφορά τις πληροφορίες που μπορούν να αποκαλύψουν για ένα έγγραφο, οι περισσότερες εφαρμογές στηρίζονται σε μετρήσεις μίας μεταβλητής χωρίς σημαντική στατιστική υποστήριξη και τεκμηρίωση.

Στην παρούσα εργασία, περιγράφεται η ανάπτυξη ενός χημειομετρικού πρωτοκόλλου για τη διαφοροποίηση μελανιών από στυλό διαρκείας, το οποίο μπορεί να επιτρέψει την τεκμηριωμένη διάκριση μεταξύ των δειγμάτων. Η μέθοδος βασίζεται στη λήψη του φάσματος απορρόφησης (περιοχή ορατού) κατάλληλα παρασκευασμένων διαλυμάτων μελάνης και την επεξεργασία των φασμάτων με ανάλυση ομαδοποίησης, ανάλυση βασικών συνιστωσών και διαχωριστική ανάλυση για την εξαγωγή συμπερασμάτων.

ΧΗΜΕΙΟΜΕΤΡΙΑ

1. Ανάλυση σε Κύριες Συνιστώσες

1.1 Εισαγωγή

Η μέθοδος των κυρίων συνιστωσών (Principal Components Analysis) είναι μια μέθοδος η οποία έχει σκοπό να δημιουργήσει γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί να είναι ασυσχέτιστοι μεταξύ τους αλλά να περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Το κέρδος από μια τέτοια διαδικασία είναι πως:

• Από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών, κάτι το οποίο για ορισμένες στατιστικές μεθόδους είναι περισσότερο χρήσιμο.

• Αν οι κύριες συνιστώσες που θα προκύψουν μπορούν να ερμηνεύσουν ένα μεγάλο ποσοστό της διακύμανσης τότε αυτό σημαίνει πως αντί να έχουμε p μεταβλητές όπως είχαμε αρχικά, έχουμε λιγότερες, με κόστος βέβαια ότι χάνουμε κάποιο (ελπίζουμε
μικρό) ποσοστό της συνολικής μεταβλητότητας.

• Ένα άλλο μεγάλο πλεονέκτημα (το οποίο από την άλλη ίσως είναι και μειονέκτημα για πολλούς) είναι πως με τη μέθοδο των κυριών συνιστωσών μπορούμε να εξετάσουμε τις συσχετίσεις ανάμεσα στις μεταβλητές και να διαπιστώσουμε πόσο οι μεταβλητές μοιάζουν ή όχι. Επίσης η μέθοδος μας επιτρέπει να αναγνωρίσουμε δίνοντας ονόματα στις καινούριες μεταβλητές (τις συνιστώσες) παρατηρώντας ποιες από τις αρχικές μεταβλητές έχουν μεγάλη επίδραση σε αυτές. Αυτό είναι πολύ χρήσιμο σε κάποιες επιστήμες καθώς μας επιτρέπουν να ποσοτικοποιήσουμε μη μετρήσιμες ποσότητες, όπως η αγάπη, η ευφυΐα, η ικανότητα ενός μπασκετμπολίστα, η εμπορευσιμότητα ενός προϊόντος κλπ αφηρημένες έννοιες. Το γεγονός βέβαια πως τέτοιες ερμηνείες εμπεριέχουν σε μεγάλο βαθμό υποκειμενικά κριτήρια έχει οδηγήσει πολλούς στο να κατηγορούν τη μέθοδο και να μην την εμπιστεύονται.

1.2 Βασική ιδέα

Πριν ξεκινήσουμε την περιγραφή της μεθόδου των κυρίων συνιστωσών είναι χρήσιμο να δούμε κάποια πράγματα από τη γραμμική άλγεβρα τα οποία και αποτέλεσαν τη βασική ιδέα πάνω στην οποία αναπτύχθηκε η μέθοδος.

Έστω ένας τετραγωνικός συμμετρικός πίνακας **A** διαστάσεων p × p. Ο πίνακας αυτός μπορεί να αναπαρασταθεί ως

$$\mathbf{A = P \Lambda P'}$$

όπου **Λ** είναι ένας p × p διαγώνιος πίνακας, όπου τα στοιχεία της διαγωνίου είναι οι ιδιοτιμές του πίνακα **A**, δηλαδή

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

και **P** ένας ορθογώνιος p × p πίνακας (δηλαδή ισχύει **P'P=1**) ο οποίος αποτελείται από τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών. Η παραπάνω αναπαράσταση του πίνακα Α ονομάζεται φασματική ανάλυση του πίνακα Α. Επομένως αφού ο πίνακας είναι ορθογώνιος θα ισχύει πως $\mathbf{P^{-1} = P'}$.

Μπορεί κανείς να δείξει με βάση τις παραπάνω ιδιότητες πως ισχύει:

$$\mathbf{\Lambda = P' A P} \qquad (1.1)$$

καθώς

$$\mathbf{A = P \Lambda P' \Leftrightarrow P^{-1} A = P^{-1} P \Lambda P' \Leftrightarrow P^{-1} A P = \Lambda P' P = \Lambda}$$

Δηλαδή αυτό που με απλά λόγια είδαμε είναι πως αν ξεκινήσουμε από έναν τετραγωνικό πίνακα Α μπορούμε να καταλήξουμε σε έναν διαγώνιο πίνακα Λ.

Γιατί αυτό όμως μας είναι τόσο χρήσιμο; Αν τώρα κοιτάξουμε την σχέση (1.1) βλέπουμε πως από έναν τετραγωνικό πίνακα μπορώ να οδηγηθώ σε έναν διαγώνιο πίνακα, πολλαπλασιάζοντας

με έναν κατάλληλο πίνακα Ρ και άρα αν ο τετραγωνικός πίνακας είναι πίνακας διακύμανσης καταλήγουμε σε έναν διαγώνιο πίνακα διακύμανσης. Δηλαδή το τυχαίο διάνυσμα που αντιστοιχεί στον πίνακα αυτόν είναι ασυσχέτιστο. Δηλαδή αυτό που μου δίνει η φασματική ανάλυση ενός πίνακα διακύμανσης είναι πως αν πολλαπλασιάσω το αρχικό διάνυσμα με έναν κατάλληλο πίνακα μπορώ να

δημιουργήσω έναν νέο διάνυσμα το οποίο να είναι ασυσχέτιστο, να έχει δηλαδή διαγώνιο πίνακα διακύμανσης.

1.3 Εύρεση των κυρίων συνιστωσών

Έστω πως έχουμε ένα σύνολο από k μεταβλητές (X_1, X_2, …, X_k) και θέλουμε να δημιουργήσουμε τις κύριες συνιστώσες (Y_1, Y_2, …, Y_k) οι οποίες να είναι γραμμικός συνδυασμός των αρχικών μεταβλητών , δηλαδή:

$$Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \ldots + \alpha_{1k}X_k$$

$$Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \ldots + \alpha_{2k}X_k$$

$$\ldots$$

$$Y_k = \alpha_{k1}X_1 + \alpha_{k2}X_2 + \ldots + \alpha_{kk}X_k$$

Υπό μορφή πινάκων μπορεί να γραφτεί ως **Y = A X** όπου **Y, X** είναι διανύσματα k x l και **A** είναι k x k πίνακας με στοιχεία

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_k \end{bmatrix}$$

όπου α_j είναι το διάνυσμα στήλη με στοιχεία $\alpha_j' = |\alpha_{j1}\ \alpha_{j2} \ldots \alpha_{jk}|$, j = 1, 2, …, k και για να μην έχουμε πρόβλημα ταυτοποίησης θέτουμε $\sum_{i=1}^{k} a_{ji}^{\ 2} = a_j' a_j = 1$.

Επομένως το πρόβλημα εύρεσης των κυρίων είναι το πρόβλημα της εύρεσης των στοιχείων του πίνακα **A**. Έχουμε όμως έναν επιπλέον περιορισμό, ότι δηλαδή οι κύριες συνιστώσες πρέπει να είναι σε φθίνουσα σειρά ως προς τη διακύμανση τους, δηλαδή η πρώτη να έχει τη μεγαλύτερη διακύμανση, η δεύτερη τη δεύτερη μεγαλύτερη και ούτω καθεξής.

Ας δουλέψουμε για την πρώτη κύρια συνιστώσα $Y_1=\alpha_1'X$. Είναι σαφές πως $Var(Y_1)=\alpha_1'\mathbf{\Sigma}\alpha_1$, όπου **Σ** ο πίνακας διακυμάνσεων του τυχαίου διανύσματος X. Επομένως για να βρούμε το α_1 θα πρέπει να μεγιστοποιήσουμε την $Var(Y_1)$ με τον περιορισμό πως $\alpha_1'\alpha_1=1$, δηλαδή θα μεγιστοποιήσουμε την συνάρτηση

$$L(\alpha_1)=\alpha_1'\mathbf{\Sigma}\alpha_1-\lambda(\alpha_1'\alpha_1-1),$$

όπου λ είναι ο πολλαπλασιαστής Lagrange.

Χρησιμοποιώντας παραγώγους διανυσμάτων βρίσκουμε πως

$$\frac{\partial L(a_1)}{\partial a_1} = 2(\mathbf{\Sigma} - \lambda \boldsymbol{I})a_1 = 0$$

και επομένως αντιστοιχεί στο να λύσουμε την εξίσωση: $\mathbf{\Sigma}\alpha_1 = \lambda\alpha_1$, η οποία είναι η εξίσωση των ιδιοδιανυσμάτων του πίνακα Σ όπου λ είναι η ιδιοτιμή.

Δηλαδή κάθε ζεύγος ιδιοτιμής και του ιδιοδιανύσματος που τη συνοδεύει είναι λύση της εξίσωσης, και άρα έχουμε k δυνατές λύσεις. Από αυτές πρέπει να διαλέξουμε ποια οδηγεί σε μεγαλύτερη διακύμανση. Η διακύμανση του Y_1 θα είναι ίση με λ, και επομένως αρκεί να διαλέξουμε το ζεύγος ιδιοτιμής και ιδιοδιανύσματος που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή.

Με παρόμοια επιχειρήματα μπορούμε να δούμε πως για όλες τις κύριες συνιστώσες τα διανύσματα α_j που χρειαζόμαστε θα αντιστοιχούν στα ιδιοδιανύσματα της j σε φθίνουσα σειρά ιδιοτιμής. Φυσικά για την εύρεση των υπόλοιπων κυρίων συνιστωσών χρειάζεται να προσθέσουμε έναν ακόμη περιορισμό: ότι οι κύριες συνιστώσες είναι ασυσχέτιστες με τις προηγούμενες τους.

Επομένως :

• Για να κατασκευάσουμε τις κύριες συνιστώσες χρειάζεται να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα **Σ** που χρησιμοποιούμε.

• Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμα της αντιστοιχούν στην πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη ιδιοτιμή στη δεύτερη κύρια συνιστώσα κλπ.

• Η διακύμανση της κάθε κύριας συνιστώσας είναι ίση με την ιδιοτιμή που της αντιστοιχεί. Έτσι αν συμβολίσουμε με λ_j την j μεγαλύτερη ιδιοτιμή τότε έχουμε πως $Var(Y_j)=\lambda_j$.

• Όπως είπαμε και πριν οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους και άρα ο πίνακας διακύμανσης τους είναι ο διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές λ_j.

• Η συνολική διακύμανση των κύριων συνιστωσών θα είναι η ίδια με τη συνολική διακύμανση των αρχικών μεταβλητών εξαιτίας των ιδιοτήτων του ίχνους συμμετρικού και τετραγωνικού πίνακα. Δηλαδή θα ισχύει $tr(\mathbf{\Sigma})=tr(\mathbf{\Lambda})$ άρα η συνολική διακύμανση διατηρείται.

• Επίσης η γενικευμένη διακύμανση των κυριών συνιστωσών είναι η ίδια με τη γενικευμένη διακύμανση των αρχικών μεταβλητών. Αυτό προκύπτει εύκολα καθώς η

ορίζουσα ενός τετραγωνικού πίνακα είναι το γινόμενο των ιδιοτιμών της και άρα ισχύει $|\Sigma| = \prod_{i=1}^{p} \lambda_i = |\Lambda|$

• Η ποσότητα $\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$ μας δείχνει το ποσοστό της συνολικής διακύμανσης που εξηγεί η j συνιστώσα. Είναι ευνόητο πως αν κάποιος πάρει όλες τις συνιστώσες τότε θα διατηρήσει όλη τη διακύμανση, ενώ αν τελικά παραλείψει κάποιες συνιστώσες κάποιο ποσοστό της διακύμανσης θα χαθεί. Προφανώς συμφέρει να διατηρούμε τις πρώτες συνιστώσες που εξηγούν μεγαλύτερο κομμάτι της διακύμανσης.

1.4 Βήματα της ανάλυσης σε Κύριες Συνιστώσες.

1.4.1 Έλεγχος συσχετίσεων.

Άσχετα με το αν θα χρησιμοποιήσουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων είναι σκόπιμο να ρίξουμε μια ματιά στον πίνακα συσχετίσεων και να δούμε αν οι αρχικές μας μεταβλητές έχουν συσχετίσεις ή όχι (αυτό γίνεται κυρίως γιατί από τον πίνακα διακύμανσης δεν είναι εύκολο να δούμε την ύπαρξη συσχετίσεων). Αν δεν υπάρχουν συσχετίσεις είναι άσκοπο να συνεχίσουμε. Μεταβλητές που εμφανίζονται ασυσχέτιστες με τις υπόλοιπες πρέπει να τις διώξουμε από την ανάλυση.

Τι εννοούμε όμως όταν λέμε να υπάρχουν συσχετίσεις; Εννοούμε πως η απόλυτη τιμή της συσχέτισης είναι μεγάλη. Αυτό δεν σημαίνει απαραίτητα πως είναι στατιστικά σημαντική, σύμφωνα με το αποτέλεσμα κάποιου ελέγχου υποθέσεων. Ακόμα και συσχετίσεις της τάξης του 0.10 τείνουν να είναι στατιστικά σημαντικές για μέτριου μεγέθους δείγματα (π.χ. 300 παρατηρήσεις). Για να είναι όμως οι συσχετίσεις ικανοποιητικές για να προχωρήσουμε σε ανάλυση σε κύριες συνιστώσες, θέλουμε να είναι της τάξης του 0.4 ή και μεγαλύτερες σε απόλυτη τιμή. Ένα μέτρο που μας επιτρέπει καλύτερα να συγκρίνουμε δύο σετ δεδομένων αλλά και να αξιολογήσουμε αν οι συσχετίσεις είναι ‘ενδιαφέρουσες’ είναι το

$$\phi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 - p}{p(p-1)}}$$

όπου r_{ij} είναι το *ij* στοιχείο του πίνακα συσχετίσεων δηλαδή η συσχέτιση της X_i με τη X_j μεταβλητή. Το στατιστικό *φ* παίρνει τιμές κοντά στο 1 αν υπάρχουν μεγάλες συσχετίσεις, καθώς όλα τα r_{ij} πλησιάζουν σε απόλυτη τιμή τη μονάδα και άρα το άθροισμα των τετραγώνων τους είναι κοντά στο p^2 και άρα ο αριθμητής τείνει να είναι ίσος με τον παρονομαστή. Αν δεν υπάρχουν συσχετίσεις η τιμή θα είναι κοντά στο 0, καθώς μόνο τα p διαγώνια στοιχεία θα είναι 1, άρα το άθροισμα τετραγώνων θα είναι p και άρα ο αριθμητής θα μηδενιστεί. Στην πράξη τιμές πάνω από 0.4 θεωρούνται ικανοποιητικές.

Το αντίστοιχο μέτρο, στην περίπτωση που δουλεύουμε με τον πίνακα διακύμανσης, είναι το

$$\phi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} s_{ij}^2 - \sum_{j=1}^{p} s_{jj}^2}{\sum_{i=1}^{p}\sum_{j\neq 1}^{p} s_{ii} s_{jj}}}$$

για το οποίο ισχύουν παρόμοια πράγματα.

Επομένως, ξεκινώντας την ανάλυση, θα ήταν χρήσιμο κανείς όχι απλά να δει αν οι συσχετίσεις είναι στατιστικά σημαντικά διάφορες του 0 αλλά αν είναι επαρκώς μεγάλες σε απόλυτη τιμή για να προχωρήσει.

1.4.2. Επιλογή πίνακα που θα δουλέψουμε.

Όπως είδαμε μπορούμε να χρησιμοποιήσουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων. Μιλήσαμε προηγουμένως πως επιλέγουμε και με ποια κριτήρια. Πρέπει να γίνει σαφές ότι τα αποτελέσματα θα διαφέρουν ανάλογα με τον πίνακα που θα επιλέξουμε για αυτό η επιλογή είναι βασική για την αξιοποίηση των αποτελεσμάτων που θα προκύψουν.

1.4.3 Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων

Ανάλογα με τον πίνακα που διαλέξαμε να στηρίξουμε την ανάλυση υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα. Τα ιδιοδιανύσματα που δίνουν τα στατιστικά πακέτα είναι κανονικοποιημένα, δηλαδή το άθροισμα τετραγώνων του είναι 1 και δεν είναι μοναδικά από την άποψη πως μπορούμε να τους αλλάξουμε πρόσημο σε όλα τα στοιχεία τους. Συνεπώς η λύση από στατιστικό πακέτο σε στατιστικό πακέτο μπορεί να διαφέρει ως προς τα πρόσημα.

1.4.4 Απόφαση για τον αριθμό των συνιστωσών που θα κρατήσουμε.

Ίσως το πιο σημαντικό κομμάτι της ανάλυσης το οποίο δυστυχώς δεν έχει εύκολη και κοινώς αποδεκτή απάντηση. Κατ' αρχάς να διευκρινίσουμε πως επιλέγοντας λιγότερες κύριες συνιστώσες από όσες μεταβλητές είχαμε αρχικά, χάνουμε αναγκαστικά πληροφορία. Αυτό είναι το κόστος για το κέρδος μας να μειώσουμε τις διαστάσεις του προβλήματος. Συνήθως λοιπόν ενδιαφερόμαστε για κάποιον μικρότερο αριθμό συνιστωσών. Πόσες όμως; Στη βιβλιογραφία υπάρχουν πολλά κριτήρια τα οποία θα προσπαθήσουμε να περιγράψουμε. Αυτά είναι:

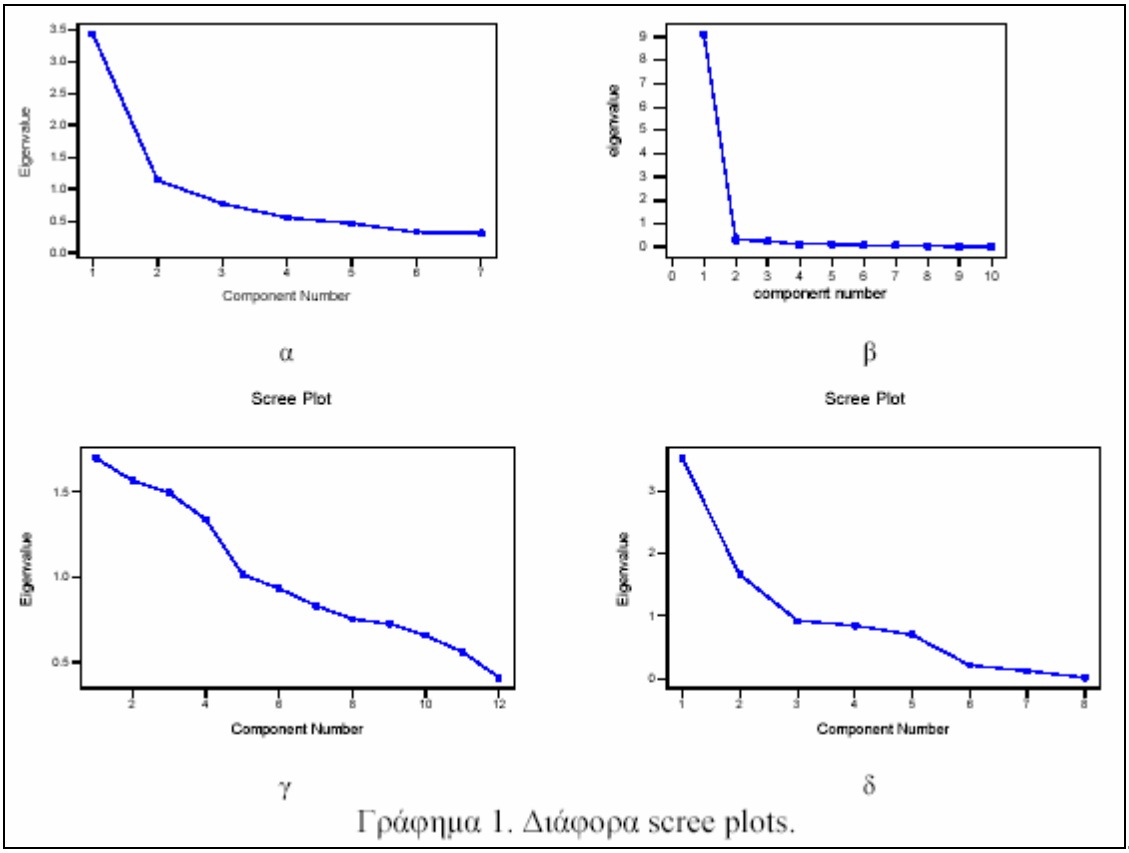
Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες. Σύμφωνα με αυτό το κριτήριο βάζουμε κάποιο όριο (π.χ. 80%) και διαλέγουμε τόσες συνιστώσες ώστε αθροιστικά να εξηγούν μεγαλύτερο ποσοστό από το στόχο που βάλαμε. Είναι πολύ απλό και εύκολο να το χρησιμοποιήσουμε αλλά δυστυχώς στην πράξη δεν δίνει τα καλύτερα αποτελέσματα, ιδίως αν ο στόχος είναι αρκετά υψηλός. Επίσης δεν είναι ξεκάθαρο ποιο ποσοστό της διακύμανσης πρέπει να βάλουμε σαν στόχο.

