



Οι βασικές αρχές μεθόδων των Εφαρμοσμένων Μαθηματικών στην ερμηνεία των πρωτογενών δεδομένων στη Κυτταρομετρία Ροής

Δρ. Βασίλης Σπυρόπουλος, Καθηγητής

Διευθυντής του Εργαστηρίου Βιοϊατρικής Τεχνολογίας (EBIT) του
Τμήματος Τεχνολογίας Ιατρικών Οργάνων του ΤΕΙ Αθήνας

Email: basile@teiath.gr **URL:** www.bmtl.bme.teiath.gr

Στατιστικές μέθοδοι στη Κυτταρομετρία Ροής

- Στη **Κυτταρομετρία Ροής**, για πρώτη φορά, η περίπλοκη Μορφή της κατανομής των πολυπληθών αποτελεσμάτων των μετρήσεων, αποτελεί **τον πυρήνα της Διαγνωστικής ερμηνείας των αποτελεσμάτων** για την λήψη Ιατρικής Απόφασης.
- Με βάση το **Syllabus** θα ασχοληθούμε **επιγραμματικά** σήμερα με τα ακόλουθα θέματα:
 - ◆ *Περιγραφική Στατιστική*
 - ◆ *Ανάλυση Ομάδων (Cluster analysis).*
 - ◆ *Γραμμική Διακριτή Ανάλυση (Linear Discriminant Analysis)*
 - ◆ *Ανάλυση κύριας συνιστώσας (Principal Component Analysis)*
 - ◆ *Στατιστική K/S (Kolmogorov–Smirnov test)*
- Επίσης θα αναφερθούν **ενδεικτικά** εμπορικά ή δωρεάν διαθέσιμα **λογισμικά** που αξιοποιούν τις μεθόδους αυτές.

Περιγραφική Στατιστική

- Η **Περιγραφική Στατιστική** (ΠΣ) περιγράφει ποσοτικά τα κύρια χαρακτηριστικά μιας συλλογής δεδομένων.
- Η ΠΣ διακρίνεται από τη **Συμπερασματική Στατιστική** (ΣΣ) ή **Επαγωγική Στατιστική** (ΕΣ), από το ότι η ΠΣ **στοχεύει να συνοψίσει ένα δείγμα**, αντί να χρησιμοποιήσει τα δεδομένα του δείγματος, για να μελετήσει **το τι πιστεύεται ότι αντιπροσωπεύουν** αυτά, για τον συνολικό πληθυσμό των δεδομένων.
- Αυτό συνήθως σημαίνει ότι η ΠΣ, σε αντίθεση με την ΣΣ, **δεν έχει αναπτυχθεί βάσει της θεωρίας πιθανοτήτων**.
- Ακόμα και όταν μια ανάλυση δεδομένων εξάγει τα κύρια συμπεράσματά **χρησιμοποιώντας ΣΣ στοιχεία**, γενικά **παρουσιάζεται επίσης μια ΠΣ**.

Μέτρα κεντρικής τάσης και μέτρα μεταβλητότητας ή διασποράς

- Για παράδειγμα, σε μια εργασία που παρουσιάζει μια μελέτη που αφορά **πληθυσμούς ανθρώπων**, εμφανίζεται συνήθως ένας πίνακας που δίνει το **συνολικό μέγεθος του δείγματος**, τα μεγέθη δείγματος **σημαντικών υποομάδων** (π.χ. για κάθε ομάδα θεραπείας ή έκθεσης) και **δημογραφικά ή κλινικά χαρακτηριστικά**, όπως ο μέσος όρος ηλικίας, το ποσοστό των ατόμων κάθε φύλου, το ποσοστό των ατόμων με συναφείς συνοσηρότητες κλπ.
- Περιγραφική Στατιστική είναι επίσης ένα σύνολο **σύντομων περιγραφικών συντελεστών**, που **συνοψίζουν ένα σύνολο δεδομένων** και αντιπροσωπεύουν, είτε ολόκληρο τον πληθυσμό ή ένα δείγμα του.
- Οι **Μετρικές** (τα μέτρα) που περιγράφουν το σύνολο των δεδομένων, είναι **μέτρα κεντρικής τάσης και μέτρα μεταβλητότητας ή διασποράς**.