Κριτήριο του Kaiser. Έστω λ_j οι ιδιοτιμές μας . Το κριτήριο αυτό λέει να πάρουμε τόσες ιδιοτιμές όσες είναι μεγαλύτερες από $\overline{\lambda} = \sum_{j=1}^{k} \lambda_j$. δηλαδή μεγαλύτερες από τη μέση τιμή των ιδιοτιμών. Στην περίπτωση που δουλεύουμε με πίνακα συσχετίσεων ισχύει $\overline{\lambda} = 1$ και άρα διαλέγουμε τόσες συνιστώσες όσες και οι ιδιοτιμές μεγαλύτερες της μονάδας. Το κριτήριο στηρίζεται στην εξής απλή υπόθεση. Αν οι μεταβλητές είναι ασυσχέτιστες και άρα δεν υπάρχει καμιά δομή στα δεδομένα, τότε ο πίνακας συσχετίσεων είναι ο μοναδιαίος και όλες οι ιδιοτιμές είναι ίσες με 1 (δουλεύουμε με πίνακα συσχέτισης). Επομένως κάθε ιδιοτιμή μεγαλύτερη της μονάδας δείχνει την παρουσία κάποιας δομής στα δεδομένα μας.

Στην πράξη η υπόθεση αυτή είναι απλοϊκή καθώς ακόμα και αν δεν υπάρχει δομή και όλες οι ιδιοτιμές είναι 1 όταν δουλέψουμε με ένα δείγμα σίγουρα κάποιες από αυτές θα είναι μεγαλύτερες από 1 αφού το άθροισμα τους πρέπει να είναι *p*. Το κριτήριο συνήθως υπερεκτιμά τον αριθμό των συνιστωσών που χρειάζονται.

Ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται. Όπως είδαμε πριν αν διατηρήσουμε *k* συνιστώσες χάνουμε κάποιο μέρος από την πληροφορία κάθε μεταβλητής και μπορούμε να βρούμε και το ποσοστό της διακύμανσης που ερμηνεύσουμε τελικά. Το κριτήριο αυτό διαλέγει τόσες συνιστώσες ώστε να ερμηνεύεται για κάθε μεταβλητή ένα υψηλό ποσοστό τουλάχιστον. Και πάλι το ποιο είναι αυτό το ποσοστό είναι υποκειμενικό. Επίσης μπορεί κάποια μεταβλητή να μην ερμηνεύεται σωστά και αυτό να οδηγήσει σε μεγάλο αριθμό συνιστωσών.

Scree plot. Το scree plot είναι ένα γράφημα που έχει στον οριζόντιο άξονα των x τη σειρά και στον κάθετο άξονα των y την τιμή της κάθε ιδιοτιμής. Το κριτήριο αυτό προτείνει να πάρουμε τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να γίνεται περίπου επίπεδο, στην ουσία μέχρι να διαπιστώσουμε ότι αρχίζει να αλλάζει η κλίση. Στα scree plot που ακολουθούν (γράφημα 1) μπορεί κανείς να δει τα προβλήματα που παρουσιάζει αυτή η μέθοδος. Στο γράφημα 1β είναι ξεκάθαρο πως θα διαλέξουμε μια μόνο συνιστώσα. Στο γράφημα 1α φαίνεται να διαλέγουμε μια συνιστώσα αλλά κάποιοι θα μπορούσαν να ισχυριστούν ότι πρέπει να πάρουμε 2. Στο γράφημα 1γ τα πράγματα φαίνονται να μην είναι καθόλου καθαρά, ενώ στο 1δ φαίνεται να έχουμε 2 φορές αλλαγή κλίσης. Από αυτά τα γραφήματα γίνεται σαφές πως δεν είναι καθόλου εύκολο να χρησιμοποιήσουμε το scree plot για να επιλέξουμε αριθμό συνιστωσών. Κατ' αρχάς υπάρχει ένα υποκειμενικό κριτήριο για το που και αν αλλάζει η κλίση. Αφετέρου μερικές φορές δεν είναι καθόλου εύκολο να διακρίνει κανείς κάτι τέτοιο για αυτό το scree plot πρέπει να χρησιμοποιείται με προσοχή.



Γράφημα 1. Διάφορα scree plots.

1.4.5 Εύρεση των συνιστωσών

Αυτό αποτελεί το πιο εύκολο ίσως κομμάτι, ιδιαίτερα στις μέρες μας που όλη τη δουλεία την κάνει ο υπολογιστής. Αρκεί να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα που επιλέξαμε για την ανάλυση, σύμφωνα με τη φασματική ανάλυση που είδαμε προηγουμένως.

1.4.6 Ερμηνεία των συνιστωσών

Αυτό το κομμάτι ίσως είναι από τα πιο δύσκολα της ανάλυσης και έχει κατηγορηθεί από πολλούς συγγραφείς. Αφού λοιπόν έχουμε κατασκευάσει τις συνιστώσες πρέπει να προσπαθήσουμε να τους δώσουμε κάποια ερμηνεία, ιδιαίτερα στις πρώτες. Αυτό εξυπηρετεί τους σκοπούς της ανάλυσης καθώς ερμηνεύει τις συσχετίσεις ανάμεσα στις μεταβλητές μας αλλά και αν όλα πάνε καλά μπορούμε να ποσοτικοποιήσουμε κάποιες μη ποσοτικές μεταβλητές.

Στα πλαίσια της ερμηνευτικότητας των συνιστωσών μπορεί κανείς να καταφύγει στην περιστροφή των αξόνων, τεχνική πιο γνωστή από την παραγοντική ανάλυση.

1.4.7 Περιστροφή συνιστωσών

Με την περιστροφή των συνιστωσών προσπαθώ να κάνω τις συνιστώσες πιο ερμηνεύσιμες. Με την περιστροφή δεν αλλάζουν κάποια από τα χαρακτηριστικά του μοντέλου όπως η καλή του προσαρμοστικότητα και το ποσό της διακύμανσης - συνδιακύμανσης που ερμηνεύει το μοντέλο παρά μόνο οι τιμές των επιβαρύνσεων. Γενικά αν **L** είναι ένας πίνακας που περιέχει τις επιβαρύνσεις και **G** ένας ορθογώνιος πίνακας (δηλαδή ισχύει **G΄G=I**) τότε ισχύει πως **LG (LG)'= LGG'L'=LL'**. κι επομένως και ο πίνακας **LG** μπορεί να θεωρηθεί ως ένας πίνακας επιβαρύνσεων. Μαθηματικά ο πίνακας **G** ορίζει έναν ορθογώνιο μετασχηματισμό.

Κάνοντας λοιπόν την περιστροφή ελπίζουμε ότι οι επιβαρύνσεις κάποιων συνιστωσών θα είναι μεγάλες σε απόλυτη κλίμακα μόνο για κάποιες από τις μεταβλητές κι έτσι βλέποντας ποιες μεταβλητές εξαρτώνται με ποιες συνιστώσες να μπορέσουμε να δώσουμε μια ερμηνεία σε αυτές. Οι βασικές μέθοδοι περιστροφής είναι

- *Varimax*: Προσπαθεί να ελαχιστοποιήσει τον αριθμό των μεταβλητών που έχουν μεγάλες επιβαρύνσεις για κάθε συνιστώσα.
- *Quartimax*: Προσπαθεί να ελαχιστοποιήσει τον αριθμό των συνιστωσών που εξηγούν μια μεταβλητή.
- *Equimax*: Συνδυασμός των varimax και quartimax.
- *Oblique*: Μη ορθογώνια περιστροφή, οι άξονες που προκύπτουν δεν είναι πια ορθογώνιοι (και άρα οι συνιστώσες δεν είναι ανεξάρτητες). Η ερμηνεία των αποτελεσμάτων είναι πιο δύσκολη. Στην πράξη τον χρησιμοποιούμε όταν δεν θέλουμε οι συνιστώσες που προκύπτουν να είναι ασυσχέτιστοι.

Η περιστροφή συνήθως καταλήγει σε κάθε συνιστώσα, οι μεταβλητές να χωρίζονται πιο έντονα σε σχέση με το πρόσημο τους, δηλαδή να υπάρχουν λίγες με μεγάλες απόλυτες τιμές ενώ οι υπόλοιπες να τείνουν να έχουν συντελεστή κοντά στο μηδέν. Αυτό βοηθά να αναγνωρίζουμε πιο εύκολα τη συνιστώσα, δηλαδή στην ευκολότερη ερμηνεία της.

1.4.8 Δημιουργία νέων μεταβλητών

Όπως είπαμε οι κύριες συνιστώσες είναι καινούριες μεταβλητές με κάποιες καλές ιδιότητες. Το ενδιαφέρον είναι πως μπορούμε για κάθε παρατήρηση να δημιουργήσουμε τόσες νέες μεταβλητές όσες και οι κύριες συνιστώσες που

αποφασίσαμε να διατηρήσουμε, με σκοπό να χρησιμοποιήσουμε τις κύριες συνιστώσες γα περαιτέρω στατιστική ανάλυση. Για να γίνει αυτό αρκεί να αντικαταστήσουμε στον τύπο της κάθε συνιστώσας τις τιμές που η παρατήρηση είχε για κάθε μεταβλητή.

1.5 Μερικά χρήσιμα αποτελέσματα

Μερικά ενδιαφέροντα αποτελέσματα σχετικά με την ανάλυση σε κύριες συνιστώσες και κάποιες ειδικές της περιπτώσεις είναι τα εξής.

• Αν μια μεταβλητή είναι ασυσχέτιστη με τις υπόλοιπες καλό είναι να την αφαιρέσουμε από την ανάλυση, αφού αν παραμείνει κάποια από τις κύριες συνιστώσες θα ταυτιστεί μαζί της. Όταν δουλεύουμε με δεδομένα αυτό σημαίνει πως δεν έχει στατιστικά σημαντικές συσχετίσεις με τις υπόλοιπες και συνεπώς δεν έχει νόημα να την συμπεριλάβουμε στην ανάλυση.

• Αν δύο ιδιοτιμές προκύψουν ίδιες τότε αυτές αντιστοιχούν σε δύο όμοιες κύριες συνιστώσες κάτι που οδηγεί σε πλεονασμό. Φυσικά στην πράξη κάτι τέτοιο είναι σπάνιο. Αν λοιπόν συμβεί πρέπει να δούμε τα δεδομένα μας μήπως υπάρχει κάποιο πρόβλημα (π.χ. στήλες που επαναλαμβάνονται). Πρέπει να τονιστεί πως για δεδομένα από δείγμα έχει αποδειχτεί πως όλες οι ιδιοτιμές είναι διαφορετικές εκτός από συγκεκριμένες προβληματικές περιπτώσεις.

• Αν έχουμε μηδενικές ιδιοτιμές αυτό σημαίνει πως ο πίνακας που στηρίξαμε την ανάλυση δεν είναι πλήρους βαθμού και άρα κάποιες μεταβλητές είναι γραμμικά εξαρτημένες και πρέπει να τις διώξουμε. Στην πράξη δεν θα συναντήσουμε μηδενικές ιδιοτιμές αλλά πολύ μικρές, κοντά στο μηδέν, ιδιοτιμές. Αυτό υπονοεί ότι κάποιες μεταβλητές είναι σχεδόν γραμμικά εξαρτημένες. Αν αναλογιστεί κανείς πως τέτοιες ιδιοτιμές αντιστοιχούν σε συνιστώσες με σχεδόν μηδενική διακύμανση μπορούμε να τις αγνοήσουμε.. Δηλαδή στην πράξη αφού δύο μεταβλητές θα παρέχουν την ίδια πληροφορία, όλη η πληροφορία θα πάει σε κάποια από τις πρώτες κύριες συνιστώσες και ότι μείνει θα πάει σε μια συνιστώσα με αμελητέα διακύμανση.

• Σε δύο διαφορετικά σετ δεδομένων μπορεί να πάρουμε τα ίδια ιδιοδιανύσματα ενώ οι ιδιοτιμές να αλλάξουν. Στην πράξη αυτό σημαίνει πως παίρνουμε τις ίδιες συνιστώσες αλλά σε κάθε περίπτωση η συνιστώσα εξηγεί

άλλο ποσοστό της διακύμανσης. Συνεπώς δεν πρέπει να περιοριζόμαστε στα ιδιοδιανύσματα αλλά να κοιτάμε και τις ιδιοτιμές.

• Στη γενική περίπτωση που ο πίνακας συσχετίσεων έχει μόνο θετικά στοιχεία (όλες οι συσχετίσεις είναι θετικές) τότε η πρώτη κύρια συνιστώσα μπορεί να εκληφθεί σαν ένας σταθμικός μέσος όρος των μεταβλητών με σταθμίσεις τους αντίστοιχους συντελεστές. Επομένως σε τέτοιες περιπτώσεις μπορούμε να κατασκευάσουμε χρήσιμους δείκτες όπου οι σταθμίσεις έχουν επιλεγεί με έναν συγκεκριμένο τρόπο και όχι εμπειρικά.

• Η βασική ιδέα στην ανάλυση σε κύριες συνιστώσες είναι να γράψουμε τις συνιστώσες ως γραμμικό συνδυασμό των αρχικών μεταβλητών. Είναι εύκολο να δει κανείς πως ομοίως λύνοντας ως προς τις αρχικές μεταβλητές παίρνουμε **X=A΄Y** επειδή ο πίνακας Α είναι ορθογώνιος δηλαδή **A'=A^{-1}.** Επομένως αν έχουμε τα σκορ των συνιστωσών μπορούμε εύκολα να βρούμε τα αρχικά δεδομένα.

1.6 Χρήση των κυρίων συνιστωσών

Η μέθοδος των κυρίων συνιστωσών μπορεί να χρησιμοποιηθεί για διάφορους σκοπούς. Μερικοί από αυτούς είναι οι ακόλουθοι:

• Στη γραμμική παλινδρόμηση όταν οι ανεξάρτητες μεταβλητές είναι συσχετισμένες έχουμε το πρόβλημα της πολυσυγγραμικότητας, όπου πια οι εκτιμήτριες ελαχίστων τετραγώνων παύουν να είναι συνεπείς, οι διακυμάνσεις τους γίνονται πολύ μεγάλες και τα πρόσημα των συντελεστών δεν έχουν κάποια φυσική ερμηνεία. Αν αντί λοιπόν για τις αρχικές συσχετισμένες μεταβλητές χρησιμοποιήσουμε τις κύριες συνιστώσες (όχι απαραίτητα όλες) οι οποίες είναι ασυσχέτιστες το πρόβλημα της πολυσυγγραμικότητας έχει αποφευχθεί. Φυσικά δεν υπάρχει πια κάποια ερμηνεία των συντελεστών αλλά
προβλέψεις μπορούν να γίνουν απλά μετασχηματίζοντας τα καινούρια δεδομένα σε κύριες συνιστώσες.

• Γενικά είναι δύσκολο κανείς να αναπαραστήσει γραφικά πολυδιάστατα δεδομένα. Αν λοιπόν αντί για τα αρχικά δεδομένα αναπαραστήσει γραφικά τις πρώτες κύριες συνιστώσες που ερμηνεύουν μεγάλο κομμάτι της μεταβλητότητας των δεδομένων επιτυγχάνει μια αξιόλογη οπτική παρουσίαση των δεδομένων

• Κοιτάζοντας τα σκορ των παρατηρήσεων στις κύριες συνιστώσες είναι μερικές φορές εύκολο να αποκτήσει κανείς μια ιδέα πως ομαδοποιούνται οι παρατηρήσεις. Αυτό έχει σχέση και με την ευκολότερη γραφική αναπαράσταση των δεδομένων που αναφέρθηκε αμέσως πριν.

• Data mining. Ένας καινούριος επιστημονικός τομέας που συνδυάζει την πληροφορική επιστήμη με τη στατιστική είναι το λεγόμενο data mining (εξόρυξη γνώσης). Η ιδέα είναι πως θα μπορέσουμε να εξάγουμε γνώση από τεράστιες βάσεις δεδομένων (όπως είναι πια οι βάσεις δεδομένων μεγάλων εταιρειών και οργανισμών). Από στατιστικής πλευράς ενδιαφερόμαστε στον να συμπυκνώσουμε την πληροφορία σε όσο γίνεται λιγότερες διαστάσεις και αυτό ακριβώς προσφέρει η ανάλυση σε κύριες συνιστώσες.

• Έλεγχος ποιότητας. Αν κάποιος παρατηρεί μια πληθώρα χαρακτηριστικών ενός προϊόντος με τη χρήση διαγραμμάτων ποιοτικού ελέγχου, είναι σχετικά δύσκολο να βρει πότε το προϊόν έχει βγει εκτός ελέγχου παρακολουθώντας τα πολλά επιμέρους χαρακτηριστικά. Αν όμως συμπυκνώσει την πληροφορία σε κάποιες κύριες συνιστώσες αυτόματα η δουλεία αυτή γίνεται πιο εύκολη.

2. Ανάλυση κατά Συστάδες

2.1 Εισαγωγή

Η *ανάλυση κατά συστάδες* (cluster analysis) είναι μια μέθοδος που σκοπό έχει να κατατάξει σε ομάδες τις υπάρχουσες παρατηρήσεις χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Μπορεί να πει κανείς πως εξετάζοντας πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών η μέθοδος τείνει να δημιουργεί ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους.

Μια επιτυχημένη ανάλυση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς αλλά παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο.

Η σημαντική διαφορά της μεθόδου από τη διαχωριστική ανάλυση είναι πως στη διαχωριστική ανάλυση γνωρίζουμε κάποια ομαδοποίηση ως προς κάποιο χαρακτηριστικό των παρατηρήσεων και θέλουμε να φτιάξουμε κάποιον κανόνα που θα μας βοηθήσει να κατατάξουμε κάποιες καινούριες παρατηρήσεις. Βλέπουμε λοιπόν πως καθώς οι 2 μέθοδοι έχουν κάποια κοινά χαρακτηριστικά ως προς τον τρόπο που λειτουργούν, μπορούν να λειτουργήσουν συμπληρωματικά.

Η ανάλυση κατά συστάδες χρησιμοποιείται σε πολλές επιστήμες για να ομαδοποιήσει δεδομένα. Για παράδειγμα διαφορετικά είδη ζώων μπορούν να ομαδοποιηθούν με βάση κάποια χαρακτηριστικά τους, όπως και οι πελάτες σε μια έρευνα αγοράς. Στην περίπτωση μας αν αγνοήσουμε την πληροφορία που έχουμε σχετικά με την κατάταξη της κατάστασης των αντικειμένων, το ενδιαφέρον θα ήταν να δούμε πως αυτά ομαδοποιούνται χρησιμοποιώντας τις πληροφορίες που έχουμε, δηλαδή τις μεταβλητές για τα χημικά συστατικά που μετρήθηκαν.

Μια πολύ βασική έννοια για την ανάλυση κατά συστάδες αλλά όχι μόνο είναι οι έννοιες της απόστασης και της ομοιότητας. Μπορείτε εύκολα να διαπιστώσετε πως αυτές οι δύο έννοιες είναι αντίθετες, παρατηρήσεις που είναι όμοιες θα έχουν μεγάλη ομοιότητα και μικρή απόσταση. Οι έννοιες αυτές είναι πολύ χρήσιμες καθώς μας επιτρέπουν να μετρήσουμε πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους και επομένως να τις τοποθετήσουμε στην ίδια ομάδα.

Στην ανάλυση κατά συστάδες υπάρχουν 3 διαφορετικές προσεγγίσεις:

- *Ιεραρχικές μέθοδοι*: Ξεκινάμε με κάθε παρατήρηση να είναι από μόνη της μια ομάδα. Σε κάθε βήμα ενώνουμε τις 2 παρατηρήσεις που έχουν πιο μικρή απόσταση. Αν 2 παρατηρήσεις έχουν ενωθεί σε προηγούμενο βήμα ενώνουμε μια προϋπάρχουσα ομάδα με μια παρατήρηση μέχρι να φτιάξουμε μια ομάδα. Κοιτώντας τα αποτελέσματα διαλέγουμε στις πόσες ομάδες θα σταματήσουμε.
- *K-Means*. Ο αριθμός των ομάδων είναι γνωστός από πριν. Με έναν επαναληπτικό αλγόριθμο μοιράζουμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποία ομάδα είναι πιο κοντά στην παρατήρηση.
- *Στατιστικές μέθοδοι*: Και οι δύο μέθοδοι που είπαμε στηρίζονται καθαρά σε αλγοριθμικές λύσεις και δεν προϋποθέτουν κάποιο μοντέλο. Υπάρχουν αρκετές μέθοδοι στατιστικές όπου ξεκινώντας από κάποιες υποθέσεις κατατάσσουμε τις παρατηρήσεις. Δυστυχώς αυτές οι μέθοδοι έχουν αρκετά υπολογιστικά προβλήματα και για αυτό δεν προσφέρονται από πολλά στατιστικά πακέτα που χρησιμοποιούνται στην πράξη.

Σε οποιαδήποτε μέθοδο θα πρέπει να τονιστεί ότι δυστυχώς υπάρχουν πολλά σημεία στα οποία ο ερευνητής μπορεί να λειτουργήσει υποκειμενικά, με αποτέλεσμα από τα ίδια δεδομένα να εξαχθούν ακόμα και αντικρουόμενα αποτελέσματα. Από την άλλη μια γενική αλήθεια είναι πως όταν στα δεδομένα υπάρχουν πραγματικά ομοιογενείς ομάδες τότε οποιαδήποτε μέθοδος θα καταφέρει να τις αναγνωρίσει. Επομένως οι αντιφατικές λύσεις είναι μάλλον μια ένδειξη ότι δεν υπάρχει η κατάλληλη δομή στα δεδομένα μου, δηλαδή δεν υπάρχουν ομοιογενείς ομάδες.

2.2 Ποίες μεταβλητές να χρησιμοποιήσω

Στην πραγματικότητα δεν υπάρχει κάποιος τρόπος για να με οδηγήσει στην επιλογή μεταβλητών πριν κάνω την ανάλυση. Επομένως διαλέγω τις μεταβλητές που πιστεύω για κάποιους λόγους ότι έχουν τη δυνατότητα να δημιουργήσουν ομοιογενείς ομάδες. Αφού κάνω την ανάλυση μπορώ εκ των υστέρων να δω να κάποιες μεταβλητές τελικά ήταν αδιάφορες με την έννοια ότι η τιμή τους είναι η ίδια για όλες τις ομάδες που δημιούργησα κι επομένως δεν μου προσφέρουν κάποια

πληροφορία. Αν μάλιστα θεωρώ ότι δεν μου προσφέρει αυτή η μεταβλητή κάτι σχετικά με την ερμηνεία που αναζητώ μπορώ να την αφαιρέσω και να χρησιμοποιήσω τις υπόλοιπες κάνοντας ξανά τη διαδικασία από την αρχή

2.2.2 Ποία απόσταση / ομοιότητα να χρησιμοποιήσω

Η επιλογή της απόστασης έχει να κάνει με τη μέθοδο που θα χρησιμοποιήσω αλλά και τον τύπο των δεδομένων μου καθώς και τα δεδομένα.

- Για δεδομένα συνεχή η ευκλείδεια απόσταση είναι συνήθως η προτιμότερη λύση. Αν κάποια από τις μεταβλητές έχει όμως τεράστια διακύμανση σε σχέση με τις υπόλοιπες, αυτή θα παίζει σπουδαιότερο ρόλο και άρα θα κατευθύνει και τα αποτελέσματα μου. Σε αυτή την περίπτωση καλό είναι να τυποποιήσω τα δεδομένα μου ώστε να έχουν ίδια μέση τιμή και διακύμανση (άρα και ειδικό βάρος).
- Αν τα δεδομένα μου είναι κατηγορικά σε ονομαστική κλίμακα δυστυχώς το SPSS δεν μου προσφέρει κάποια έγκυρη απόσταση. Πρέπει να τονίσουμε πως κάθε απόσταση μπορεί και πρέπει να χρησιμοποιείται με συγκεκριμένο τύπο δεδομένων. Τέλος αν τα δεδομένα μας είναι δυαδικά (0-1, παρουσία χαρακτηριστικού, απουσία χαρακτηριστικού) τότε:
    - Αν η κοινή απουσία ενός χαρακτηριστικού από 2 άτομα δείχνει ομοιότητα τότε ο Simple Matchin συντελεστής είναι καλή επιλογή.
    - Επειδή όμως σε πολλές εφαρμογές η κοινή απουσία δεν σημαίνει τίποτα (πχ στην ιατρική η απουσία κάποιου συμπτώματος δεν λέει κάτι για την αρρώστια) τότε ο συντελεστής του Jaccard είναι μια καλή επιλογή
- Τέλος αν τα δεδομένα περιέχουν και συνεχή και δυαδικά δεδομένα τότε μια καλή πρόταση είναι να χρησιμοποιήσετε την απόσταση Block αφού πρώτα μετασχηματίσετε τα δεδομένα να παίρνουν τιμές στο διάστημα από 0 έως 1.

2.2.3 Πόσες ομάδες θα φτιάξω.

Ανάλογα με τη μέθοδο που θα χρησιμοποιήσω ο αριθμός των ομάδων μπορεί να είναι γνωστός από πριν (K-Means) ή αλλιώς θα τον επιλέξω αφού δω τα αποτελέσματα μου (Hierarchical clustering). Στην πραγματικότητα τα κριτήρια επιλογής του αριθμού των ομάδων είναι πολλά, αλλά μερικές φορές η ερμηνεία που μπορώ να δώσω είναι ο καλύτερος οδηγός για να επιλέξω αυτόν τον αριθμό. Μια

καλή ιδέα είναι να τρέξω πρώτα μια ιεραρχική ανάλυση και αφού δω όλες τις λύσεις κι επιλέξω τον αριθμό των ομάδων να τρέξω μια μέθοδο K-means για να δημιουργήσω τις ομάδες

2.2.4 Ποία μέθοδο να χρησιμοποιήσω;

Το τελευταίο ερώτημα έχει να κάνει με την επιλογή ανάμεσα στις 2 μεθόδους που έχω διαθέσιμες. Γενικά οι ιεραρχικές μέθοδοι δεν είναι καλή ιδέα να χρησιμοποιούνται για μεγάλο πλήθος δεδομένων καθώς απαιτούν πολύ χρόνο και υπολογιστική ισχύ. Επίσης υπάρχει η τάση να δημιουργούνται ομάδες με ανομοιογενές μέγεθος. Από την άλλη η μέθοδος K-means ενώ αποφεύγει αυτά τα προβλήματα και δουλεύει ικανοποιητικά με μεγάλα δείγματα και δημιουργεί ομάδες παραπλήσιου μεγέθους, εξαρτάται πολύ από τις αρχικές τιμές που θα χρησιμοποιήσουμε.

2.3 Η μέθοδος K-Means

Πρέπει να έχω επιλέξει εκ των προτέρων τον αριθμό των ομάδων που θα προκύψουν. Η μέθοδος δουλεύει επαναληπτικά. Χρησιμοποιεί την έννοια του κεντροειδούς (centroid) και στη συνέχεια κατατάσσει τις παρατηρήσεις ανάλογα με την απόσταση τους από τα κεντροειδή όλων των ομάδων. Το κεντροειδές δεν είναι τίποτα άλλο από τη μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας. Όπως είπαμε και πριν ο αλγόριθμος αυτός δουλεύει ικανοποιητικά για μεγάλα σετ δεδομένων επειδή σε αυτή την περίπτωση δουλεύει πολύ πιο γρήγορα από την ιεραρχική ομαδοποίηση. Αυτός είναι και ο λόγος που η μέθοδος μερικές φορές καλείται και γρήγορη ομαδοποίηση (Quick Clustering). Ο αλγόριθμος είναι ο εξής:

- Βήμα $1^{ο}$. Βρες τα αρχικά κεντροειδή.
- Βήμα $2^{ο}$. Κατάταξε κάθε παρατήρηση στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση.
- Βήμα $3^{ο}$. Από τις παρατηρήσεις που είναι μέσα στην ομάδα υπολόγισε τα νέα κεντροειδή.

- Βήμα 4^{ο}. Αν τα νέα κεντροειδή δεν διαφέρουν από τα παλιά σταμάτα αλλιώς πήγαινε στο βήμα 2.

Τα αρχικά κεντροειδή μπορούν είτε να οριστούν από το χρήστη είτε υπολογίζονται με κάποιο συγκεκριμένο αλγόριθμο από το πακέτο. Τα κριτήρια τερματισμού (βήμα 4) μπορούν να οριστούν από το χρήστη καθώς για μεγάλα σετ δεδομένων με πολύπλοκη δομή ο αλγόριθμος μπορεί να καθυστερήσει πολύ αν το κριτήριο τερματισμού είναι τόσο αυστηρό.

2.4 Ιεραρχική ομαδοποίηση

Στην ιεραρχική ομαδοποίηση, ο αριθμός των ομάδων δεν είναι γνωστός από πριν. Οι μέθοδοι λειτουργούν ιεραρχικά με την έννοια ότι ξεκινούν χρησιμοποιώντας κάθε παρατήρηση σας μια ομάδα και σε κάθε βήμα ενώνουν ομάδες οι παρατηρήσεις που βρίσκονται πιο κοντά. Επειδή χρησιμοποιούν έναν πίνακα αποστάσεων (δηλαδή τις αποστάσεις όλων των παρατηρήσεων από τις υπόλοιπες χρειάζονται πολύ χρόνο και χώρο στον υπολογιστή και για αυτό είναι ασύμφορες για μεγάλα σετ δεδομένων. Ο αλγόριθμος που δουλεύουν είναι ο εξής:

- Βήμα 1. Δημιούργησε τον πίνακα αποστάσεων για όλες τις παρατηρήσεις.
- Βήμα 2. Βρες τη μικρότερη απόσταση και ένωσε τις παρατηρήσεις που την έχουν. Δηλαδή δημιουργώ μια ομάδα με τις παρατηρήσεις που είναι πιο κοντά. Αν η μικρότερη απόσταση αφορά μια ήδη δημιουργηθείσα ομάδα και μια παρατήρηση απλά βάζω αυτή την παρατήρηση σε αυτή την ομάδα ή αν αφορά 2 ομάδες που ήδη υπάρχουν τις ενώνω.
- Βήμα 3. Αν δεν έχουν όλες οι παρατηρήσεις μπει σε μια ομάδα πήγαινε στο βήμα 1 αλλιώς σταμάτα.

Το κρίσιμο σημείο για τον αλγόριθμό είναι πως θα υπολογίσω την απόσταση της ομάδας που έφτιαξα (είτε από συγχώνευση άλλων ομάδων είτε από συγχώνευση παρατηρήσεων). Υπάρχουν πολλές μέθοδοι, όπως

- Η μέθοδος του κοντινότερου γείτονα (nearest neighbour or single linkage)
- Η μέθοδος του μακρινότερου γείτονα (furthest neighbour or complete linkage)

- Η μέθοδος του μέσου ανάμεσα στις ομάδες (Average between groups)
- Η μέθοδος του μέσου μέσα στις ομάδες (Average within groups)
- Η μέθοδος του Ward's και άλλες

Από αυτές η πιο απλή είναι η μέθοδος του κοντινότερου γείτονα η οποία όμως είχε το μειονέκτημα πως δίνει ομάδες με μεγάλες διαφορές ως προς το μέγεθος τους. Η μέθοδος του Ward έχει το πλεονέκτημα ότι μας δίνει περίπου ισοπληθείς ομάδες και για αυτό καλό είναι να την προτιμάμε.

Τα μειονεκτήματα της ιεραρχικής ομαδοποίησης είναι ότι δεν συμφέρει από άποψη πως έχουν κάποια καλή διαχωριστική ικανότητα.

3. Διαχωριστική Ανάλυση

3.1 Εισαγωγή

Ας υποθέσουμε ότι έχουμε Κ πληθυσμούς (ομάδες) Π_1, Π_2, …, Π_K με Κ ≥ 2. Τότε για κάθε πληθυσμό Π_k έχουμε και μία κατανομή $f_k(x)$. Σκοπός της *Διαχωριστικής Ανάλυσης* (ή *Διακριτικής Ανάλυσης*) είναι να «διαχωρίσει» ή να κατανείμει κάθε παρατήρηση στους Κ γνωστούς πληθυσμούς – ομάδες. Προφανώς ψάχνουμε για ένα διαχωριστικό κανόνα που μπορεί να καταχωρίσει σωστά όσο τον δυνατόν περισσότερες παρατηρήσεις.

Ενώ η *διαχωριστική ανάλυση* μοιάζει με την *ανάλυση κατά συστάδες* έχει σημαντικές διαφορές. Η πρώτη και ποιο σημαντική είναι ότι στη διαχωριστική ανάλυση οι ομάδες είναι γνωστές ενώ στην ανάλυση κατά συστάδες δεν είναι. Για το λόγο αυτό ο στόχος είναι διαφορετικός. Στη διαχωριστική ανάλυση κύριο μέλημα μας είναι η κατασκευή ενός κανόνα που θα μας βοηθήσει να λάβουμε αποφάσεις στο μέλλον ενώ στην ανάλυση κατά συστάδες ο κύριος στόχος μας είναι να δημιουργήσουμε ομοειδής ομάδες με κύριο στόχο την κατανόηση των ήδη υπαρχόντων στοιχείων και τη μείωση της διασποράς σε επιμέρους ομάδες.

3.2 Ο Βασικός Κανόνας Διαχωρισμού Δυο Ομάδων

Στην ουσία έχουμε να αντιμετωπίσουμε ένα πρόβλημα θεωρίας αποφάσεων. Έτσι, όταν μπορούμε να ποσοτικοποιήσουμε τις απώλειες λόγω λανθασμένης κατάταξης μπορούμε να γράψουμε το αναμενόμενο κόστος ταξινόμησης μιας παρατήρησης που προέρχεται από την *k* ομάδα (ECM: expected cost of misclassification) δίδεται ως εξής: $ECM_k = \pi_k \sum_{l=1}^{K} C(l \mid k) \cdot P(l \mid k)$, όπου *C(l|k)* είναι το κόστος να κατατάξουμε την παρατήρηση στη *l* ομάδα ενώ ανήκει στην *k*, αν *k=l* τότε το κόστος είναι μηδενικό, *P(l|k)* είναι η πιθανότητα να κατατάξουμε την παρατήρηση στη *l* ομάδα ενώ ανήκει στην *k*, π_l είναι η εκ των προτέρων πιθανότητα (prior probability) να ανήκει μια παρατήρηση στον *l* πληθυσμό (ομάδα) και $f_l(x)$ είναι η

πιθανότητα (η πυκνότητα πιθανότητας) να παρατηρηθούν οι τιμές (χαρακτηριστικά) του διανύσματος **x** όταν βρισκόμαστε στην *l* ομάδα. Το συνολικό κόστος είναι ίσο με το άθροισμα των επιμέρους ECM_k. Φυσικά επιλέγουμε να κατατάξουμε την παρατήρηση στην ομάδα με το μικρότερο αναμενόμενο κόστος λανθασμένης κατάταξης το οποίο είναι ισοδύναμο με ελαχιστοποίηση του συνολικού κόστους λανθασμένης κατάταξης.

Όταν έχουμε δύο ομάδες (K=2) τότε:

$ECM_1 = \pi_1 [C(1|1) P(1|1) + C(2|1) P(2|1)] = \pi_1 C(2|1) P(2|1)$

$ECM_2 = \pi_2 [C(1|2) P(1|2) + C(2|2) P(2|2)] = \pi_2 C(1|2) P(1|2)$

εφόσον C(2|2) = C(2|2) = 0. Άρα ο βασικός κανόνα διαχωρισμού γίνεται: *επιλέγω να κατατάξω την παρατήρηση μου στην 1η ομάδα αν ECM1 ≤ ECM2 αλλιώς στη 2η ομάδα.*

Ο παραπάνω κανόνας συνεπάγεται, μετά από πράξεις τον κανόνα: *Αν* $\frac{f_1(x_{(i)})}{f_2(x_{(i)})} \geq \frac{\pi_2}{\pi_1} \times \frac{C(1|2)}{C(2|1)}$ *τότε κατατάσσουμε την i παρατήρηση στην 1η ομάδα, διαφορετικά κατατάσσουμε την i παρατήρηση στη 2η ομάδα.* Όπου *x(i)* είναι το διάνυσμα με τα χαρακτηριστικά (μεταβλητές) της *i* παρατήρηση, *C(1|2)* είναι το κόστος που προέρχεται από την λανθασμένη καταχώρηση μιας παρατήρησης στην 1η ομάδα (ενώ πραγματικά ανήκει στη 2η) και *C(2|1)* είναι το κόστος που προέρχεται από την λανθασμένη καταχώρηση μιας παρατήρησης στην 2η ομάδα (ενώ πραγματικά ανήκει στη 1η).

3.2 Η Διαχωριστική Συνάρτηση του Fisher

Ο διαχωριστικός κανόνας του Fisher βασίζεται στην μετατροπή των χαρακτηριστικών **x** σε μονοδιάστατα σκορ μέσω μιας συνάρτησης η οποία λέγεται *διαχωριστική συνάρτηση*. Τα σκορ των δύο ομάδων θα πρέπει να είναι όσο το δυνατόν πιο απομακρυσμένα έτσι ώστε να μπορούμε εύκολα με βάση αυτά τα σκορ να κάνουμε διαχωρισμό και ταξινόμηση των δύο ομάδων. Έτσι λοιπόν ο Fisher πρότεινε τη χρήση γραμμικών συνδυασμών για τη δημιουργία αυτών των σκορ χωρίς

να γίνει κάποια υπόθεση για την κατανομή των ομάδων. Η γραμμικότητα υιοθετήθηκε για λόγους ευκολίας. Παρόλα αυτά υπέθεσε ισότητα των πινάκων συνδιακύμανσης αφού χρησιμοποίησε τη συνδυασμένη κοινή (pooled) εκτίμηση $\mathbf{S_p}$.

Έστω λοιπόν ότι τα σκορ δίνονται ως U_1 για την 1η ομάδα και ως U_2 για τη 2η ομάδα. Τότε ένα μέτρο του πόσο κοντά είναι τα σκορ των δύο ομάδων δίνεται από την απόσταση των μέσων τιμών ($\overline{U}_1 - \overline{U}_2$). Ο Fisher μέτρησε την απόσταση σε τυπικές αποκλίσεις κατά απόλυτες τιμές, δηλαδή πήρε σαν μέτρο απόστασης των δύο ομάδων την ποσότητα:

$$D = \frac{\left|\overline{U}_1 - \overline{U}_2\right|}{S_u} \text{ με } S_u = \frac{\sum_{i\in G_1}\left(U_i - \overline{U}_1\right)^2 + \sum_{i\in G_2}\left(U_i - \overline{U}_2\right)^2}{n_1 + n_2 - 2}$$

όπου $i \in G_i$ σημαίνει ότι λαμβάνουμε υπόψη τις παρατηρήσεις που ανήκουν στην i ομάδα. Σκοπός είναι να μεγιστοποιήσουμε την ποσότητα D ή αντίστοιχα την απόσταση D2.

Έστω ο γραμμικός συνδυασμός $\mathbf{L}^T\mathbf{x}$ τότε πρέπει να μεγιστοποιήσουμε την ποσότητα $D^2 = \frac{\left[\mathbf{L}^T\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)\right]^2}{\mathbf{L}^T\mathbf{S}_p\mathbf{L}}$. Από την ανισότητα Cauchy-Schwarz έχουμε ότι για κάθε px1 διανύσματα **a** και **b** ισχύει ότι $(\mathbf{a}^T\mathbf{b})^2 \le (\mathbf{a}^T\mathbf{a})(\mathbf{b}^T\mathbf{b})$. Εφόσον ο πίνακας συνδιακυμάνσεων είναι θετικά ορισμένος μπορούμε να θέσουμε $\mathbf{a} = \mathbf{S}_p^{1/2}\mathbf{L}$ και $\mathbf{b} = \mathbf{S}_p^{1/2}\left(\bar{x}_1 - \bar{x}_2\right)$ τότε έχουμε:

$$\left[\mathbf{L}^T\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)\right]^2 \le \left(\mathbf{L}^T\mathbf{S}_p^{1/2}\mathbf{S}_p^{1/2}\mathbf{L}\right)\left[\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)^T\mathbf{S}_p^{-1/2}\mathbf{S}_p^{-1/2}\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)\right] \Leftrightarrow$$
$$\left[\mathbf{L}^T\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)\right]^2 \le \left(\mathbf{L}^T\mathbf{S}_p\mathbf{L}\right)\left[\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)^T\mathbf{S}_p^{-1}\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)\right] \Leftrightarrow$$
$$\mathrm{D}^2 = \frac{\left[\mathbf{L}^T\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)\right]^2}{\left(\mathbf{L}^T\mathbf{S}_p\mathbf{L}\right)} \le \left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)^T\mathbf{S}_p^{-1}\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)$$

Άρα για $\mathbf{L} = \mathrm{c}\,\mathbf{S}_p^{-1}\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)$, όπου c > 0, έχουμε $\mathrm{D}^2 = \left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)^T\mathbf{S}_p^{-1}\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)$ δηλαδή τη μέγιστη απόσταση μεταξύ των μέσων και τον καλύτερο δυνατόν διαχωρισμό (συνήθως παίρνουμε c=1). Ο διαχωριστικός κανόνας ολοκληρώνεται ορίζοντας την κρίσιμη τιμή η οποία δεν είναι άλλη από την μέση τιμή των $\overline{U}_1$και $\overline{U}_2$, δηλαδή η ποσότητα

$$m = \left(\overline{U}_1 + \overline{U}_2\right)/2 = \mathbf{L}^T\left(\overline{\mathbf{x}}_\mathbf{1} + \overline{\mathbf{x}}_\mathbf{2}\right)/2$$

η οποία ισαπέχει από τα $\overline{U}_1$ και $\overline{U}_2$. Έτσι ο διαχωριστικό κανόνας γίνεται: *Αν* $\boldsymbol{L^T x} \geq m$ *(ή* $\boldsymbol{L^T x}-m \geq 0$*) τότε κατατάσσουμε στην* $1^η$ *ομάδα αλλιώς στην* $2^η$*.*

3.3 Γενίκευση Διαχωριστικής Ανάλυσης σε Κ ομάδες

Για να γενικεύσουμε τη μέθοδο σε διαχωρισμό Κ ομάδων πρέπει να υπολογίσουμε τα σκορ:

$$W_k = -\sum_{l=1}^{K} \pi_l C(k \mid l) f_1(\mathbf{x})$$

τα οποία αντιστοιχούν σε ελαχιστοποίηση του συνολικού κόστους λανθασμένης κατάταξης, και καταχωρούμε την παρατήρηση μας στην ομάδα με το μεγαλύτερο σκορ. Στην περίπτωση των κανονικών κατανομών και όταν έχουμε ίσα κόστη μπορούμε εναλλακτικά να υπολογίσουμε τα σκορ $W_k = \mathbf{L}_k^T \mathbf{x} - \frac{1}{2}\mathbf{L}_k^T \overline{x}_k + \ln(\pi_k)$ για

k = 1, 2, …, K, όπου Κ ο αριθμός των ομάδων, $\overline{x}_k$ ο δειγματικός μέσος της k ομάδας, $\mathbf{L}_k = \mathbf{S}_p^{-1}\overline{x}_k$ και $\boldsymbol{S_p}$ είναι ο κοινός συνδυασμένος εκτιμητής του πίνακα διακύμανσης-συνδιακύμανσης που δίδεται ως:

$$\mathbf{S_p}=\omega_1\mathbf{S_1}+ \omega_2\mathbf{S_2} + \ldots+\omega_K\mathbf{S_K}$$

με $\boldsymbol{S_k}$ την εκτίμηση του πίνακα διακυμάνσεων της k ομάδας, $\omega_k=(n_k-1)/(n-K)$ και n_k ο αριθμός των παρατηρήσεων στην k ομάδα. Οι γραμμικές συναρτήσεις W_k λέγονται και γραμμικές διαχωριστικές συναρτήσεις του Fisher και οι τιμές που τελικά παίρνουν λέγονται σκορ των διαχωριστικών συναρτήσεων του Fisher. Τα κεντροειδή τους είναι οι αντίστοιχες μέσες τιμές ενώ οι κανονικοποιημένες διαχωριστικές συναρτήσεις είναι μειωμένες κατά μια (δηλαδή Κ-1) και είναι ανάλογες των διαφορών $Z_k= W_k - W_K$ για k = 1, 2, …, Κ-1 (δηλαδή σε κάθε περίπτωση συγκρίνουν τη κάθε ομάδα με κάποια βασική ομάδα η οποία συνήθως είναι η τελευταία ή η πρώτη).