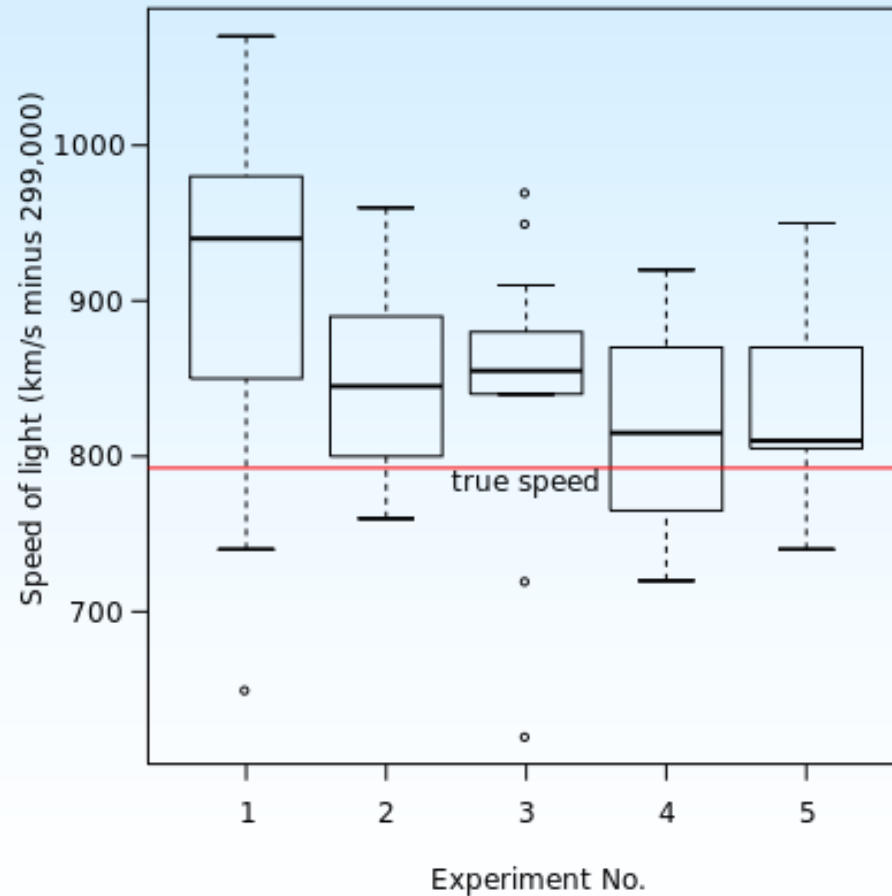
Μια απλή σύνοψη σχετικά με το δείγμα και τις παρατηρήσεις που έγιναν

- Τα μέτρα κεντρικής τάσης περιλαμβάνουν τη μέση τιμή, το διάμεσο και τη λειτουργία (mode), ενώ τα μέτρα διασποράς περιλαμβάνουν την τυπική απόκλιση (ή διακύμανση), την ελάχιστη και τη μέγιστη τιμή μιας μεταβλητής, τη κύρτωση και την ασυμμετρία (η παραμόρφωση).
- Η ΠΣ παρέχει μια απλή σύνοψη σχετικά με το δείγμα και τις παρατηρήσεις που έγιναν.
- Η σύνοψη μπορεί να είναι ποσοτική, δηλαδή σύνοψη στατιστικών, είτε οπτική, δηλαδή απλά στην κατανόηση γραφήματα.
- Οι περιλήψεις αυτές μπορούν, είτε να αποτελούν τη βάση της αρχικής περιγραφής των δεδομένων, ως μέρος μιας πιο εκτεταμένης στατιστικής ανάλυσης, ή μπορεί να είναι αυτές καθ' εαυτές επαρκείς, για μια συγκεκριμένη έρευνα.