3.4 Γενίκευση Διαχωριστικής Ανάλυσης του Fisher σε Κ ομάδες

Ο *Fisher* εναλλακτικά πρότεινε μια επέκταση της μεθόδου του για τα διαχωρισμό Κ ομάδων. Έτσι λοιπόν προτείνει τη χρήση Κ-1 γραμμικών συνδυασμών

της μορφής $\mathbf{L}_k^T\mathbf{x}$ με $\mathbf{L}_k$ να είναι τα διανύσματα του πίνακα $\mathbf{\Delta} = (n-K)\,\mathbf{S}_p^{-1}\mathbf{W}$ (υπό τον περιορισμό ότι $\mathbf{L}^T\mathbf{S}_p\mathbf{L}=1$) με σειρά που αντιστοιχεί στο μέγεθος των ιδιοτιμών. Δηλαδή $\mathbf{L_1}$ είναι το ιδιοδιάνυσμα που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή, $\mathbf{L_2}$ είναι το ιδιοδιάνυσμα που αντιστοιχεί στην $2^η$ μεγαλύτερη ιδιοτιμή κ.ο.κ. Όπου $W = \sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T$ είναι ένα μέτρο της διακύμανσης των μέσων τιμών των K ομάδων. Σημείωση ότι $\mathbf{L_1}$ μεγιστοποιεί την ποσότητα $D^2=(n-K)\mathbf{L}^T\mathbf{W}\mathbf{L}/\mathbf{L}^T\mathbf{S}_p\mathbf{L}$ η οποία είναι ένα μέτρο της απόστασης μεταξύ των μέσων δηλαδή ένα μέτρο διαχωρισμού των ομάδων κατά αντιστοιχία με την απόσταση που είχαμε όταν K=2.

Η ερμηνεία των παραπάνω διαχωριστικών συναρτήσεων είναι ότι η $1^η$ διαχωριστική συνάρτηση μεγιστοποιεί τις διαφορές των μέσων σε μια διάσταση. Η $2^η$ διαχωριστική συνάρτηση μεγιστοποιεί την απόσταση των μέσων σε μια κατεύθυνση ορθογώνια στην $1^η$, η $3^η$ μας δείχνει την απόσταση σε μια $3^η$ διάσταση ανεξάρτητη των άλλων 2 κ.ο.κ. Μπορούμε να περιγράψουμε τις διαχωριστικές συναρτήσεις σαν παράγοντες (factors) που διαχωρίζουν βέλτιστα τα κεντροειδή (μέσες τιμές) σε σχέση με τη διασπορά μέσα σε κάθε ομάδα.

Ο διαχωριστικός κανόνας αν κρατήσουμε r διαχωριστικές συναρτήσεις γίνεται:

Ταξινομούμε την παρατήρηση x στην k ομάδα αν $\sum_{l=1}^{r}\left[\mathbf{L}_l\left(\mathbf{x}-\bar{\mathbf{x}}_k\right)\right]^2 \le \sum_{k=1}^{r}\left[\mathbf{L}_l\left(\mathbf{x}-\bar{\mathbf{x}}_i\right)\right]^2$ *για όλα τα i διαφορετικά του k.*

ΕΞΕΤΑΣΗ ΕΓΓΡΑΦΩΝ - ΜΕΛΑΝΙΑ

1. Εισαγωγή

Η ανάλυση μελανιού είναι μια σημαντική δικανική διαδικασία που μπορεί να αποκαλύψει τις χρήσιμες πληροφορίες για τα εξεταζόμενα έγγραφα. Οι περισσότερες από τις εφαρμογές της θεωρούν την ανίχνευση και την επιβεβαίωση των αλλαγών στα έγγραφα με σημαντική οικονομική αξία όπως οι ασφαλιστικές αξιώσεις, οι διαθήκες, οι συμβάσεις και οι φορολογικές επιστροφές. Αυτές οι τροποποιήσεις μπορούν να επιβεβαιωθούν από τη σύγκριση των μελανιών που χρησιμοποιούνται για να παραγάγουν το ψευδές έγγραφο ή τον προσδιορισμό του χρόνου στον οποίο τα διάφορα τμήματα του εγγράφου γράφτηκαν. Είναι επομένως εμφανές ότι υπάρχει μια μεγάλη ανάγκη για την ανάπτυξη ενόργανων μεθόδων που θα επιτρέψουν μια εις βάθος εξέταση των μελανιών που χρησιμοποιούνται για να παραγάγουν ένα έγγραφο και παράλληλα συγκεκριμένα στατιστικά πρωτόκολλα είναι απαραίτητο να ακολουθηθούν έτσι ώστε να προκύψουν αντικειμενικά συμπεράσματα σχετικά με την ομοιότητα του μελανιού, με προκαθορισμένα επίπεδα εμπιστοσύνης.

2. Τα έγγραφα

Οι προαναφερθείσες ανάγκες αυξάνονται από το γεγονός ότι τα έγγραφα είναι μάλλον σύνθετα συστήματα, ειδικά όσον αφορά στη σύνθεση μελανιού. Τα υλικά από τα οποία τα έγγραφα αποτελούνται μπορούν να διαιρεθούν σε κύρια και δευτερεύοντα. Τα κύρια υλικά περιλαμβάνουν το υλικό υποστήριξης του (π.χ. χαρτί, χαρτόνι, πολυμερές, κλπ…) και το κείμενο (π.χ. μελάνι, «καρμπόν», toner φωτοτυπικών μηχανημάτων, μολύβι, κλπ…). Τα δευτερεύοντα υλικά, που δεν είναι ουσιαστικά στην ύπαρξη του εγγράφου, εμφανίζονται συνήθως ως αποτέλεσμα των διορθώσεων και της διαχείρισης του εγγράφου. Αυτά μπορούν να περιλάβουν τα διορθωτικά υλικά, τα υπολείμματα σβησιμάτων, τις κόλλες, τους λεκέδες, τα δακτυλικά αποτυπώματα, κλπ.

3. Μελάνια

3.1 Σύσταση

Η σύσταση του μελανιού είναι πολύπλοκη, δεδομένου ότι τα σύγχρονα μελάνια περιέχουν έναν μεγάλο αριθμό ουσιών που στοχεύουν να βελτιώσουν τα χαρακτηριστικά του. Προφανώς, το σημαντικότερο συστατικό είναι το χρωματίζον υλικό. Αυτό βρίσκεται υπό μορφή απλής βαφής (dye based ink) ή/και χρωστικής ύλης (pigmented ink). Οι βαφές (dyes) μπορούν να είναι όξινες ή βασικές και είναι διαλυτές στο υγρό σώμα του μελανιού που είναι επίσης γνωστό ως όχημα. Αφ' ετέρου, οι χρωστικές ύλες (pigments) είναι λεπτά αλεσμένοι πολυμοριακοί κόκκοι που είναι αδιάλυτοι στο όχημα. Το όχημα, η του οποίου σύνθεση επηρεάζει τα χαρακτηριστικά ροής και ξήρανσης του μελανιού, αποτελείται συνήθως από έλαια (λιναρόσπορου, σόγιας, ορυκτού ή άλλου τύπου ελαίου), διαλύτες (οργανικούς ή ύδωρ) και ρητίνες (μη κρυστάλλινο υλικό υψηλού μοριακού βάρους). Ανάλογα με το βαθμό κορεσμού τους, τα έλαια επηρεάζουν την ταχύτητα ξήρανσης του μελανιού, ενώ το διαλυμένο περιεχόμενο αφορά την αντοχή του μελανιού στο χρόνο. Μια τρίτη κατηγορία ουσιών χρησιμοποιείται για να συμπληρώσει τα χαρακτηριστικά του μελανιού. Παραδείγματος χάριν, τα χαρακτηριστικά ξήρανσης μπορούν να επηρεαστούν περαιτέρω από την παρουσία ξηραντικών (συνήθως ανόργανα άλατα που καταλύουν την οξείδωση των υπό ξήρανση ελαίων). Καλή σταθερότητα του μελανιού επιτυγχάνεται και μέσω της χρήσης πλαστικοποιητών (διαλύτες, αρκετά σταθεροί, που αυξάνουν την αντοχή του μελανιού). Τα κεριά και τα λίπη (π.χ. ζελατίνα πετρελαίου) έχουν επίσης μια παρόμοια επίδραση στη σταθερότητα και του μελανιού.

3.2 Τεχνικές Ανάλυσης

Οι τεχνικές σχετικά με την ανάλυση των μελανιών μπορούν να διαιρεθούν σε καταστρεπτικές και μη καταστρεπτικές όσον αφορά τις αλλαγές που επιφέρουν στο εξεταζόμενο έγγραφο. Στις καταστρεπτικές μεθόδους, μια μερίδα του μελανιού πρέπει να αφαιρεθεί από το έγγραφο πριν από την ανάλυση. Αυτό μπορεί να

περιλάβει την κοπή ενός κομματιού του υλικού υποστήριξης που φέρει το δείγμα ή γρατσούνισμα της επιφάνειας του εγγράφου για να αφαιρεθεί μέρος του μελανιού. Από την άλλη, οι μη καταστρεπτικές μέθοδοι περιλαμβάνουν την παρατήρηση του μελανιού στο έγγραφο με τη βοήθεια μιας τεχνικής ανάκλασης που επιτρέπει την καταγραφή των φασματικών χαρακτηριστικών μελανιού χωρίς αφαίρεση του δείγματος από το υλικό υποστήριξης.

Οι χρωματογραφικές και ηλεκτροφορητικές μέθοδοι έχουν πάντα προτιμούνται από τους επιστήμονες όσον αφορά στην ανάλυση των μελανιών. Αυτή η προτίμηση προέρχεται από το γεγονός ότι τα μελάνια είναι μάλλον σύνθετα μίγματα που απαιτούν το χωρισμό των συστατικών τους για επιτυγχάνεται καλή διάκριση. Η χρωματογραφία εγγράφου είναι μία από τις παλαιότερες καταστρεπτικές μεθόδους που υιοθετούνται στην ανάλυση μελανιού και έχει χρησιμοποιηθεί ειδικά για τα οργανικά βασισμένα στη χρωστική ουσία μελάνια, δεδομένου ότι τα παλαιότερα σιδηρο-γαλλοτανικά μελάνια είναι δύσκολο να διαχωριστούν με τη χρωματογραφία. Η χρωματογραφία λεπτής στοιβάδας (TLC) και οι παραλλαγές της, συμπεριλαμβανομένης της χρωματογραφίας δίσκου και της χρωματογραφίας λεπτής στοιβάδας υψηλής απόδοσης (HPTLC), έχουν αντικαταστήσει βαθμιαία τη χρωματογραφία εγγράφου και έχουν αποδειχθεί ότι είναι ικανοποιητικοί αντικαταστάτες της για το χωρισμό των συστατικών του μελανιού. Η παρατήρηση των χρωματογραφημάτων λεπτής στοιβάδας κάτω από εναλλακτικές πηγές φωτός, η χρήση υπέρυθρης φωταύγειας και μικροφασματοφωτομετρίας έχουν υιοθετηθεί επίσης σε μία προσπάθεια να επιτευχθεί ο καλύτερος χαρακτηρισμός των ζωνών TLC. Αν και η TLC παραμένει η συνιστώμενη μέθοδος για την ανάλυση μελανιού λόγω του χαμηλότερου κόστους και της σχετικής απλότητάς της, η υγρή χρωματογραφία υψηλής απόδοσης (HPLC) έχει χρησιμοποιηθεί ως εναλλακτική καταστρεπτική τεχνική που έχει το πλεονέκτημα της υψηλότερης ανάλυσης και είναι επίσης ικανή να ανιχνεύσει τα άχρωμα συστατικά του μελανιού. Ο χαρακτηρισμός των μελανιών έχει επιτευχθεί επίσης με τη βοήθεια των παραδοσιακών ηλεκτροφορητικών μεθόδων, αλλά με αυξανόμενο κόστος εξοπλισμού και απαίτηση για μεγαλύτερα δείγματα. Τα καλύτερα αποτελέσματα από την άποψη διακριτικότητας, ποσοτικοποίησης και χρόνου ανάλυσης έχουν επιτευχθεί με τη χρήση της τριχοειδούς ηλεκτροφόρησης (CE).

Οι μη καταστρεπτικές τεχνικές, αν και όντας οι πιο χρήσιμες όσον αφορά την ακεραιότητα του εγγράφου, δεν έχουν αναπτυχθεί και χρησιμοποιηθεί λεπτομερώς. Η

παρατήρηση των εγγράφων κάτω από τις εναλλακτικές πηγές φωτός με το γυμνό μάτι είναι πιθανώς η συνηθέστερα χρησιμοποιημένη μέθοδος, αλλά η αξία της περιορίζεται στην παροχή μόνο των ποιοτικών ενδείξεων για τα χρησιμοποιούμενα μελάνια. Τίποτα δεν μπορεί να ειπωθεί για τη σύνθεση μελανιών ή την προέλευσής τους. Αυτό το πρόβλημα έχει ξεπεραστεί με τη χρήση των κατάλληλων ανιχνευτών που μπορούν να μετρήσουν την ανακλώμενη ακτινοβολία από τα δείγματα σε διαφορετικά μήκη κύματος, προσφέροντας κατά συνέπεια συμπληρωματικές πληροφορίες σχετικά με τη φύση και τη σύνθεση του μελανιού.

ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

1. Εισαγωγή

Αν και πολλή έρευνα έχει διεξαχθεί για την ανάπτυξη των αποδοτικών αναλυτικών μεθόδων σχετικά με τη σύνθεση των μελανιών, η χημειομετρία είναι μια περιοχή που δεν έχει χρησιμοποιηθεί εκτενώς για να ερευνήσει και να υποστηρίξει τα αναλυτικά αποτελέσματα. Δεδομένου ότι η δικανική επιστήμη είναι πειθαρχική και πρέπει να βγάλει τα συμπεράσματά της σε καθαρώς αντικειμενική βάση, όποτε αυτό είναι δυνατό, είναι υποχρεωτικό για τους δικανικούς επιστήμονες να ακολουθήσουν άκαμπτα στατιστικά πρωτόκολλα στην επίτευξη των αποφάσεων σχετικά με τα πειραματικά στοιχεία. Επομένως, στην παρούσα εργασία, έχουμε προσπαθήσει να ερευνήσουμε τη χρησιμότητα της χημειομετρίας πολλών μεταβλητών στη διάκριση των μπλε μελανιών από στυλό με σφαιρική άκρη. Αυτό επιτεύχθηκε με την εξαγωγή των χρωστικών ουσιών του μελανιού σε αιθανόλη, την καταγραφή των φασμάτων ορατού (Vis) των εκχυλισμάτων και την εφαρμογή της ανάλυσης κατά συστάδες (Cluster Analysis, CA), της ανάλυσης κύριων συνιστωσών (Principal Component Analysis, PCA) και της διαχωριστικής ανάλυσης (Discriminant Analysis, DA) στα φάσματα. Αν και η χρήση της χημειομετρίας πολλών μεταβλητών με τα φάσματα UV-Vis αποφεύγεται συνήθως, αφού αυτά τα φάσματα παρουσιάζουν ευρείες ζώνες με προβλήματα πολυσυγγραμικότητας, έχουμε ακολουθήσει ένα στατιστικό πρωτόκολλο που μας επέτρεψε σε για να αποφασίσουμε σχετικά με τον αριθμό και την ποιότητα των μεταβλητών (αρχικές ή συνιστώσες) που θα χρησιμοποιηθούν έτσι ώστε μια αποτελεσματική διάκριση των μελανιών να μπορεί να επιτευχθεί.

Μια λεπτομερής έρευνα του συστήματος έδειξε ότι η χρήση των συνιστωσών αντί των αρχικών τιμών απορροφητικότητας σε ορισμένα μήκη κύματος ήταν προτιμητέα εάν επρόκειτο να επιτευχθεί στη διαχωριστική ανάλυση χαμηλότερη συσχέτιση μεταξύ των μεταβλητών. Οι πρώτες τρεις κύριες συνιστώσες αποδείχθηκαν επαρκείς στην εξήγηση του μεγαλύτερου μέρους (96,97%) της διαφοροποίησης των αρχικών στοιχείων και χρησιμοποιήθηκαν ως «αφανείς» μεταβλητές στη διαχωριστική ανάλυση. Το διαχωριστικό μοντέλο που υπολογίστηκε με αυτόν τον τρόπο κατείχε άριστα χαρακτηριστικά, με το Wilk's λ να παίρνει τιμή $8,98 \times 10^{-5}$, ενώ ήταν και ιδιαίτερα στατιστικά σημαντικό στο επίπεδο P=0,05. Η μέση

συσχέτιση μεταβλητών ήταν 33,6% και η ταξινόμηση του συνόλου των δεδομένων κατάρτισης ήταν 100% σωστή. Η δυνατότητα διαχωρισμού αποδόθηκε στη μέτρηση των διαφορών που παρατηρήθηκαν στη μορφή των φασματικών ζωνών και των σχετικών εντάσεών τους. Επομένως βγήκε το συμπέρασμα ότι τα φάσματα ορατού των μελανιών μπορούν να παρέχουν στους δικανικούς εξεταστές εγγράφων σημαντικά αποτελέσματα σχετικά με την ομοιότητα του μελανιού και αντικειμενικά συμπεράσματα μπορούν να συναχθούν υπό τον όρο ότι ακολουθούνται αυστηρές στατιστικές διαδικασίες. Αυτές οι διαδικασίες μπορούν να αποκαλύψουν χρήσιμες πληροφορίες για τα φασματικά χαρακτηριστικά των μελανιών που διαφέρουν μεταξύ των εμπορικών επωνυμιών, αλλά παραμένουν σχεδόν σταθερά μέσα στην ίδια εμπορική επωνυμία.

2. Πειραματική Πορεία

2.1 Δείγματα, προετοιμασία και μετρήσεις.

Πέντε εμπορικά διαθέσιμες μάρκες (κωδικοποιημένες ως: BI, FC, PE, PI και ST) από στυλό μπλε μελανιών χρησιμοποιήθηκαν για μελέτη. Για κάθε μάρκα, λάβαμε δείγμα μελανιού από δέκα στυλό, με τη βοήθεια μιας βελόνας ανοξείδωτου χάλυβα που χρησιμοποιήθηκε για να διαπεράσει τον τοίχο της πλαστικής δεξαμενής μελανιού και να μεταφέρει το δείγμα στο διαλύτη. Κάθε δείγμα λήφθηκε χρησιμοποιώντας μια διαφορετική βελόνα και η «λεκιασμένη» με μελάνι άκρη της βελόνας εκπλύθηκε σε 15ml απόλυτης αιθανόλης (MERCK 99,8% v/v) περιεχόμενη σε δοκιμαστικούς σωλήνες. Μετά από ανάδευση, τα δείγματα υποβλήθηκαν σε φυγοκέντρηση σε 2000 στροφές ανά λεπτό και η απορρόφηση του υπερκείμενου υγρού μετρήθηκε με φασματοφωτόμετρο JENWAY σε κυψελίδα χαλαζία 1,00cm με απόλυτη αιθανόλη ως τυφλό δείγμα. Το όριο ανίχνευσης ήταν μεταξύ 400-750nm ανά διαστήματα 1 nm.

2.2 Στατιστικές διαδικασίες

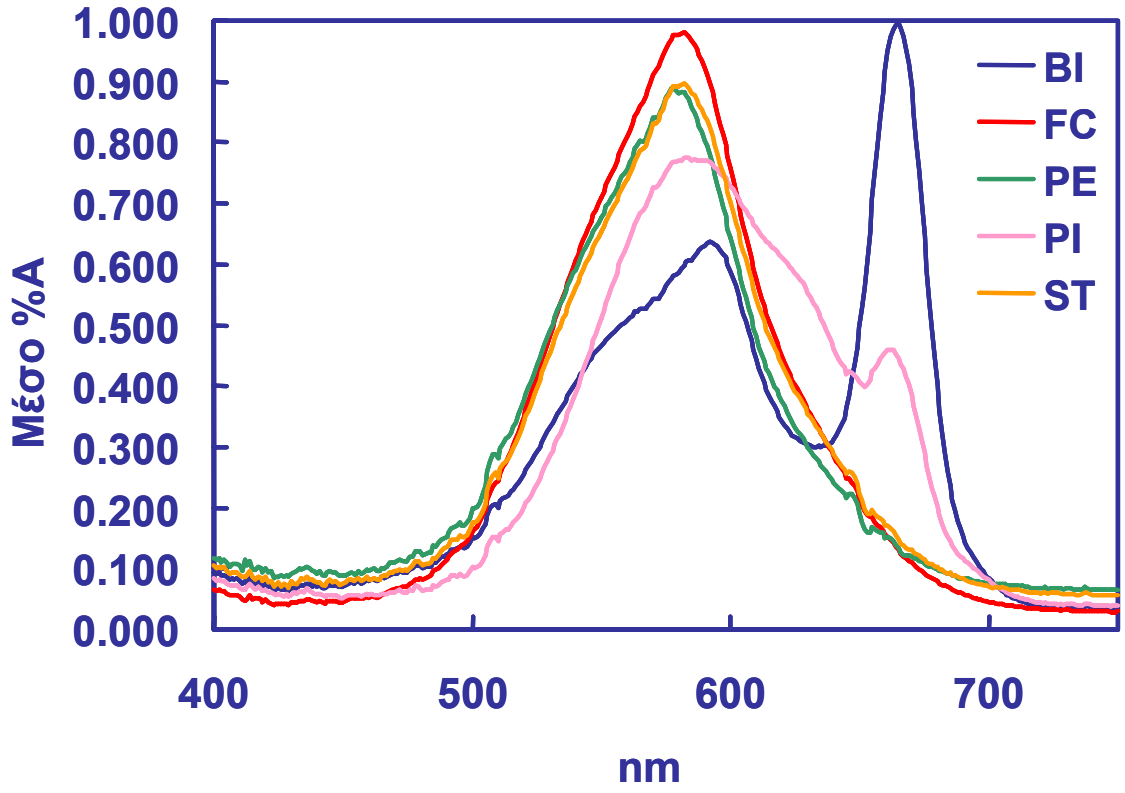
Οι τιμές απορροφητικότητας για κάθε δείγμα διαιρέθηκαν με τη συνολική απορροφητικότητα του κάθε φάσματος και τα αποτελέσματα πολλαπλασιάστηκαν με 100 (κανονικοποίηση). Υπολογίστηκαν οι τιμές του λογαρίθμου της επί τοις εκατό απορρόφησης ($\log_{10}$(%A)) για να εξασφαλιστεί η κανονικότητα των δεδομένων. Όλες οι στατιστικές αναλύσεις εκτελέσθηκαν σε έναν προσωπικό υπολογιστή που έτρεχε το λογισμικό STATISTICA, έκδοση 4.3 για Windows της εταιρίας StatSoft Inc., 1993. Χρησιμοποιήθηκαν διαδοχικά η *ανάλυση κατά συστάδες* (μέθοδος K-Means), η *ανάλυση κυρίων συνιστωσών* (με περιστροφή Varimax) και η *διαχωριστική ανάλυση*.