Παραδείγματα

- Το **ποσοστό ευστοχίας βολών** ενός παίκτη σε μια ομάδα καλαθοσφαίρισης είναι ένας προσεγγιστικός δείκτης ΠΣ που συνοψίζει την απόδοση ενός παίκτη ή μιας ομάδας.
- Παρόμοια λειτουργεί **ο μέσος όρος βαθμολογίας του πτυχίου** ενός αποφοίτου μιας σχολής ενός ΑΕΙ, που περιγράφει τη γενική απόδοση του φοιτητή στη διάρκεια των σπουδών του.
- Η χρήση των **περιγραφικών στατιστικών** και οι **συνόψεις**, έχει μια μακρά ιστορία και, πράγματι, η **απλή στηλοθέτηση** των πληθυσμών και άλλων δεδομένων, ήταν ο πρώτος τρόπος που εμφανίστηκαν αυτά στις στατιστικές.
- Πιο πρόσφατα, έχει διατυπωθεί μια **συλλογή τεχνικών σύνοψης**. κάτω από τον τίτλο **exploratory data analysis** (διερευνητική ανάλυση δεδομένων).
- Ένα παράδειγμα είναι η τεχνική **box plot**.

Box plot δεδομένων από το περίφημο πείραμα των Michelson–Morley για την μέτρηση του c



Μονοπαραμετρική Ανάλυση

- Η Μονοπαραμετρική Ανάλυση περιλαμβάνει την περιγραφή της κατανομής μιας μόνο μεταβλητής, και συμπεριλαμβάνει την κεντρική τάση και τη διασπορά.
- Το σχήμα της κατανομής μπορεί επίσης να περιγραφεί μέσω δεικτών όπως η ασυμμετρία και η κύρτωση.
- Χαρακτηριστικά της κατανομής μιας μεταβλητής μπορεί επίσης να απεικονίζονται σε μορφή γραφικών ή σε μορφή πίνακα, συμπεριλαμβανομένων ιστογραμμάτων και stem and leaf χρονοπινάκων.