2.3 Αποτελέσματα και συζήτηση

Η χρήση των %A τιμών αντί των μετρηθέντων τιμών απορρόφησης ήταν απαραίτητη, δεδομένου ότι τα προκύπτοντα φάσματα θα έπρεπε να κανονικοποιηθούν ως προς την συνολική απορρόφηση του δείγματος προκειμένου να είναι συγκρίσιμα. Το μέγεθος των δειγμάτων μελανιού που χρησιμοποιήθηκαν σε αυτήν την μελέτη ήταν εξαιρετικά μικρό και πολύ δύσκολο να ζυγιστεί και επομένως το κρίναμε απαραίτητο να χρησιμοποιήσουμε τον τύπο:

$$\%A_i = \frac{A_i}{\sum_{i=400}^{750} A_i} \times 100\%$$

Αυτός ο μετασχηματισμός υπερνικά το πρόβλημα των διαφορών σε βάρη των δειγμάτων που χρησιμοποιούνται για τα διαλύματα και τα φάσματα γίνονται συγκρίσιμα. Το μέσο φάσμα $\%\overline{A}$ για κάθε στυλό παρουσιάζεται στο σχήμα 1.



Σχήμα 1. Μέσα %A φάσματα απορρόφησης μελανιών

Από τα φάσματα, μπορεί να φανεί ότι τα μελάνια που μελετήθηκαν ενέπεσαν σε δύο κύριες κατηγορίες: τα μελάνια FC, PE, ST κατέδειξαν ένα φάσμα *απλών κορυφών*, ενώ τα μελάνια BI και PI *διπλών κορυφών*. Η πρώτη ζώνη απορρόφησης είχε την κορυφή της στην περιοχή 570-600nm και ήταν κοινή για όλα τα μελάνια. Η ηλεκτρομαγνητική ακτινοβολία αυτών των μηκών κύματος εμφανίζεται *κίτρινη-πορτοκαλί* στο γυμνό μάτι και απορροφάται από τις μπλε-κυανές ουσίες. Η προαναφερθείσα περιοχή επομένως αποδόθηκε στις μπλε βαφές (dyes) και χρωστικές ουσίες (pigments) που περιλήφθηκαν σε όλα τα δείγματα. Η δεύτερη περιοχή, που ήταν παρούσα μόνο στην περίπτωση των δειγμάτων BI και PI, είχε την κορυφή της στη σειρά 660-670nm. Η ακτινοβολία σε αυτήν την περιοχή είναι *πορτοκαλί* στο χρώμα και απορροφάται από τις *γαλαζοπράσινες* ουσίες. Πράγματι, όταν παρατηρούνται με το γυμνό μάτι, τα διαλύματα αυτών των μελανιών είχαν χρώματα που διέφεραν ελαφρώς από τια «καθαρά» μπλε διαλύματα των μελανιών FC, ST και PE. Αν και αυτά τα μέσα φάσματα επέτρεψαν μια εύκολη πρώτη διάκριση των δειγμάτων, τίποτα δεν θα μπορούσε να ειπωθεί για τις διαφορές που παρατηρήθηκαν στα φάσματα των μελανιών FC, ST και PE, δεδομένου ότι αυτά τα

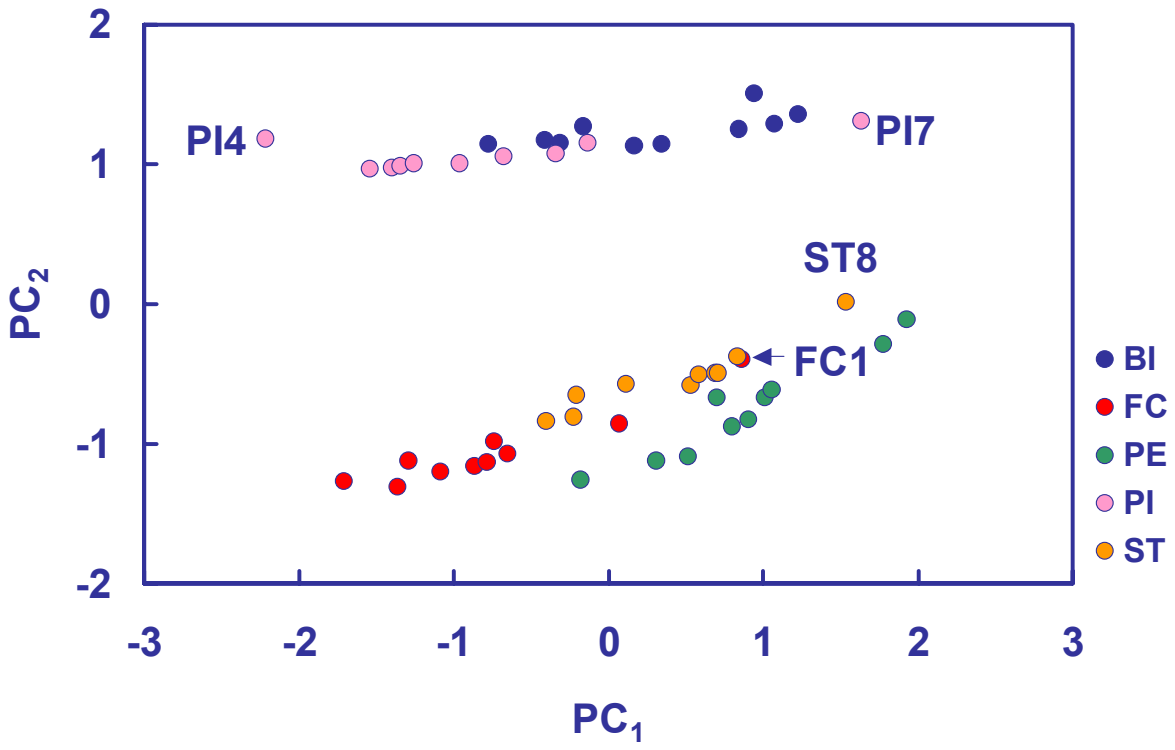
φάσματα αντιπροσωπεύουν μόνο την μεταξύ των ομάδων δειγμάτων διακύμανση, χωρίς όμως τίποτα να γίνεται γνωστό για την διακύμανση των δειγμάτων μέσα στην ίδια την ομάδα. Επομένως, προηγμένες στατιστικές δοκιμές κρίθηκαν απαραίτητες για την καθιέρωση ενός απολύτως αντικειμενικού πρωτοκόλλου διάκρισης.

Δεδομένου ότι και η *Ανάλυση σε Κύριες Συνιστώσες* και η *Διαχωριστική Ανάλυση* είναι παραμετρικές στατιστικές τεχνικές, ήταν απαραίτητο για τα στοιχεία να λογαριθμηθούν ($\log_{10}$(%A)) για να εξασφαλίσει η κανονικότητα. Δεδομένου ότι τα δείγματα για κάθε μάρκα θεωρήθηκαν πως ανήκουν σε διαφορετικούς πληθυσμούς και μόνο δέκα δείγματα από κάθε πληθυσμό ήταν διαθέσιμα, καμία δοκιμή κανονικότητας δεν πραγματοποιήθηκε, δεδομένου ότι αυτές οι δοκιμές είναι γνωστές να έχουν καλά αποτελέσματα για μεγάλα σύνολα δειγμάτων μόνο. Από τώρα και στο εξής, οι τιμές που θα αναφέρονται ως αρχικές μεταβλητές θα είναι οι ($\log_{10}$(%A)).

Το STATISTICA έκδοση 4.3 για Windows δεν επιτρέπει σε περισσότερες από 300 μεταβλητές για να χρησιμοποιηθούν σε οποιαδήποτε από τις αναλύσεις του. Εντούτοις, τα φάσματα μελανιού που καταγράφηκαν στην περιοχή 400-750nm σε διαστήματα 1nm και οι ($\log_{10}$(%A)) μετασχηματισμένες %A τιμές σε κάθε μήκος κύματος επρόκειτο να ληφθούν ως αρχικές μεταβλητές. Αυτό σήμανε ότι 351 μεταβλητές ήταν διαθέσιμες και μερικές από αυτές θα έπρεπε να αφαιρεθούν από το σύνολο δεδομένων για να συμμορφωθούν με τους περιορισμούς το STATISTICA. Αυτού του είδους η μείωση των μεταβλητών μπορεί να επιτευχθεί με την εκτέλεση μιας ανάλυσης κατά συστάδες (K-Means) στις *μεταβλητές ως προς τα αντικείμενα* (δείγματα μελανιού). Σύμφωνα με αυτήν την τεχνική, οι μεταβλητές που φέρνουν παρόμοιες πληροφορίες για τα αντικείμενα αναμένονται να διαμορφώσουν τις συστάδες από τις οποίες οι αντιπροσωπευτικότερες μεταβλητές (αυτές πιο κοντά στα κέντρα των συστάδων) μπορούν να επιλεχτούν. Αν και ο μέγιστος αριθμός συστάδων που μπορεί να διαμορφωθεί από το STATISTICA είναι 50, είμαστε υποχρεωμένοι να υπολογίσουμε μόνο 20 συστάδες και να χρησιμοποιήσουμε τις σχετικές μεταβλητές, δεδομένου ότι ο σχηματισμός 50, 40 ή 30 συστάδων παρήγαγε μεταβλητές που αποδείχθηκε ότι ήταν υψηλής συσχέτισης και όλες οι μετέπειτα προσπάθειες να εφαρμοστεί η ανάλυση κυρίων συνιστωσών οδήγησαν σε ένα μοναδικό πίνακα συσχετισμού.

Αρχικά, η ανάλυση σε κύριες συνιστώσες ήταν απαραίτητη για την αφαίρεση οποιασδήποτε τιμής με μεγάλη απόκλιση (outliers) στα στοιχεία. Η παρουσία outliers μπορεί να έχει επιπτώσεις στα αποτελέσματα της διαχωριστικής ανάλυσης μέσω μιας

υπερεκτίμησης της μεταξύ των δειγμάτων διακύμανσης και της λανθασμένης διατήρησης της *μηδενικής υπόθεσης* για την διαφορά μεταξύ των μέσων τιμών των ομάδων. Στα συστήματα πολλών μεταβλητών, η ανάλυση σε κύριες συνιστώσες μπορεί να βοηθήσει την παρατήρηση των outliers από την προβολή των στοιχείων σε ένα γράφημα μετά από την περιστροφή Varimax των πρώτων δύο κυρίων συνιστωσών (PC_1 και PC_2). Αυτό το γράφημα φαίνεται στο σχήμα 2 από το οποίο μπορεί να φανεί ότι τα δείγματα FC1, PI4, PI7 και ST8 πρέπει να θεωρηθούν outliers και έπρεπε να αφαιρεθούν.



Σχήμα 2. Οι δύο πρώτες κύριες συνιστώσες. Διακρίνονται οι outliers.

Το επόμενο βήμα στη μελέτη που αφορά τις ιδιότητες του διακρίνοντος προτύπου που υπολογίστηκε στη διαχωριστική ανάλυση. Η ανάλυση σε κύριες συνιστώσες εφαρμόστηκε διαδοχικά στο αρχικό σύνολο δεδομένων και κάθε φορά εξήχθησαν όλο και λιγότερες συνιστώσες και περιστράφηκαν. Για κάθε συνιστώσα, η μεταβλητή με την υψηλότερη *φόρτωση* διατηρήθηκε και το προκύπτον νέο υποσύνολο των μεταβλητών χρησιμοποιήθηκε στη διαχωριστική ανάλυση. Αυτή η μέθοδος έχει το πλεονέκτημα της ευκολότερης ερμηνείας των αποτελεσμάτων της διαχωριστικής ανάλυσης, δεδομένου ότι η αναφορά στα αρχικά μήκη κύματος μπορεί

να γίνει άμεσα, αλλά η πολυσυγγραμικότητα είναι ακόμα παρούσα και στη διαχωριστική ανάλυση παρατηρείται υψηλή συσχέτιση των μεταβλητών. Τα χαρακτηριστικά του μοντέλου διάκρισης ελέγχθηκαν επίσης με χρήση των συνιστωσών αντί των αρχικών μεταβλητών. Όταν αυτή η μέθοδος ακολουθείται, η ερμηνεία των αποτελεσμάτων είναι δυσκολότερη δεδομένου ότι πρέπει να εξεταστεί η φόρτωση των αρχικήν μεταβλητών στις συνιστώσες για να αποφασίσει ποιες μεταβλητές αντιπροσωπεύουν αυτές. Εντούτοις, η τεχνική χειρίζεται το πρόβλημα της πολυσυγγραμικότητας αποτελεσματικά, το οποίο οφείλεται στην ορθογώνια φύση των παραγόμενων συνιστωσών.

Για να αποφευχθεί ο παράγοντας τύχη, πρέπει να χρησιμοποιηθούν ως όσο το δυνατόν λιγότερες μεταβλητές κατά την εκτέλεση μιας διαχωριστικής ανάλυσης. Μια χρήσιμη εμπειροτεχνική μέθοδος υπαγορεύει ότι *m* μεταβλητές πρέπει να χρησιμοποιηθούν όταν υπάρχουν *n* αντικείμενα έτσι ώστε:

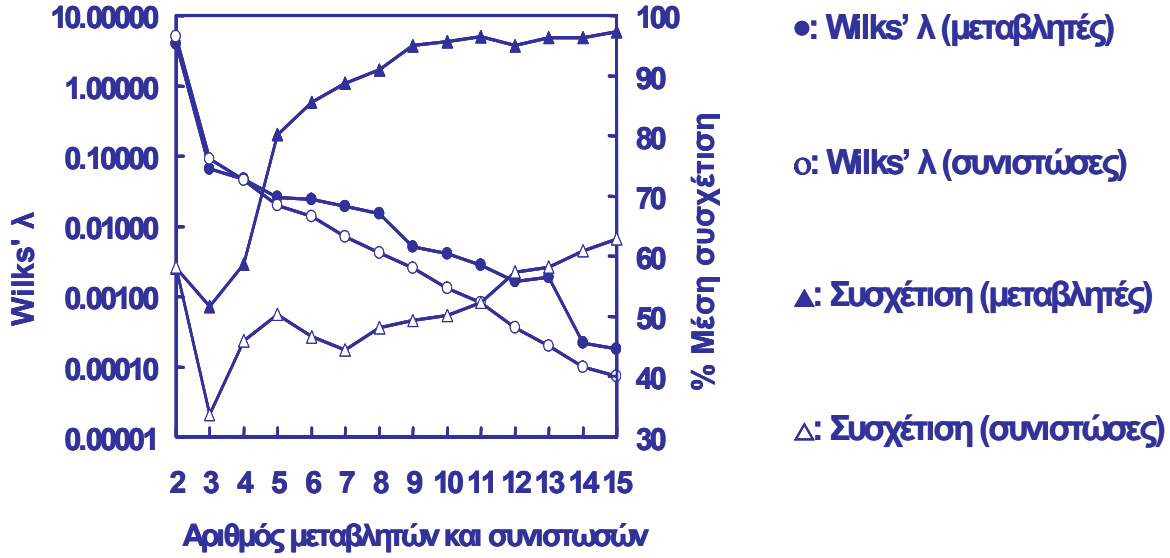
$$m < \frac{n}{3}$$

Λαμβάνοντας τα αφαιρούμενα outliers υπόψη, 46 αντικείμενα παρέμειναν και αυτό σήμανε ότι μόνο μέχρι 15 μεταβλητές (ή συνιστώσες) θα μπορούσαν να εισαχθούν στην διαχωριστική ανάλυση. Τα δύο κύρια κριτήρια (κριτήριο Kaiser και scree plot) που χρησιμοποιήθηκαν στην απόφαση του πληθυσμού των συνιστωσών που θα εξαγάγει η ανάλυση σε κύριες συνιστώσες, λήφθηκαν υπόψη επίσης κατά την ερμηνεία των αποτελεσμάτων με την διαχωριστική ανάλυση. Το scree plot (σχήμα 3) για το σύνολο δεδομένων έδειξε ότι τα πρώτα τρία συστατικά είχαν ιδιοτιμές μεγαλύτερες από το σύνολο (κριτήριο Kaiser) ενώ τα πρώτα τέσσερα ή πέντε συστατικά ικανοποιούσαν το κριτήριο του scree plot.

Σχήμα 3. Scree plot

Οι αλλαγές στις τιμές του *Wilk's λ* ως προς τον αριθμό των αρχικών μεταβλητών ή των χρησιμοποιούμενων συνιστωσών παρουσιάζονται στο σχήμα 4. Η χρήση μιας λογαριθμικής κλίμακας για τον άξονα του Wilk's λ ήταν υποχρεωτική δεδομένου ότι ένα ευρύ φάσμα των τιμών έπρεπε να καλυφθεί. Όπως αναμενόταν θεωρητικά, η μεροληπτική δύναμη του στατιστικού μοντέλου αυξήθηκε καθώς όλο και περισσότερες μεταβλητές (ή συνιστώσες) χρησιμοποιήθηκαν. Η αλλαγή στην κλίση που παρατηρείται στο σημείο που αντιστοιχούσε σε τρεις μεταβλητές (ή συνιστώσες) έδειξε ότι λίγη βελτίωση επιτεύχθηκε με περαιτέρω αύξηση του αριθμού μεταβλητών (ή συνιστωσών). Αυτό το σημείο ήταν σε συμφωνία με το αποτέλεσμα του κριτηρίου Kaiser. Ελαφρώς καλύτερες (μικρότερες) τιμές *Wilk's λ* παρατηρήθηκαν όταν χρησιμοποιήθηκαν οι συνιστώσες αντί των αρχικών μεταβλητών. Συγχρόνως, ο μέσος επί τοις εκατό συσχετισμός των αρχικών μεταβλητών στο διαχωριστικό μοντέλο αυξήθηκε γρήγορα όταν χρησιμοποιήθηκαν περισσότερες από τέσσερις μεταβλητές και έφθασαν τελικά σε ένα «πλατώ» γύρω στο 96%. Αξίζει επίσης να αναφερθεί ότι τέσσερις μεταβλητές είναι ο προβεπόμενος αριθμός μεταβλητών από το κριτήριο του scree plot. Όταν οι κύριες συνιστώσες χρησιμοποιήθηκαν στο

μοντέλο, η ίδια αύξηση παρατηρήθηκε στο σημείο που αντιστοιχούσε σε τέσσερις συνιστώσες, αλλά το τελικό «πλατώ» δεν υπερέβη το 63%. Αυτό αναμενόταν δεδομένου ότι τα κύρια συστατικά είναι ορθογώνια και ασύνδετα εξ ορισμού. Η χρήση των κύριων συστατικών ως νέες «αφανείς» μεταβλητές στη διαχωριστική ανάλυση ήταν επομένως προτιμητέα.



Σχήμα 4. Wilk's λ και συσχετίσεις μεταβλητών και συνιστωσών

Η εξέταση του αριθμού των στατιστικά ασήμαντων (*P*=0,05) μεταβλητών ή συνιστωσών στο διαχωριστικό μοντέλο παρουσίασε ότι όταν εισήχθησαν πέντε ή λιγότερες μεταβλητές ή συνιστώσες στην διαχωριστική ανάλυση, όλες τους ήταν στατιστικά σημαντικές στο επίπεδο *P*=0,05. Πάλι η χρήση των κύριων συνιστωσών αντί των αρχικών μεταβλητών αποδείχθηκε επαρκέστερη. Εν πάση περιπτώσει, το ποσοστό της σωστής ταξινόμησης των στοιχείων στο σύνολο δεδομένων κατάρτισης ήταν 100%, εκτός αν χρησιμοποιήθηκαν μόνο δύο μεταβλητές ή συνιστώσες, κάτι που οδήγησε σε ποσοστό σωστής ταξινόμησης 91,3% και 84,8% αντίστοιχα.

Με βάση τα προαναφερθέντα συμπεράσματα, αποφασίσαμε να χρησιμοποιήσουμε τα αποτελέσματα των πρώτων τριών κύριων συνιστωσών στο μοντέλο της διαχωριστικής ανάλυσης. Αυτό μας επέτρεψε να θυσιάσουμε όσο το δυνατόν λιγότερη διαχωριστική και να αποφύγουμε επίσης πάρα πολλές νέες «αφανείς» μεταβλητές στο μοντέλο. Οι τρεις κύριες συνιστώσες που χρησιμοποιήθηκαν ερμήνευσαν το 96,97% της συνολικής διαφοροποίησης στα αρχικά στοιχεία. Το *Wilk's λ* του υπολογιστικού διαχωριστικού μοντέλου ήταν

$8{,}98 \times 10^{-5}$ και το μοντέλο βρέθηκε για να είναι στατιστικά σημαντικό σε επίπεδο *P*=0,05 με μια μέση μεταβλητή συσχέτιση 33,6%. Η μετά απ' αυτό ταξινόμηση των στοιχείων στο σύνολο δεδομένων κατάρτισης βρέθηκε για να είναι 100% σωστή.

Οι τρεις διαχωριστικές συναρτήσεις (κανονικές ρίζες) που υπολογίστηκαν για το πρότυπο ελέγχθηκαν για τη σημασία τους με τη βοήθεια της δοκιμής x^2 και με διαδοχική αφαίρεση των ριζών μέχρι να παραμείνει μια μόνο ρίζα. Από τις τιμές του *P* στον πίνακα 1 μπορεί να φανεί ότι και οι τρεις ρίζες ήταν στατιστικά σημαντικές στο επίπεδο του *P*=0,05.