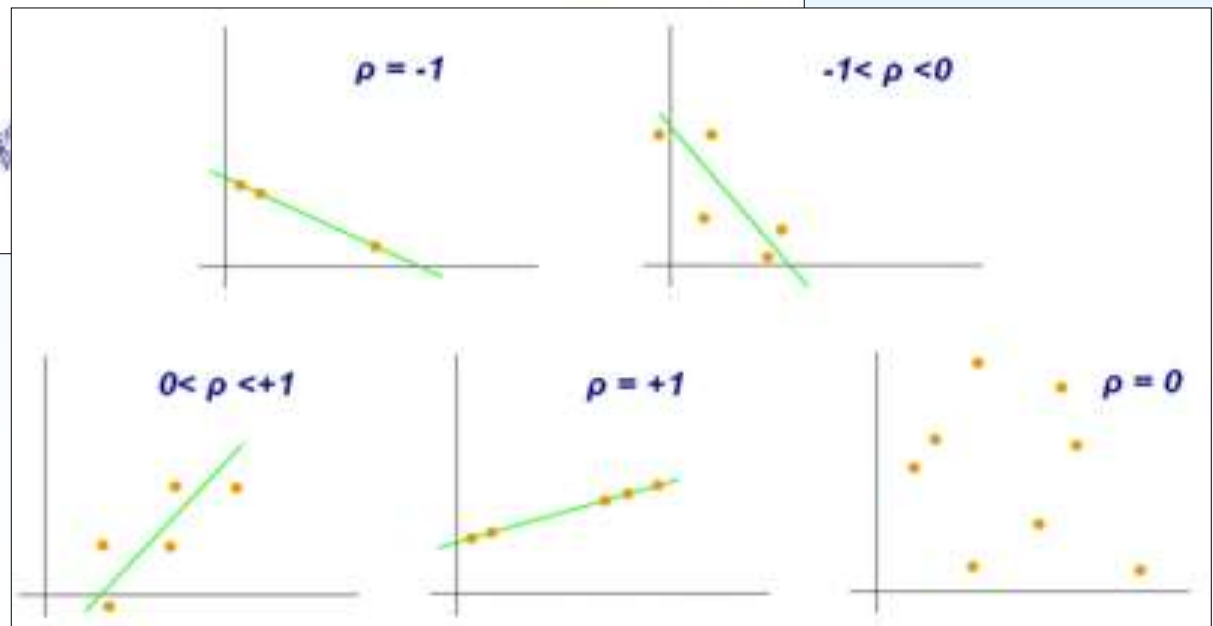
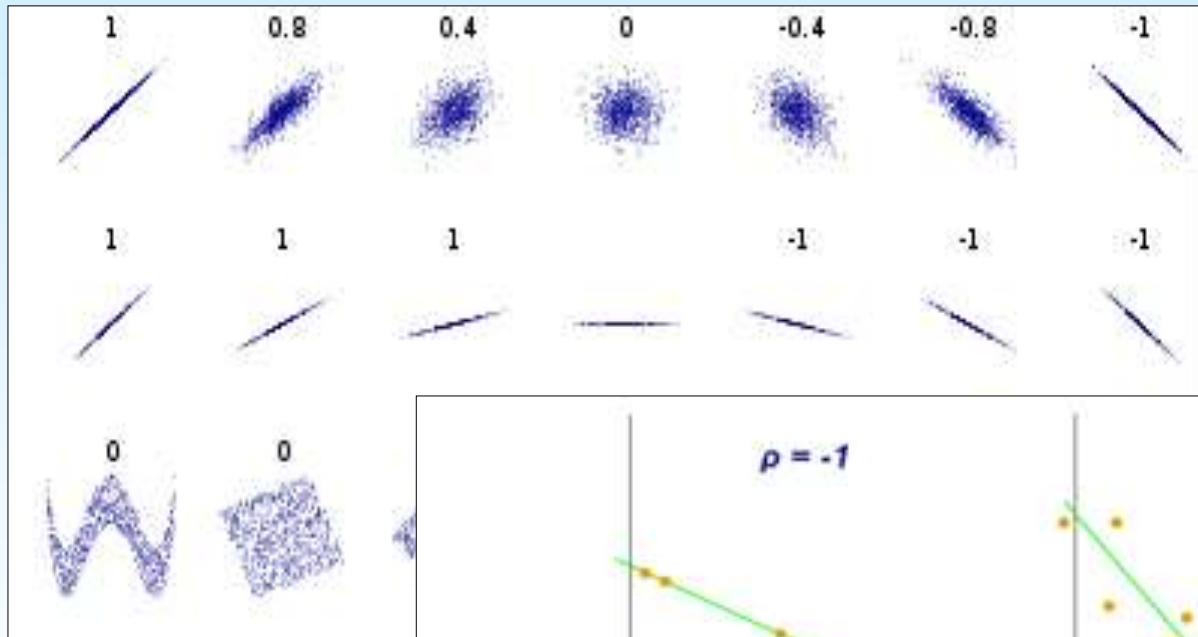
Χρονοπίνακες (stem and leaf)

平日 Weekdays												土曜日・休日 Saturdays, Sundays and Holidays											
5	10	17	29	39	47	54						5	10	17	29	39	47	54					
6	1	7	13	22	26	30	32	36	41	48	48	6	1	7	14	22	30	32	36	41	48	48	58
7	1	4	10	13	17	20	22	25	28	31	34	7	1	6	10	14	22	30	35	40	45	50	54
8	2	5	11	14	17	21	24	27	31	35	38	8	0	9	12	15	20	24	27	32	36	39	42
9	2	6	10	14	17	24	28	31	35	38	41	9	0	9	12	15	20	24	26	30	34	39	41
10	3	11	16	20	27	30	33	39	41	45	48	10	0	9	11	15	20	24	26	30	34	39	41
11	0	9	11	15	19	24	26	30	34	39	41	11	0	9	11	15	19	24	26	30	34	39	41
12	0	9	11	15	19	24	26	30	34	39	41	12	0	9	11	15	19	24	26	30	34	39	41
13	0	9	11	15	19	24	26	30	34	39	41	13	0	9	11	15	19	24	26	30	34	39	41
14	0	9	11	15	19	24	26	30	34	39	41	14	0	9	11	15	19	24	26	30	34	39	41
15	0	9	11	15	19	24	26	30	34	39	41	15	0	9	11	15	19	24	26	30	34	39	41
16	0	9	11	15	19	24	26	30	32	35	39	16	0	9	11	15	19	24	26	30	34	39	41
17	1	9	13	16	20	24	28	32	36	40	44	17	0	9	11	15	18	24	27	30	34	39	43
18	1	4	12	16	19	20	27	33	34	38	42	18	0	9	11	16	20	23	27	30	35	39	43
19	1	5	13	16	20	20	28	31	35	36	43	19	2	10	13	18	24	27	31	36	39	44	50
20	0	5	12	16	20	20	27	31	34	38	42	20	1	6	10	15	20	24	28	31	36	39	44
21	1	5	8	17	24	25	29	32	38	42	46	21	1	6	10	15	20	24	28	32	36	40	45
22	2	7	15	19	20	27	31	37	42	47	54	22	2	6	16	23	26	30	37	44	50	56	
23	3	9	14	19	22	26	30	34	40	45	48	23	0	6	14	22	28	32	40	49	55	59	
0	4	5	12	16	21	30	34	48				0	6	16	28	38							

Διπαραμετρική ή πολυπαραμετρική ανάλυση

- Όταν ένα δείγμα αποτελείται από **περισσότερες από μία μεταβλητές**, η ΠΣ πρέπει να χρησιμοποιείται για να περιγράψει τη σχέση μεταξύ **ζευγών μεταβλητών**.
- Στην περίπτωση αυτή, τα **περιγραφικά στατιστικά στοιχεία** περιλαμβάνουν:
 - ◆ *Διασταυρούμενους πίνακες (cross-tabulations).*
 - ◆ *Γραφήματα διασποράς (scatter plots).*
 - ◆ *Ποσοτικές μετρικές της εξάρτησης.*
 - ◆ *Περιγραφές των υπό όρους κατανομών.*
- Οι ποσοτικές μετρικές της εξάρτησης περιλαμβάνουν συσχετίσεις, όπως του **Pearson r** όταν δύο μεταβλητές είναι **συνεχείς**, ή του **Spearman ρ** εάν η μία ή και οι δύο δεν είναι **συνεχείς** και της **συνδιακύμανσης**, η οποία αντικατοπτρίζει **τις μεταβλητές της κλίμακας στην οποία μετράται**.

Παραδείγματα διαγραμμάτων διασποράς με διαφορετικούς συντελεστές συσχετισμού ρ



Το z-score

- Η **κλίση** στην ανάλυση παλινδρόμησης, αντανακλά επίσης **τη σχέση μεταξύ των μεταβλητών**.
- Η **μη τυποποιημένη κλίση** υποδεικνύει τη **μονάδα αλλαγής στη μεταβλητή-κριτήριο**, για μία μονάδα μεταβολής της πρόβλεψης.
- Η **τυποποιημένη κλίση** δηλώνει αυτήν την αλλαγή **στο τυποποιημένο z - score**.

ATM – INVESTORS ASSOCIATION

ALTMAN Z-SCORE

TYPE OF COMPANY: CYCLICAL COMPANY

PUBLICLY LISTED COMPANY
PRIVATE FIRMS
CYCLICAL COMPANY

1	Total Assets	
2	Total Liabilities	€ 80,944.00
3	Current Assets	€ 28,291.00
4	Current Liabilities	€ 50,255.00
5	EBIT	(€ 6,027.00)
6	Retained Earnings	(€ 40,419.00)
7	Net Sales ⁽¹⁾	€ 25,201.00
8	Market Capitalization ⁽²⁾	€ 7,425.00

-3.67

BANKRUPT ZONE

Z > 2.60	SAFETY ZONE
1.1 < Z < 2.60	GREY ZONE
Z < 1.1	BANKRUPT ZONE

(1) In Cyclical Companies is not considered the value of Net Sales.
(2) In Private Firms use the Shareholders' Equity.