Πίνακας 1. Σημαντικότητα των κανονικών ριζών

Αφαιρούμενες ρίζες	**χ^2**	**Κανονικές ρίζες**	**Τιμή P**
0	382.0371	12	<0.001
1^{η}	194.9673	6	<0.001
1^{η}, 2^{η}	27.3722	2	<0.001

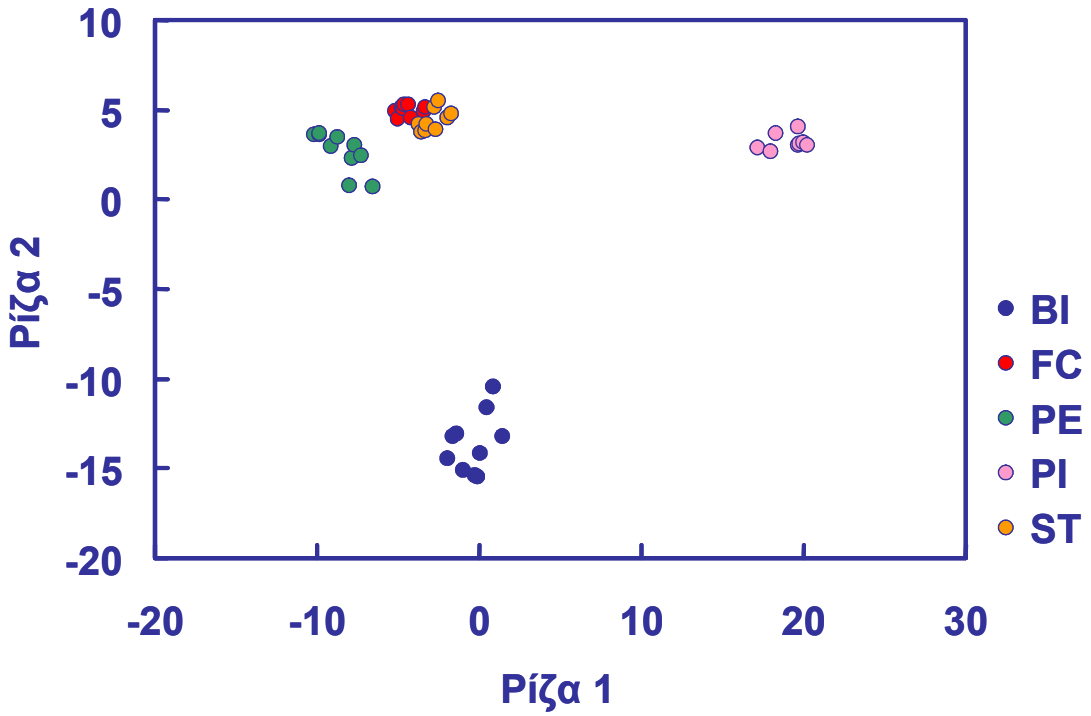
Οι διαχωριστικές συναρτήσεις που υπολογίστηκαν ήταν της μορφής

$$L = b_0 + \sum_{i=1}^{3} b_i PC_i$$

Οι αρχικοί (b_i) και μετασχηματισμένοι (B_i) συντελεστές των κανονικών ριζών δίνονται στον πίνακα 2. Οι αρχικοί συντελεστές χρησιμοποιήθηκαν για τον υπολογισμό των τιμών των κανονικών ριζών (σχήμα 5). Μόνο οι πρώτες δύο ρίζες παρουσιάζονται, δεδομένου ότι εκείνες ήταν αυτές που έπαιξαν σημαντικότερο ρόλο στη διάκριση. Η ρίζα 1 ήταν υπεύθυνη για το διαχωρισμό των PE, [FC-ST-BI] και PI, ενώ η ρίζα 2 περαιτέρω βοήθησε το διαχωρισμό [FC-ST] και BI. Η ρίζα 3, που δεν παρουσιάζεται εδώ, ήταν υπεύθυνη για τον τελικό διαχωρισμό των FC και ST.

Οι μετασχηματισμένοι συντελεστές χρησιμοποιήθηκαν για την εκτίμηση της συμβολής των συνιστωσών στη διάκριση που επιτεύχθηκε από την αντίστοιχη συνάρτηση δι'ακρισης. Στον πίνακα 2 φαίνεται ότι η PC_3 διαδραμάτισε το σημαντικότερο ρόλο στη διάκριση που επιτεύχθηκε από τη ρίζα 1. Από τις στατιστικά

σημαντικές (P<0,05) φορτώσεις των συνιστοσών (πίνακας 3) φαίνεται ότι η PC_3 κυρίως αντιπροσωπεύει την απορρόφηση στα 622nm. Σε εκείνο το μήκος κύματος, τα φάσματα που ανήκουν στην ομάδα PI παρουσιάζουν έναν «ώμο» στο φάσμα απορρόφησης (σχήμα 1) ο οποίος έλειπε από τις άλλες ομάδες. Η διαχωριστική ικανότητα της ρίζας 1 είναι ιδιαίτερα εμφανής στο σχήμα 5 στο οποίο η ομάδα PI βρίσκεται άπω δεξιά της γραφικής παράστασης, και διαχωρίζεται εντελώς από τις άλλες ομάδες. Ο πίνακας 2 επίσης δείχνει ότι η ρίζα 2 επηρεάστηκε κυρίως από τη PC_2. Από τον πίνακα 3 φαίνεται ότι αυτή η συνιστώσα αντιπροσωπεύει την απορρόφηση στην περιοχή 660-686nm και στα 553nm και 570nm. Το σχήμα 1 δείχνει ότι η πρώτη περιοχή αντιστοιχεί στη δεύτερη κορυφή που χαρακτηρίζει τα μελάνια BI και PI, με τις τιμές της %A να είναι σημαντικά υψηλότερες στην περίπτωση των μελανιών BI. Η απορρόφηση στα 553nm και 570nm ήταν επίσης διαφορετική για τα μελάνια BI και αντιστοιχούσε στο ανερχόμενος μέρος της καμπύλης με μια κλίση που ήταν μικρότερη από τις κλίσεις των άλλων καμπυλών. Αυτά τα χαρακτηριστικά γνωρίσματα ήταν υπεύθυνα για το διαχωρισμό της ομάδας BI με τη βοήθεια της ρίζας 2, κάτι που είναι επίσης εμφανές στο σχήμα 5. Ομοίως, στη PC_1 οφείλεται η διάκριση που επιτεύχθηκε από τη ρίζα 3 (πίνακας 2). Αυτή η συνιστώσα αντιπροσώπευε τις τιμές απορρόφησης στην περιοχή 405-444nm και στα 707nm και 728nm (πίνακας 3). Όλα αυτά τα μήκη κύματος αντιστοιχούσαν στις πολύ χαμηλές τιμές απορρόφησης στα άκρα φασμάτων (σχήμα 1). Αν και μη παρατηρήσιμες με γυμνό μάτι, αυτές οι λεπτές διαφορές στην απορρόφηση ήταν αυτές που πέτυχαν το διαχωρισμό της μικρής αλληλεπικάλυψης των ομάδων FC και ST.



Σχήμα 5. Κανονικές ρίζες της Διαχωριστικής ανάλυσης

Πίνακας 2. Ακατέργαστοι και τυποποιημένοι συντελεστές διαχωριστικών συναρτήσεων

	Ρίζα 1	Ρίζα 2	Ρίζα 3
Αρχικοί συντελεστές			
b_0	7.06 x 10^{-15}	5.70 x 10^{-15}	-1.33 x 10^{-16}
b_1	-1.2365	0.2430	1.3203
b_2	3.8685	-6.6783	0.1139
b_3	-8.4158	-3.1056	-0.1416
Μετασχηματισμένοι συντελεστές			
B_1	-0.9193	0.1807	0.9816
B_2	0.5633	-0.9724	0.0166
B_3	-1.1575	-0.4271	-0.0195

Πίνακας 3. Φόρτωση στατιστικά σημαντικών συνιστωσών

PC_1			PC_2			PC_3		
Αρχική μεταβλητή	**Φόρτωση**	**Τιμή *P***	**Αρχική μεταβλητή**	**Φόρτωση**	**Τιμή *P***	**Αρχική μεταβλητή**	**Φόρτωση**	**Τιμή *P***
$\log_{10}(A_{418})$	0.9614	<0.001	$\log_{10}(A_{677})$	0.9980	<0.001	$\log_{10}(A_{495})$	0.7999	<0.001
$\log_{10}(A_{405})$	0.9551	<0.001	$\log_{10}(A_{660})$	0.9858	<0.001	$\log_{10}(A_{511})$	0.7374	<0.001
$\log_{10}(A_{443})$	0.9513	<0.001	$\log_{10}(A_{686})$	0.9748	<0.001	$\log_{10}(A_{482})$	0.7106	<0.001
$\log_{10}(A_{444})$	0.9483	<0.001	$\log_{10}(A_{696})$	0.7583	<0.001	$\log_{10}(A_{524})$	0.6330	<0.001
$\log_{10}(A_{707})$	0.9296	<0.001	$\log_{10}(A_{643})$	0.6645	<0.001	$\log_{10}(A_{537})$	0.4773	0.001
$\log_{10}(A_{728})$	0.9092	<0.001	$\log_{10}(A_{495})$	-0.3392	0.021	$\log_{10}(A_{464})$	0.4599	0.001
$\log_{10}(A_{464})$	0.8742	<0.001	$\log_{10}(A_{511})$	-0.5995	<0.001	$\log_{10}(A_{643})$	-0.7040	<0.001
$\log_{10}(A_{482})$	0.6910	<0.001	$\log_{10}(A_{524})$	-0.7661	<0.001	$\log_{10}(A_{622})$	-0.9469	<0.001
$\log_{10}(A_{696})$	0.5609	<0.001	$\log_{10}(A_{537})$	-0.8754	<0.001			
$\log_{10}(A_{495})$	0.4881	0.001	$\log_{10}(A_{570})$	-0.9760	<0.001			
$\log_{10}(A_{511})$	0.3059	0.039	$\log_{10}(A_{553})$	-0.9803	<0.001			

Η γενική συμβολή των κύριων συνιστωσών στο διαχωριστικό μοντέλο παρουσιάζεται στον πίνακα 4. Η συμβολή ακολούθησε τη σειρά PC_3>PC_2>PC_1 και όλες οι συνιστώσες βρέθηκαν για να είναι στατιστικά σημαντικές στο επίπεδο *P*=0,05. Αυτή η σειρά είναι σε τέλεια συμφωνία με αυτό που αναφέρθηκε ανωτέρω για το ρόλο και τη φύση κάθε διαχωριστικής συνάρτησης.

Πίνακας 4. Συμβολή των κυρίων συνιστωσών στο διαχωριστικό μοντέλο

Συνιστώσα	**Μερικό *Wilks' λ***	***P*-value**	**Συσχέτιση (%)**
PC_1	0.2735	<0.001	45.7
PC_2	0.0153	<0.001	20.8
PC_3	0.0113	<0.001	34.3

Η διαχωριστική ανάλυση ολοκληρώθηκε με τον υπολογισμό των αποκαλούμενων συναρτήσεων ταξινόμησης. Αυτές οι συναρτήσεις επιτρέπουν τη μετέπειτα ταξινόμηση των στοιχείων του συνόλου δεδομένων κατάρτισης ή την ταξινόμηση των νέων στοιχείων. Για να ταξινομήσουμε νέα δείγματα, πρώτον θα

έπρεπε να υπολογίσουμε τις αντίστοιχες τιμές των συνιστωσών και να εισαγουμει έπειτα τα αποτελέσματα στις συναρτήσεις ταξινόμησης. Η συνάρτηση που θα παρήγαγε το υψηλότερο αποτέλεσμα θα έδειχνε την ομάδα στην οποία το νέο δείγμα άνηκε. Πέντε συναστήσεις ταξινόμησης υπολογίστηκαν (μια για κάθε ομάδα) και ήταν της μορφής:

$$C = b_0' + \sum_{i=1}^{3} b_i' PC_i$$

Οι συντελεστές τους δίνονται στον πίνακα 5 και η μετέπειτα ταξινόμηση του συνόλου δεδομένων κατάρτισης με τη βοήθεια αυτών των συναρτήσεων βρέθηκε να είναι 100% σωστή (καμία λάθος ταξινόμηση δεν εμφανίστηκε).

Πίνακας 5. Συντελεστές συναρτήσεων ταξινόμησης.

Συντελεστές	**Ομάδες μελανιών**				
	BI	**FC**	**PE**	**PI**	**ST**
b_0'	-94.3730	-24.9465	-41.5852	-189.4998	-15.4220
b_1'	-3.1026	4.4046	12.4671	-22.5842	5.2652
b_2'	89.6619	-50.2556	-50.4302	52.7334	-40.2094
b_3'	45.1379	21.7563	63.0824	-170.7646	9.7896

ΣΥΜΠΕΡΑΣΜΑΤΑ

Αυτή η μελέτη έδειξε ότι άριστη διάκριση (100% σωστή ταξινόμηση του συνόλου δεδομένων κατάρτισης) μεταξύ μελανιών 5 διαφορετικών εμπορικών εταιριών θα μπορούσε να επιτευχθεί από την εξέταση των φασμάτων ορατού των αιθανολικών διαλυμάτων του μελανιού και της εφαρμογής ενός πολυμεταβλητού πρωτοκόλλου χημειομετρίας για τη μελέτη των αναλυτικών δεδομένων. Ένα διαχωριστικό μοντέλο υπολογίστηκε με Wilks' λ ίσο με $8,98 \times 10^{-5}$ και βρέθηκε να είναι στατιστικά σημαντικό στο επίπεδο P=0,05. Τα προβλήματα της συγραμμικότητας των δεδομένων αντιμετωπίστηκαν αποτελεσματικά με τη βοήθεια της *Ανάλυσης σε Κύριες Συνιστώσες*, η οποία χρησιμοποιήθηκε για τον υπολογισμό τριών νέων μεταβλητών (κύριες συνιστώσες) που χρησιμοποιήθηκαν στη *Διαχωριστική Ανάλυση*. Η μέση συσχέτιση των μεταβλητών αυτών στο διαχωριστικό μόντέλο ήταν 33,6%. Η μελέτη έδειξε ότι η διάκριση των δειγμάτων βασίστηκε σε διαφορές που αφορούσαν (α) τη μορφή της πρώτης φασματικής ζώνης, η οποία ήταν παρούσα σε όλες τις περιπτώσεις, (β) την ένταση της δεύτερης ζώνης, η οποίο εμφανίστηκε μόνο στην περίπτωση δύο μελανιών και (γ) την απορρόφηση των μελανιών στα άκρα των φασμάτων. Πιστεύουμε ότι μια περαιτέρω μελέτη του συστήματος που περιλαμβάνει τις μη καταστρεπτικές τεχνικές (ενδεχομένως φασματοσκοπία, συντελεστή ανάκλασης, κτλ) ακολουθούθμενες από μια παρόμοια στατιστική προσέγγιση θα επέτρεπε στους δικανικούς εξεταστές εγγράφων να συναγάγουν αντικειμενικά συμπεράσματα σχετικά με την ομοιότητα των μελανιών που χρησιμοποιήθηκαν στην εγγραφή των διαφόρων τμημάτων του εγγράφου. Αυτή η μέθοδος, που είναι απλή και σχετικά ανέξοδη, συνδυασμένη με τις τεχνικές που υπάρχουν ήδη, θα μπορούσε να αποτελέσει τη βάση μιας ιεραρχίας σύμφωνα με την οποία τα δείγματα θα μπορούσαν πρώτα να εξεταστούν με το γυμνό μάτι (πιθανώς κάτω από τις εναλλακτικές πηγές φωτός) και σε περίπτωση που καμία διαφορά δεν θα παρατηρούνταν θα μπορούσε να συνεχίσει με πιο σύνθετες μεθόδους ανάλυσης. Εν πάση περιπτώσει, η ανάγκη για χρήση καθορισμένων με σαφήνεια στατιστικών πρωτοκόλλων παραμένει προφανής εφόσον η δικανική επιστήμη βασίζεται στην αντικειμενικότητα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Ι. Ντζούφρας, Δ. Καρλής, *Στοιχεία Πολυμεταβλητής Ανάλυσης Δεδομένων*, Εκδόσεις Πανεπιστημίου Αιγαίου, Χίος 2001, 39-158.
2. StatSoft Inc, *Electronic Statistics Textbook*, ©1984-2002 http://www.statsoft.com/textbook/stathome.html
3. S. Balakrishnama, A. Ganapathiraju, *Linear Discriminant Analysis – A Brief Tutorial*, Mississippi State University
4. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, California, 1990.
5. S. Axler, *Linear Algebra Done Right*, Springer-Verlag New York Inc., New York, New York, 1995.
6. C. Vogt, J. Vogt, A, Becker, E. Rohde, *Separation, comparison and identification of fountain pen inks by capillary electrophoresis with UV-visible and fluorescence detection and by proton-induced X-ray emission*, J. Chromatogr. A 781 (1997) 391-405.
7. C. Roux, M. Novotny, I. Evans, C. Lennard, *A study to investigate the evidential value of blue and black ball-point pen inks in Australia*, Forensic Sci. Int. 101 (1999) 167-176.
8. J.A. Zlotnick, F.P. Smith, *Chromatographic and electrophoretic approaches in ink analysis*, J. Chromatogr. B 733 (1999) 265-272.
9. R. Saferstein, *Criminalistics: An Introduction to Forensic Science*, seventh ed., Prentice Hall, New Jersey, 2001, p. 468.
10. J.W. Brackett Jr., L.W. Bradford, *Comparison of ink writing on documents by means of paper chromatography*, J. Crim. Law Crim. Police Sci. 43 (1952) 530-539.
11. C. Brown, P.L. Kirk, *Horizontal paper chromatography in the identification of ball-point pen inks*, J. Crim. Law Crim. Police Sci. 45 (1954) 334-339.
12. D.A. Crown, J.V.P. Conway, P.L. Kirk, *Differentiation of blue ball-point pen inks*, J. Crim. Law Crim. Police Sci. 52 (1961) 338-343.

13. P.S. Raju, R.C. Banerjee, N.K. Lyengar, *Comparison of inks by paper chromatography*, J. Forensic Sci. 8 (1963) 268-285.
14. M. Lederer, M. Schudel, *Adsorption chromatography on cellulose V: A simple chromatographic system for the identification of inks*, J. Chromatogr. 475 (1989) 451-456.
15. G.R. Nakamura, S.C. Shimoda, *Examination of micro-quantity of ball-point inks from documents by thin-layer chromatography*, J. Crim. Law Crim. Police Sci. 56 (1965) 113-118.
16. K.W. Smalldon, *Comparison of ink dyestuffs using minimal quantities of writing*, J. Forensic Sci. Soc. 9 (1969) 151-152.
17. R.L. Brunelle, M.J. Pro, *A systematic approach to ink identification*, J. Assoc. Off. Anal. Chem. 55 (1972) 823-826.
18. R.L. Kuranz, *Technique for the separation of ink dyestuffs with similar Rf values*, J. Forensic Sci. 19 (1974) 852-854.
19. R.S. Verma, K.N. Prasad, G.J. Misra, *Thin-layer chromatographic analysis of fiber-tip pen inks*, Forensic Sci. Int. 13 (1979) 65-70.
20. A. Siouffi, G. Guiochon, *Use of reversed-phase thin-layer chromatography for the identification of black inks from board felt markers and ball-point pens*, J. Chromatogr. 209 (1981) 441-445.
21. R.L. Kuranz, *Technique for transferring ink from a written line to a thin-layer chromatographic sheet*, J. Forensic Sci. 31 (1986) 655-657.
22. G. Matysik, E. Soczewinski, *On-line extraction and preconcentration of solid samples in equilibrium sandwich chambers for thin-layer chromatography: Analysis of ink from ball-point pens*, J. Chromatogr. 355 (1986) 363-366.
23. V.N. Aginsky, *Forensic examination of 'slightly soluble' ink pigments using thin-layer chromatography*, J. Forensic Sci. 38 (1993) 1131-1133.
24. J.A. Lewis, *Thin-layer chromatography of writing inks: quality control considerations*, J. Forensic Sci. 41 (1996) 874-877.
25. D. Doud, *Chromatographic analysis of inks*, J. Forensic Sci. 3 (1958) 486-492.
26. J.A. Tappolet, *High-performance thin-layer chromatography: Its application to the examination of writing inks*, Forensic Sci. Int. 22 (1983) 99-109.
27. R.N. Totty, M.R. Ordidge, L.J. Onion, *Comparison of the use of visible microspectrometry and high performance thin-layer chromatography for the*

discrimination of aqueous inks used in porous tip and roller ball pens, Forensic Sci. Int. 28 (1985) 137-144.

28. K.M. Varshney, T. Jettappa, V.K. Mehrotra, T.R. Baggi, *Ink analysis from typed script of electronic typewriters by high performance thin layer chromatography*, Forensic Sci. Int. 72 (1995) 107-115.
29. R.M. Kevern, *Infrared luminescence from thin layer chromatograms of inks*, J. Forensic Sci. Soc. 13 (1973) 25-28.
30. R.D. Blackledge, M. Iwan, *Differentiation between inks of the same brand by infrared luminescence photography of their thin-layer chromatography*, Forensic Sci. Int. 21 (1983) 165-173.
31. V.N. Aginsky, *Comparative examination of inks by using instrumental thin-layer chromatography and microspectrophotometry*, J. Forensic Sci. 38 (1993) 1111-1130.
32. L.F. Colwell, B.L. Karger, *Ball-point pen ink examination by high pressure liquid chromatography*, J. Assoc. Off Anal. Chem. 60 (1977) 613-618.
33. A.H. Lyter, *Examination of ball-point pen ink by high pressure liquid chromatography*, J. Forensic Sci. 27 (1982) 154-160.
34. P.C. White, B.B. Wheals, *Use of a rotating disc multiwavelength detector operating in the visible region of the spectrum for monitoring ball pen inks separated by high-performance liquid chromatography*, J. Chromatogr. 303 (1984) 211-216.
35. I.R. Tebbett, C. Chen, M. Fitzgerald, L. Olson, *The use of HPLC with multiwavelength detection for the differentiation of non ball pen inks*, J. Forensic Sci. 37 (1992) 1149-1157.
36. A. Löfgren, J. Andrasko, *HPLC analysis of printing inks*, J. Forensic Sci. 38 (1993) 1151-1160.
37. J.W. Thompson, *Identification of ink by electrophoresis*, J. Forensic Sci. Soc. 7 (1967) 199-202.
38. H.W. Moon, *Electrophoretic identification of felt-tip pen inks*, J. Forensic Sci. 25 (1980) 146-149.
39. S. Fanali, M. Schudel, *Some separations of black and red water-soluble fiber-tip pen inks by capillary zone electrophoresis and thin-layer chromatography*, J. Forensic Sci. 36 (1991) 1192-1197.