Optimized for 1280x1024 resolution and works best with Microsoft Excel 2008 and up.
Designed and developed by Octávio Viana © 2007-2008 ATM | associação de investidores All Rights Reserved

Περιγραφική στατιστική κυτταρομετρικών μεταβλητών συλλεχθεισών σε διαφορετικές ομάδες μελετών

Table 1. Descriptive statistics of the cytometric variables collected in different study groups,^a and results of the reproducibility between-investigator variance, s^2_{bi} and repeatability (within-investigator variance, s^2_{wi}) analysis^b

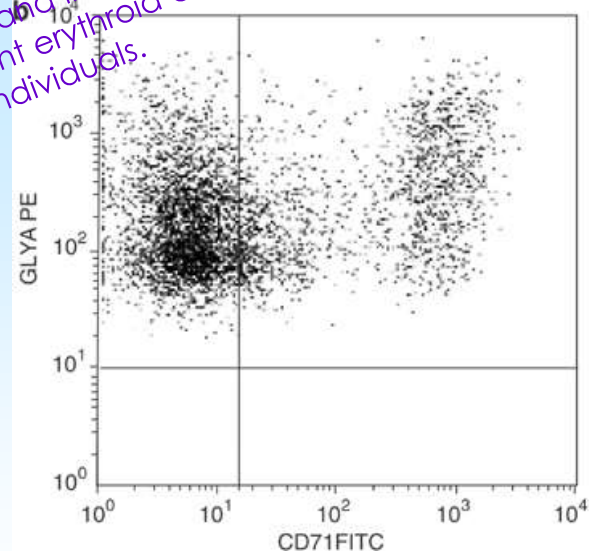
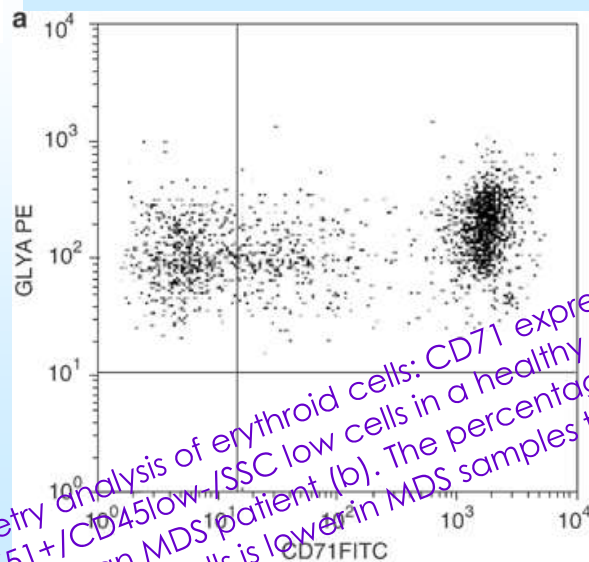
▲ Figure and tables index						Next
<i>Cytometric variables</i>	<i>Healthy donors</i>	<i>Pathologic controls</i>	<i>MDS</i>	<i>p^c</i>	<i>s²_{bi}</i>	<i>s²_{wi}</i>
Nucleated red cells	11.5 (7.4–14.3)	11 (7.5–20.0)	26.5 (13–37)	<0.001	0.62	2.66
GlyA+CD71+ nucleated red cells	78.6 (63–87.9)	66 (31–78)	53.0 (38.7–66)	0.02	15.94	0.52
CD10+ granulocytes	36 (25–38)	36 (28.9–43.6)	12 (7–34)	<0.001	20.40	17.87
CD56+ granulocytes	0.77 (0.6–1)	1 (0.7–1.2)	3 (1–5)	<0.001	0.21	0.47
CD33/CD16 immature/mature cell ratio	0.9 (0.7–1.2)	1.1 (0.8–1.4)	1.9 (1.1–3.2)	0.007	0.018	0.018
CD34+ cells	1.5 (1–2)	1.5 (1–2.2)	3.0 (1.4–6.4)	0.003	0.61	0.08
CD34+CD33+/CD34+CD33- ratio	1.0 (0.6–1)	1.0 (0.7–1.3)	2.3 (1.5–3.8)	<0.001	0.006	0.004

^a Immunophenotypic data are expressed as percentage of cells positive for antigens, and summarized by their median and quartiles.