40. E. Rohde, A.C. McManus, C. Vogt, W.R. Heineman, *Separation and comparison of fountain pen inks by capillary electrophoresis*, J. Forensic Sci. 42 (1997) 1004-1011.
41. J.A. Zlotnick, F.P. Smith, *Separation of some black roller-ball pen inks by capillary electrophoresis: preliminary data*, Forensic Sci. Int. 92 (1998) 269-280.
42. A. Giles, *The forensic examination of documents, in: P. White (Ed.), Crime Scene to Court: The Essentials of Forensic Science, The Royal Society of Chemistry*, Cambridge, 1998, pp. 123–125.
43. N.C. Thanasoulias, E.T. Piliouris, M.S.E. Kotti, N.P. Evmiridis, *Application of multivariate chemometrics in forensic soil discrimination based on the UV-Vis spectrum of the acid fraction of humus*, Forensic Sci. Int. 130 (2002) 73-82.
44. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, *Chemometrics: A Textbook*, Elsevier, Amsterdam, 1988, pp. 407–409.
45. H.F. Kaiser, *The application of electronic computers to factor analysis*, Educ. Psychol. Meas. 20 (1960) 141–151.
46. R.B. Cattell, *The scree test for the number of factors*, Multivar. Behav. Res. 1 (1966) 245–276.

ELSEVIER

Forensic Science International 138 (2003) 75–84

Forensic Science International

www.elsevier.com/locate/forsciint

Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their Vis spectra

Nicholas C. Thanasoulias*, Nikolaos A. Parisis, Nicholaos P. Evmiridis

Laboratory of Analytical Chemistry, Department of Chemistry, University of Ioannina, University Campus, 451 10 Ioannina, Greece

Received 16 April 2003; received in revised form 24 August 2003; accepted 25 August 2003

Abstract

Fifty blue ball-point pen inks of five different brands were examined on the basis of the Vis spectrum of their ethanolic solutions with a view to achieving good discrimination between them. Samples were dissolved in absolute ethanol and their absorbance values in the range of 400–750 nm, after appropriate transformations, were used as variables in the multivariate statistical techniques of cluster analysis (CA), principal component analysis (PCA) and discriminant analysis (DA). These techniques were used successively so that an effective and meaningful discriminant model was calculated in the final step. The initial 351 variables ($\log_{10}$ transformed ink absorption values at each wavelength) were subjected to a *K*-means CA over the objects (samples) and only 20 variables were retained. Principal component analysis was used to detect any outliers (four samples were removed) and the remaining samples were re-subjected to PCA to decide how many variables to enter into DA and whether original variables or components should be used. It was found that the first three principal components (in accordance with the Kaiser criterion) were good descriptors of the 20 original variables (96.97% of the data variance was explained) and their use as latent variables in DA lead to low average variable redundancy (33.6%) in the discriminant model. The calculated model had a Wilks' λ of 8.98×10^{-5} and was statistically significant at the $P = 0.05$ level. The post hoc classification of the training dataset was 100% correct. From the DA results and the component loadings it was found that discrimination was achieved on the basis of differences in the shape of the absorption bands as well as their relative intensities. The method was therefore deemed appropriate for supporting exclusionary forensic purposes.

Keywords: Forensic chemistry; Forensic document examination; Principal component analysis; Discriminant analysis; Vis spectrophotometry; Blue ball-point pen ink

1. Introduction

Ink analysis is an important forensic procedure that can reveal useful information about questioned documents. Most of its applications regard the detection and confirmation of alterations to documents with significant financial value such as insurance claims, wills, contracts and tax returns. These modifications can be confirmed by comparison of the inks used to produce the questioned document or estimation of the time at which the various sections of the document were written [1]. It is therefore evident that there is a great need for the development of instrumental methods that will allow an in-depth examination of the inks used to produce a document and at the same time rigid statistical protocols are necessary to be followed so that conclusions regarding ink similarity can be drawn on an objective basis at predefined confidence levels.

The aforementioned needs rise from the fact that documents are rather complex systems that consist of primary and secondary materials [2]. Primary materials comprise the document support (e.g. paper, cardboard, polymer, etc.) and the text (e.g. ink deposits, carbon copies, photocopier toner, pencil, etc.). Secondary materials usually appear on a document as a result of corrections and manipulations and include correcting materials, erasure residues, adhesives, stains, fingerprints, etc.

* Corresponding author. Tel.: +30-26510-98399; fax: +30-26510-44831.
E-mail address: nthanas@cc.uoi.gr (N.C. Thanasoulias).

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2003.08.014

Modern inks contain a plethora of substances that aim to improve the ink characteristics [1,3]. Obviously, the most important component is the coloring material. This comes in the form of dyes, pigments or a combination of both. Dyes can be acidic or basic and are soluble in the liquid body of the ink that is also known as the vehicle. On the other hand, pigments are finely ground multimolecular granules that are insoluble in the vehicle. The vehicle, whose composition affects the flowing and drying characteristics of the ink, can consist of oils, solvents and resins. Another class of substances is used to finely tune the characteristics of the ink. These substances include driers, plasticizers, waxes, greases, soaps and detergents.

The techniques regarding the analysis of inks can be divided into destructive and non-destructive ones with regard to the changes they bring about to the questioned document. In destructive methods, a portion of the ink has to be removed from the document prior to the analysis [4]. On the other hand, non-destructive methods involve the observation of ink on the document by means of a reflectance technique that allows the recording of the ink spectral characteristics without removing the sample from its support.

Chromatographic and electrophoretic methods have always been favored by scientists as far as the analysis of inks is concerned. This preference stems from the fact that inks are rather complex mixtures that require separation of their constituents if good discrimination is to be achieved. Paper chromatography has been among the oldest destructive methods employed in ink analysis and has been used especially for organic dye based inks [5–7] since the older iron-gallotanate inks are difficult to separate by chromatography. Thin layer chromatography (TLC) [2,8–10] and its variants, including disk chromatography [11] and high performance thin layer chromatography (HPTLC) [12–14], have gradually replaced paper chromatography and have proved to be a satisfactory equivalent for separating ink components. Observation of thin layer chromatograms under alternative light sources, the use of infrared luminescence and microspectrophotometry have also been employed in an attempt to achieve better characterization of the TLC bands [15–17]. Although TLC remains the preferred method for ink analysis due to its low cost and relative simplicity, high performance liquid chromatography (HPLC) has been used as an alternative destructive technique that has the advantage of higher resolution and is also capable of detecting colorless components in the ink matrix [18–20]. Characterization of inks has also been achieved by means of traditional electrophoretic methods, but with increased equipment cost and a demand for larger samples [21,22]. Better results in terms of resolution, quantitation and analysis time have been achieved with the use of capillary electrophoresis (CE) [1,23–25].

Non-destructive techniques, although being the most useful ones with regard to document integrity, have not been developed and exploited thoroughly. The observation of documents under alternative light sources with the naked eye is probably the most commonly used method [26]. Problems associated with the subjectivity of the human eye have been surpassed with the use of suitable detectors that can measure the reflected radiation from the samples at different wavelengths, thus offering further information regarding the nature and composition of the ink [2].

Although much research has been carried out for the development of efficient analytical methods regarding the composition of inks, chemometrics is an area that has not been extensively used to explore and support the analytical results. Multivariate chemometrics in particular is a powerful tool when dealing with multi-component systems and allows the extraction of maximum information from complicated datasets. Since forensic science is a discipline that must draw its conclusions on a purely objective basis whenever this is possible, it is mandatory for forensic scientists to follow rigid statistical protocols in reaching decisions regarding experimental data. Therefore, in our present work, we have tried to explore the usefulness of multivariate chemometrics in the discrimination of blue ball-point pen inks. This was achieved by extracting ink dyes in ethanol, recording the Vis spectra of the extracts and applying cluster analysis (CA), principal component analysis (PCA) and discriminant analysis (DA) to the spectral data. Although the use of multivariate chemometrics with UV-Vis spectra is usually avoided, as these spectra show broad bands that cause multicollinearity problems, we have followed a statistical protocol presented in previously published work [27] that allowed us to decide on the number and quality of the variables (original ones or principal components) to be used so that an effective discrimination of the inks could be achieved. Readers who are not familiar with the statistical techniques presented herein are referred to Appendix A at the end of the paper where the fundamentals of cluster analysis, principal component analysis and discriminant analysis are given.

2. Materials and methods

2.1. Samples, their preparation and measurements

Five commercially available brands of blue ball-point pen inks (coded as: BI, FC, PE, PI and ST) were used for the study. For each brand, 10 pens of the same batch were sampled once by means of a stainless steel needle that was used to penetrate the wall of the plastic ink reservoir and transfer a small portion of the sample (less than 1 mg) into the solvent. Each pen was sampled using a different needle and the ink stained tip of the needle was submerged in a test tube containing 10.0 ml absolute ethanol (MERCK, 99.8% v/v). After part of the ink had dissolved in the solvent, the solution was agitated and centrifuged at 2000 rpm for 10 min and the absorbance of the supernatant liquid was measured on a JENWAY 6405 UV-Vis spectrophotometer in

a 1.00 cm quartz cell against absolute ethanol as the blank. The scanning range was 400–750 nm at 1 nm intervals.

2.2. Statistical procedures

The absorbance values for each sample were divided by the total absorbance resulting from each spectrum and the results were multiplied by 100. The $\log_{10}(\%A)$ values were calculated to ensure normality of the data. All statistical analyses were performed on a personal computer running STATISTICA Version 4.3 for Windows by StatSoft Inc., 1993. The modules used were cluster analysis, factor analysis (FA) and discriminant analysis. Varimax rotation was used in PCA.

3. Results and discussion

The use of $\%A$ values instead of raw absorbance data was necessary since the resulting spectra would have to be normalized per unit mass of sample if they were to be comparable. However, ink samples of the size used in this study were extremely small (less than 1 mg) and very difficult to weigh and we therefore judged it necessary to use the formula:

$$\%A_k = \frac{A_k}{\sum_{i=400}^{750} A_i} \times 100$$

This transformation overcomes the problem of differences in the weights of the samples used to prepare the solutions and the spectra become comparable. The average spectrum for each pen brand (as $\%\bar{A}$ versus λ) is shown in Fig. 1. Each spectrum represents the average absorption of inks coming from the same batch. The wavelength range chosen in the study was that of the visible region in an attempt to imitate the response of the human eye to the coloring constituents of inks. From the spectra, it can be seen that the inks studied fell into two main categories: inks FC, PE, ST demonstrated a single-peaked spectrum, whereas inks BI and PI possessed double-peaked spectra. The first absorption band had its peak in the range of 570–600 nm and was common for all inks. Electromagnetic radiation of these wavelengths appears yellow–orange to the naked eye and is absorbed by blue–cyan substances. The band mentioned above was therefore attributed to the blue dyes and pigments contained in all samples. The second band, which was present only in the case of samples BI and PI, had its peak in the range of 660–670 nm. Radiation in this range is orange-red in color and is absorbed by blue–green substances. Indeed, when observed with the naked eye, the solutions of these inks had colors that differed slightly from those of inks FC, ST and PE. Although these average spectra allowed a rough discrimination of the samples, nothing could be said about the differences observed in the spectra of inks FC, ST and PE as these spectra represented only the between sample variance with nothing being known about the within sample variance. Therefore, advanced statistical tests were judged necessary for the establishment of a completely objective discriminating protocol.

As both PCA and DA are parametric statistical techniques, it was necessary for the data ($\%A$ values) to be $\log_{10}$ transformed to ensure normality. Since samples for each brand were considered to belong to different populations and only 10 samples from each population were available, no normality tests were carried out, as these tests are known to work well for large sample sizes only. From now on, the $\log_{10}(\%A)$ values will be referred to as the original variables.

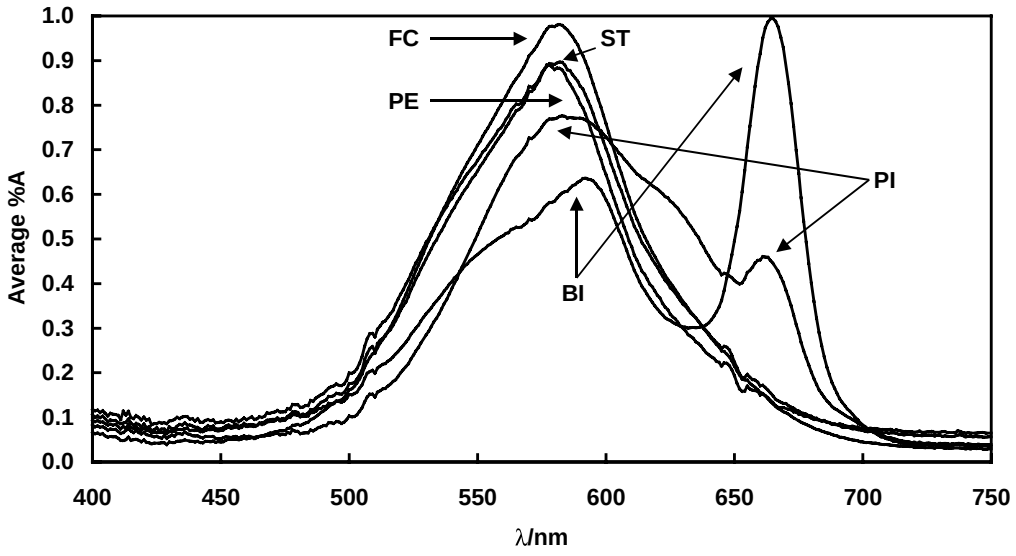


Fig. 1. Average $\%A$ vs. λ for all pen brands.

STATISTICA Version 4.3 for Windows does not allow more than 300 variables to be used in any of its analyses. However, the ink spectra were recorded in the range of 400–750 nm at 1 nm intervals and the $\log_{10}$ transformed $\%A$ values at every wavelength were to be taken as the original variables. This meant that 351 variables were available and some of them would have to be removed for the dataset to comply with the STATISTICA limitations. Feature reduction of this kind can be achieved by performing a cluster analysis (K-means) on the variables over the objects (ink samples) [28]. According to this technique, variables carrying similar information about the objects are expected to form clusters from which the most representative variables (the ones closest to the cluster centroids) can be chosen. Although the maximum number of clusters that can be formed by STATISTICA is 50, we were able to calculate only 20 clusters and use the relevant variables, since the formation of 50, 40 or 30 clusters yielded variables that proved to be highly correlated and all subsequent attempts to run a PCA resulted in a singular correlation matrix.

In the first place, PCA was necessary for the removal of any outliers in the data. The presence of outliers can affect the results of DA through an overestimation of the within sample variance, something that can reduce the effectiveness of the discriminant model. Although the deletion of some data may appear to introduce bias in the analysis, in this case, such a procedure was necessary since for each brand all 10 pens were of the same batch and their spectra were expected to come from the same population. In multivariate systems, PCA can aid the observation of outliers by projection of the data on a plane after Varimax rotation of the first two extracted components (PC_1 and PC_2). This plane is shown in Fig. 2 from which it can be seen that samples FC1, PI4, PI7 and ST8 should be considered outliers and their existence was attributed to measurement problems. These outliers were removed from the dataset.

The next step in the study regarded the quality of the discriminant model calculated in DA. Principal component analysis was successively applied to the original dataset and each time fewer and fewer components were extracted and rotated. For each component, the variable with the highest loading was retained and the resulting new subset of variables was used in DA. This method has the advantage of easier interpretation of the DA results since reference to the original wavelengths can be made directly, but multicollinearity is still present and high variable redundancy is observed in DA. The characteristics of the discriminant model were also checked when components were used instead of the original variables. When this method is followed, the interpretation of the results is more difficult as one has to look at the component loadings to decide which variables they represent. However, the technique handles the problem of multicollinearity effectively due to the orthogonal nature of the extracted components.

To avoid capitalizing on chance, one has to use as few variables as possible when running a DA. A useful rule of thumb [28] dictates that m variables should be used when n objects exist so that:

$$m < \frac{n}{3}$$

Taking the removed outliers into account, 46 objects remained and that meant that only up to 15 variables (or components) should be entered into the DA module. The two main criteria (Kaiser criterion [29] and the scree test [30]) used in deciding how many components to extract in PCA were also taken into account when interpreting the DA results. The scree plot (Fig. 3) for the given dataset showed

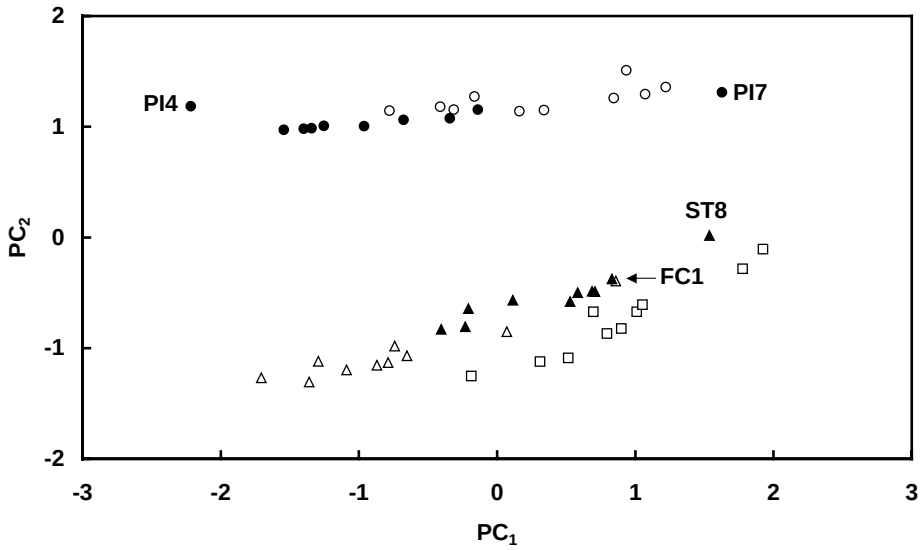


Fig. 2. Principal component graph for the detection of outliers ((○) BI, (△) FC, (□) PE, (●) PI, (▲) ST).

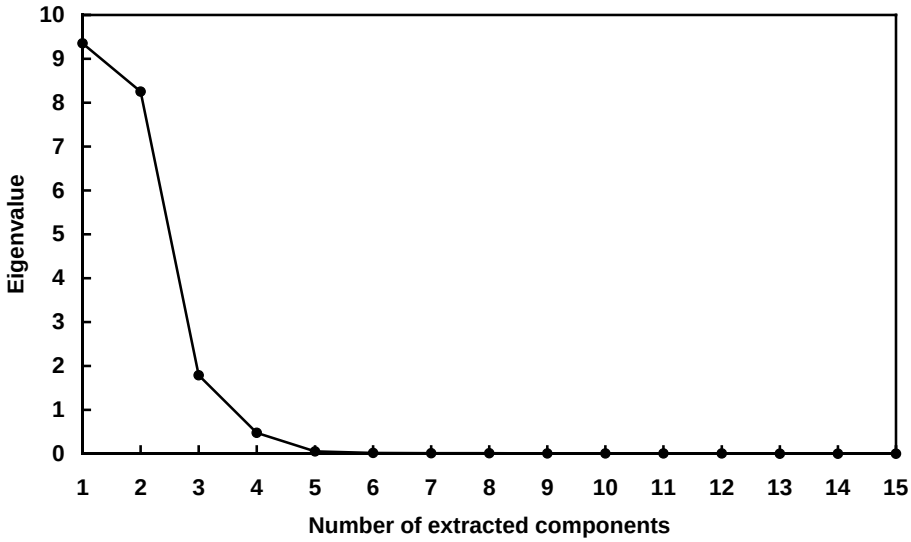


Fig. 3. Scree plot for deciding on the number of variables or components to be used in discriminant analysis ((●) eigenvalue of extracted component).

that the first three components had eigenvalues greater than unity (Kaiser criterion) while the first four or five components satisfied the scree test.

Changes in Wilks' λ values versus the number of original variables or components used are shown in Fig. 4. The use of a logarithmic scale for the Wilks' λ axis (showing values divided by 10^{-3}) was mandatory since a wide range of values had to be covered. As it was theoretically expected, the discriminatory power of the model increased as more and more variables (or components) were used. The change in the slope observed at the point that corresponded to three variables (or components) denoted that little improvement was achieved by further increasing the number of variables (or components). This point was in agreement with the Kaiser criterion result. Slightly better (smaller) Wilks' λ values were observed when components were used instead of original variables. At the same time, the percent average redundancy of the original variables in the discriminant

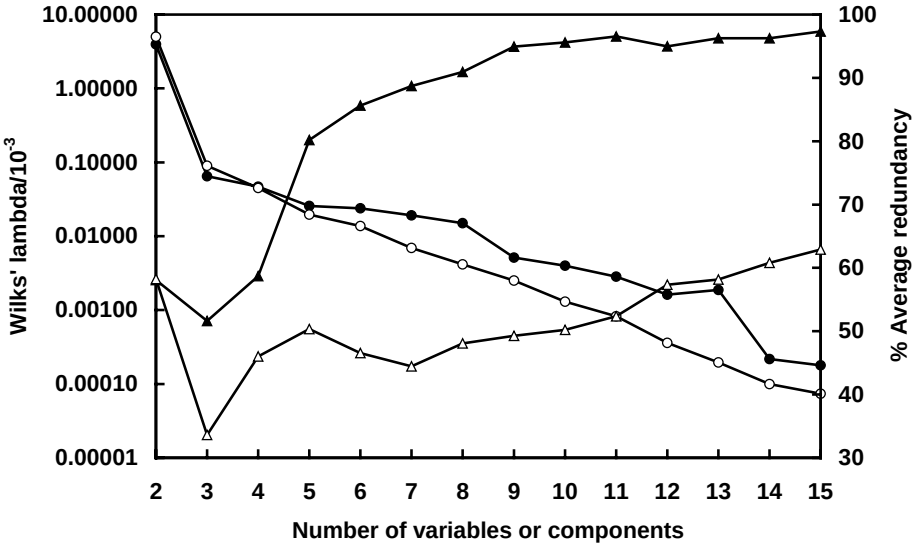


Fig. 4. Wilks' λ and percent average redundancy of variables or components in discriminant analysis ((●) Wilks' λ based on original variables, (○) Wilks' λ based on principal components, (▲) percent average redundancy of original variables, (△) percent average redundancy of principal components).

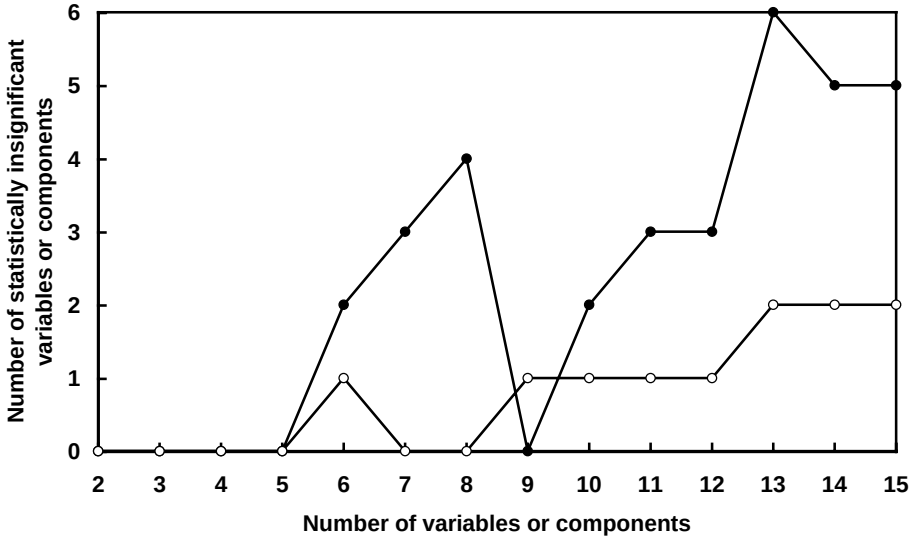


Fig. 5. Number of statistically insignificant variables or components in discriminant analysis ((●) number of statistically insignificant original variables, (○) number of statistically insignificant principal components).

model increased rapidly when more than four variables were used and eventually reached a plateau at ca. 96%. It is also worth mentioning that four variables is the predicted number of variables by the scree test. When principal components were used in the model, the same increase was observed at the point that corresponded to four components, but the final plateau did not exceed 63%. This was expected since principal components are orthogonal and uncorrelated by definition. The use of principal components as new latent variables in DA was therefore preferable.

Examination of the number of statistically insignificant (at $P = 0.05$) variables or components in the discriminant model (Fig. 5) showed that when five or less variables or components were entered into the DA module, they were all statistically significant at the $P = 0.05$ level. Again the use of principal components instead of original variables proved to be more adequate. In every case, the percent correct post hoc classification of the items in the training dataset was 100% except when only two variables or components were used, something that resulted in percent correct classification values of 91.3 and 84.8%, respectively.

Based on the findings mentioned above, we decided to use the scores of the first three principal components in the DA model. This allowed us to sacrifice as little discriminatory power as possible and also avoid entering too many new latent variables in the model. The three components used accounted for 96.97% of the total variance in the original data. The Wilks' λ of the calculated discriminant model was 8.98×10^{-5} and the model was found to be statistically significant at the $P = 0.05$ level with an average variable redundancy of 33.6%. The post hoc classification of the items in the training dataset was found to be 100% correct.

The three discriminant functions (canonical roots) calculated for the model were checked for their significance by means of a χ^2-test and successive removal of the roots until only one root remained. From the P-values in Table 1 it can be seen that all three roots were statistically significant at the $P = 0.05$ level.

The discriminant functions that were calculated were of the form:

$$L = b_0 + \sum_{i=1}^{3} b_i PC_i$$

The raw (b_i) and standardized (B_i) coefficients of the canonical roots are given in Table 2. Raw coefficients were

Table 1
Significance of discriminant functions

Removed roots	χ^2	d.f.	P-value
0	382.0371	12	<0.001
First	194.9673	6	<0.001
First, second	27.37222	2	<0.001

Table 2
Raw and standardized coefficients of discriminant functions

	Root 1	Root 2	Root 3
Raw coefficient			
b_0	7.06×10^{-15}	5.70×10^{-15}	-1.33×10^{-16}
b_1	−1.2365	0.2430	1.3203
b_2	3.8685	−6.6783	0.1139
b_3	−8.4158	−3.1056	−0.1416
Standardized coefficient			
B_1	−0.9193	0.1807	0.9816
B_2	0.5633	−0.9724	0.0166
B_3	−1.1575	−0.4271	−0.0195

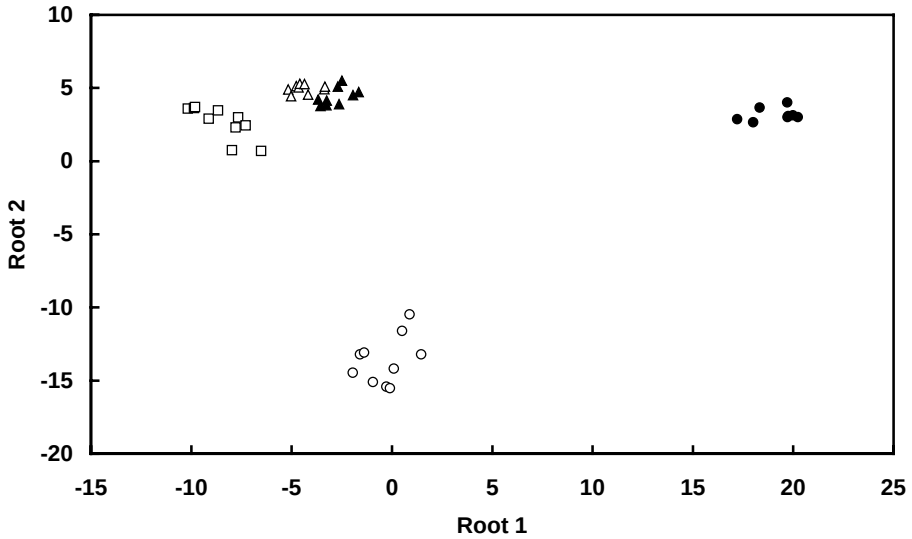


Fig. 6. Canonical score graph of the first two discriminant functions for the post hoc classification of the training dataset ((○) BI, (△) FC, (□) PE, (●) PI, (▲) ST).

Table 3
Statistically significant component loadings

PC_1			PC_2			PC_3		
Original variable	Loading	*P*-value	Original variable	Loading	*P*-value	Original variable	Loading	*P*-value
$\log_{10}(\%A_{418})$	0.9614	<0.001	$\log_{10}(\%A_{677})$	0.9980	<0.001	$\log_{10}(\%A_{495})$	0.7999	<0.001
$\log_{10}(\%A_{405})$	0.9551	<0.001	$\log_{10}(\%A_{660})$	0.9858	<0.001	$\log_{10}(\%A_{511})$	0.7374	<0.001
$\log_{10}(\%A_{443})$	0.9513	<0.001	$\log_{10}(\%A_{686})$	0.9748	<0.001	$\log_{10}(\%A_{482})$	0.7106	<0.001
$\log_{10}(\%A_{444})$	0.9483	<0.001	$\log_{10}(\%A_{696})$	0.7583	<0.001	$\log_{10}(\%A_{524})$	0.6330	<0.001
$\log_{10}(\%A_{707})$	0.9296	<0.001	$\log_{10}(\%A_{643})$	0.6645	<0.001	$\log_{10}(\%A_{537})$	0.4773	0.001
$\log_{10}(\%A_{728})$	0.9092	<0.001	$\log_{10}(\%A_{495})$	−0.3392	0.021	$\log_{10}(\%A_{464})$	0.4599	0.001
$\log_{10}(\%A_{464})$	0.8742	<0.001	$\log_{10}(\%A_{511})$	−0.5995	<0.001	$\log_{10}(\%A_{643})$	−0.7040	<0.001
$\log_{10}(\%A_{482})$	0.6910	<0.001	$\log_{10}(\%A_{524})$	−0.7661	<0.001	$\log_{10}(\%A_{622})$	−0.9469	<0.001
$\log_{10}(\%A_{696})$	0.5609	<0.001	$\log_{10}(\%A_{537})$	−0.8754	<0.001			
$\log_{10}(\%A_{495})$	0.4881	0.001	$\log_{10}(\%A_{570})$	−0.9760	<0.001			
$\log_{10}(\%A_{511})$	0.3059	0.039	$\log_{10}(\%A_{553})$	−0.9803	<0.001			

used for the calculation of the sample canonical scores and the drawing of the respective graph (Fig. 6). Only the first two roots are given since those were the ones that played the most important role in the discrimination. Root 1 was responsible for the separation ||PE||FC-ST-BI||PI||, whereas root 2 further aided the separation ||FC-ST||BI||. Root 3, which is not shown here, was responsible for the final separation ||FC||ST||.

Standardized coefficients were used for the estimation of the component contribution to the discrimination achieved by the respective discriminant function. From Table 2 it can be seen that PC_3 played the most important role in the discrimination achieved by root 1. From the statistically significant ($P < 0.05$) component loadings (Table 3) it can be seen that PC_3 mainly represented absorbance at 622 nm. At that wavelength, spectra belonging to the PI group demonstrated an absorbance shoulder (Fig. 1) that was absent in the other groups. The discriminatory power of root 1 is especially evident in Fig. 6 in which the PI group lies on the far right, completely separated from the other

Table 4
Contribution to discrimination and redundancy of components

Component	Partial Wilks' λ	*P*-value	Tolerance	Redundancy (%)
PC_1	0.2735	<0.001	0.5431	45.7
PC_2	0.0153	<0.001	0.7917	20.8
PC_3	0.0113	<0.001	0.6568	34.3

Table 5
Classification function coefficients

Coefficient	Ink group				
	BI	FC	PE	PI	ST
b_0'	−94.3730	−24.9465	−41.5852	−189.4998	−15.4220
b_1'	−3.1026	4.4046	12.4671	−22.5842	5.2652
b_2'	89.6619	−50.2556	−50.4302	52.7334	−40.2094
b_3'	45.1379	21.7563	63.0824	−170.7646	9.7896

groups. Table 2 also shows that root 2 was mainly affected by PC_2. From Table 3 it can be seen that this component represented absorbance in the range of 660–686 nm and at 553 and 570 nm. Fig. 1 shows that the first range corresponded to the second peak that characterized BI and PI inks with *%A* values being significantly higher in the case of BI inks. Absorption at 553 and 570 nm was also different for BI inks and corresponded to the ascending part of the band with a slope that was smaller than the slopes of the other curves at those wavelength points. These features were responsible for the separation of the BI group by means of root 2, something that is also evident in Fig. 6. Similarly, PC_1 scores affected the discrimination achieved by root 3 (Table 2). This component represented absorbance values in the range of 405–444 nm and at 707 and 728 nm (Table 3). All of these wavelengths corresponded to very low absorbance values at the spectra extremes (Fig. 1). Although unnoticeable to the naked eye, these subtle differences in absorption were the ones that achieved the separation of the somewhat overlapping FC and ST groups.

The overall contribution of the principal components to the discriminant model is shown in Table 4. The contribution followed the order $PC_3 > PC_2 > PC_1$ and all components were found to be statistically significant at the $P = 0.05$ level. This order is in perfect agreement with what was mentioned above about the role and nature of each discriminant function.

The DA was completed with the calculation of the so-called classification functions. These functions allow the post hoc classification of the items in the training dataset or the classification of new items. To classify new samples one would have to calculate the relevant component scores first and then enter the results into the classification functions. The function yielding the highest result would indicate the group to which the new sample belonged. Five classification functions were calculated (one for each group) and were of the form:

$$C = b_0' + \sum_{i=1}^{3} b_i' PC_i$$

Their coefficients are given in Table 5 and the post hoc classification of the training dataset by means of these functions was found to be 100% correct (no misclassifications occurred).

4. Conclusions

This study showed that excellent discrimination (100% correct classification of the training dataset) between inks of different brands could be achieved by examination of the Vis spectra of ethanolic ink solutions and application of a multivariate chemometrics protocol for the study of the analytical data. A discriminant model was calculated with a Wilks' λ of 8.98×10^{-5} and was found to be statistically significant at the $P = 0.05$ level. Problems with data multicollinearity were effectively dealt with by means of PCA, which was used for the calculation of three new latent variables (principal components) that were used in DA. The average redundancy of the latent variables in the discriminant model was 33.6%. The study showed that discrimination of the samples was based on differences that regarded (a) the shape of the first spectral band, which was present in all cases, (b) the intensity of the second band, which appeared only in the case of two inks and (c) the absorption of inks at the spectra extremes. The proposed method is similar to chromatography, in the sense that it detects the coloring materials in the sample, but avoids all the time consuming steps that characterize separation techniques. We believe that a more thorough study of the system could be based on the use of many more ink samples representative of the blue ballpoint pen ink population. Such an experiment should also regard the variability between inks of the same brand but of different batches and should involve either the extraction of inks from documents (destructive technique) or direct observation of the radiation reflected by the ink on the document (non-destructive technique). The calculated models should also be tested on new samples instead of items from the training dataset to ensure a better assessment of their usefulness. This would help establish an appropriate protocol that could allow forensic document examiners to draw objective conclusions regarding the similarity of inks used to write various sections of a document.

Appendix A. Statistical background

A.1. Cluster analysis

The term *cluster analysis* is used to describe a number of different classification algorithms. Generally, these

algorithms allow the organization of observed data into meaningful structures, thus promoting the development of taxonomies. In its simplest form, CA is based on a joining or tree clustering algorithm. Its purpose is to join objects into successively larger clusters (hierarchical tree) using some measure of similarity between the objects. The most commonly employed similarity measure in this technique is the Euclidian distance between the objects. If the objects have been measured on m variables, an m-dimensional space exists in which the Euclidian distance between objects k and l is given by the formula

$$D_{kl} = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2}$$

Another form of CA is the K-means algorithm that addresses a different problem, namely that of which objects belong to a certain predefined number of clusters. To answer this question one needs an iterative method according to which k random clusters are formed initially and then objects are moved between these clusters so that the within cluster variance is minimized, while the between cluster variance is maximized. Although it is more common to employ CA to classify objects on the basis of their variables, it is possible to run a K-means algorithm on the variables over the objects, thus achieving the grouping of variables that carry similar information about the objects.

A.2. Principal component analysis

Principal component analysis is a technique that belongs to the broader field of factor analysis. Generally, FA aims at (a) reducing the number of variables on which the objects of a dataset were measured and (b) detecting structure in the relationships between the variables. Principal component analysis achieves the aims mentioned above by using linear combinations of the original variables (manifest variables) to yield new variables called principal components or PCs (latent variables). The extraction of the PCs is successive with the first PC explaining most of the variance in the original data. The second PC can then be extracted to explain most of the remaining unexplained variance. This procedure can be repeated m times for m manifest variables until all the original variance has been explained. By definition, the extracted components are orthogonal and therefore uncorrelated. By extracting only the first two or three PCs one can project the objects of a dataset on a plane or in a three-dimensional space respectively and visualize an otherwise unperceivable m-dimensional space.

The correlations of the PC scores with the original scores on the m manifest variables are called component loadings and form the basis for the qualitative interpretation of the extracted components. A technique called Varimax rotation is used to ensure that the loading of a manifest variable is maximized on one component while it is minimized on all other components. This results in a much easier interpretation of the PCs. If some of the original variables are already correlated, they are expected to load highly on the same component.

Each extracted component is characterized by its eigenvalue which roughly corresponds to the number of manifest variables this component represents. For the decision concerning the number of PCs that should be extracted for a given dataset two criteria have been extensively used: the Kaiser criterion and the scree test. According to the Kaiser criterion, only components with eigenvalues greater than unity should be extracted, the rationale being that components representing less than one variable should not be taken into account. On the other hand, the scree test requires the plotting of eigenvalues against the number of extracted components and the determination of the point where the plot levels off. Beyond this point, no further improvement in variance explanation can be achieved and more components are not needed.

A.3. Discriminant analysis

Discriminant analysis resembles PCA in the sense that new latent variables are formed from the original manifest variables, but the requirement is that maximum separation of the objects is achieved. The new latent variables which are also called discriminant functions or canonical roots are a linear combination of the manifest variables and are orthogonal.

The discriminatory power of the model calculated this way is assessed by means of the Wilks' λ statistic, which is given by the formula

$$\lambda = \frac{\det(W)}{\det(T)}$$

where $\det(W)$ is the determinant of the within-groups variance–covariance matrix and $\det(T)$ is the determinant of the total variance–covariance matrix. The smaller the Wilks' λ value is, the more effective the model is. Another measure of the individual contribution of each variable in the discriminant model is the partial Wilks' λ statistic, which is the ratio of the Wilks' λ value after adding the respective variable over the Wilks' λ value before adding the variable. Smaller values of the partial Wilks' λ statistic denote a higher contribution of the respective variable.

The effectiveness of the discriminant model can also be checked by running a post hoc classification of the training dataset objects. According to this technique, the original objects are treated as new ones and are classified by means of the classification functions calculated for the respective model. A variant of this technique is based on the classification of entirely new objects and observation of any misclassifications.

References

[1] C. Vogt, J. Vogt, A. Becker, E. Rohde, Separation, comparison and identification of fountain pen inks by

capillary electrophoresis with UV-visible and fluorescence detection and by proton-induced X-ray emission, J. Chromatogr. A 781 (1997) 391–405.
[2] C. Roux, M. Novotny, I. Evans, C. Lennard, A study to investigate the evidential value of blue and black ball-point pen inks in Australia, Forensic Sci. Int. 101 (1999) 167–176.
[3] J.A. Zlotnick, F.P. Smith, Chromatographic and electrophoretic approaches in ink analysis, J. Chromatogr. B 733 (1999) 265–272.
[4] R. Saferstein, Criminalistics: An Introduction to Forensic Science, seventh ed., Prentice-Hall, Englewood Cliffs, NJ, 2001, p. 468.
[5] J.W. Brackett Jr., L.W. Bradford, Comparison of ink writing on documents by means of paper chromatography, J. Crim. Law Crim. Police Sci. 43 (1952) 530–539.
[6] D.A. Crown, J.V.P. Conway, P.L. Kirk, Differentiation of blue ball-point pen inks, J. Crim. Law Crim. Police Sci. 52 (1961) 338–343.
[7] M. Lederer, M. Schudel, Adsorption chromatography on cellulose V: a simple chromatographic system for the identification of inks, J. Chromatogr. 475 (1989) 451–456.
[8] G.R. Nakamura, S.C. Shimoda, Examination of microquantity of ball-point inks from documents by thin-layer chromatography, J. Crim. Law Crim. Police Sci. 56 (1965) 113–118.
[9] R.S. Verma, K.N. Prasad, G.J. Misra, Thin-layer chromatographic analysis of fiber-tip pen inks, Forensic Sci. Int. 13 (1979) 65–70.
[10] V.N. Aginsky, Forensic examination of 'slightly soluble' ink pigments using thin-layer chromatography, J. Forensic Sci. 38 (1993) 1131–1133.
[11] D. Doud, Chromatographic analysis of inks, J. Forensic Sci. 3 (1958) 486–492.
[12] J.A. Tappolet, High-performance thin-layer chromatography: its application to the examination of writing inks, Forensic Sci. Int. 22 (1983) 99–109.
[13] R.N. Totty, M.R. Ordidge, L.J. Onion, Comparison of the use of visible microspectrometry and high performance thin-layer chromatography for the discrimination of aqueous inks used in porous tip and roller ball pens, Forensic Sci. Int. 28 (1985) 137–144.
[14] K.M. Varshney, T. Jettappa, V.K. Mehrotra, T.R. Baggi, Ink analysis from typed script of electronic typewriters by high performance thin layer chromatography, Forensic Sci. Int. 72 (1995) 107–115.
[15] R.M. Kevern, Infrared luminescence from thin layer chromatograms of inks, J. Forensic Sci. Soc. 13 (1973) 25–28.
[16] R.D. Blackledge, M. Iwan, Differentiation between inks of the same brand by infrared luminescence photography of their thin-layer chromatography, Forensic Sci. Int. 21 (1983) 165–173.
[17] V.N. Aginsky, Comparative examination of inks by using instrumental thin-layer chromatography and microspectrophotometry, J. Forensic Sci. 38 (1993) 1111–1130.
[18] L.F. Colwell, B.L. Karger, Ball-point pen ink examination by high pressure liquid chromatography, J. Assoc. Off. Anal. Chem. 60 (1977) 613–618.
[19] P.C. White, B.B. Wheals, Use of a rotating disc multi-wavelength detector operating in the visible region of the spectrum for monitoring ball pen inks separated by high-performance liquid chromatography, J. Chromatogr. 303 (1984) 211–216.
[20] A. Löfgren, J. Andrasko, HPLC analysis of printing inks, J. Forensic Sci. 38 (1993) 1151–1160.
[21] J.W. Thompson, Identification of ink by electrophoresis, J. Forensic Sci. Soc. 7 (1967) 199–202.
[22] H.W. Moon, Electrophoretic identification of felt-tip pen inks, J. Forensic Sci. 25 (1980) 146–149.
[23] S. Fanali, M. Schudel, Some separations of black and red water-soluble fiber-tip pen inks by capillary zone electrophoresis and thin-layer chromatography, J. Forensic Sci. 36 (1991) 1192–1197.
[24] E. Rohde, A.C. McManus, C. Vogt, W.R. Heineman, Separation and comparison of fountain pen inks by capillary electrophoresis, J. Forensic Sci. 42 (1997) 1004–1011.
[25] J.A. Zlotnick, F.P. Smith, Separation of some black roller-ball pen inks by capillary electrophoresis: preliminary data, Forensic Sci. Int. 92 (1998) 269–280.
[26] A. Giles, The forensic examination of documents, in: P. White (Ed.), Crime Scene to Court: The Essentials of Forensic Science, The Royal Society of Chemistry, Cambridge, 1998, pp. 123–125.
[27] N.C. Thanasoulias, E.T. Piliouris, M.S.E. Kotti, N.P. Evmiridis, Application of multivariate chemometrics in forensic soil discrimination based on the UV-Vis spectrum of the acid fraction of humus, Forensic Sci. Int. 130 (2002) 73–82.
[28] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988, pp. 407–409.
[29] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1960) 141–151.
[30] R.B. Cattell, The scree test for the number of factors, Multivar. Behav. Res. 1 (1966) 245–276.