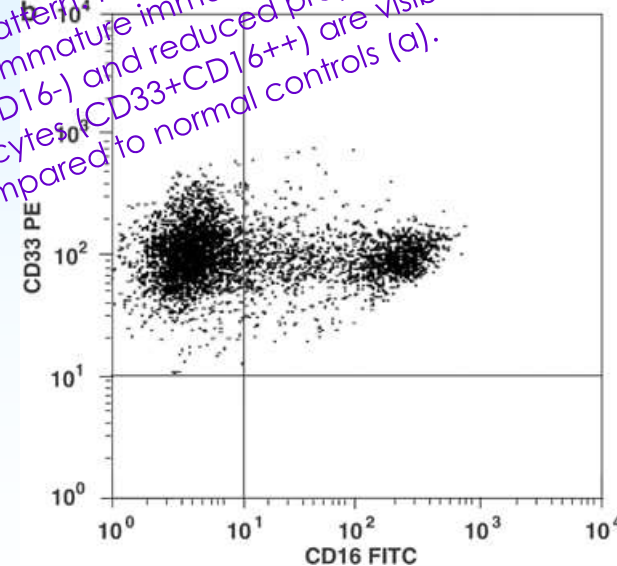
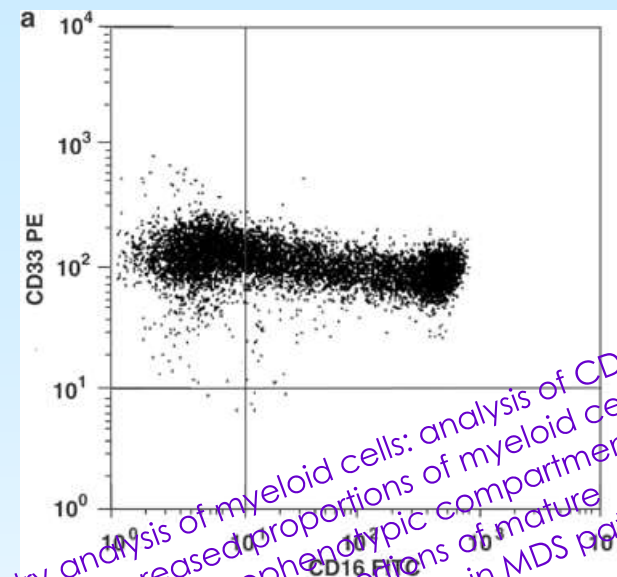
^b These results were obtained via a repeated-measure analysis of variance on data from three MDS samples evaluated by four blinded independent investigators, who analyzed each sample three times.

^c All the variables were log-transformed before the analysis of variance to normalize the distribution.

L Malcovati et al. “Flow cytometry evaluation of erythroid and myeloid dysplasia in patients with myelodysplastic syndrome Leukemia”, (2005), 19, 776–783



Flow Cytometry analysis of erythroid cells: CD71 expression on GYA+/LDS751+/CD45low/SSC low cells in a healthy stem cell donor (a) and in an MDS patient (b). The percentage of GYA+ CD71bright erythroid cells is lower in MDS samples than in normal individuals.



Flow Cytometry analysis of myeloid cells: analysis of CD33 vs CD16 pattern. Increased proportions of myeloid cells in the more immature immunophenotypic compartments (CD33+CD16-) and reduced proportions of mature granulocytes (CD33+CD16++) are visible in MDS patients (b) compared to normal controls (a).

Η Ανάλυση Ομάδων (Cluster Analysis)

- Η **Ανάλυση Ομάδων (Clusters) ή Ομαδοποίηση** είναι η κατάταξη ενός συνόλου αντικειμένων, με τέτοιο τρόπο ώστε **αντικείμενα της ίδιας ομάδας (συμπλέγματος) ομοιάζουν περισσότερο μεταξύ τους**, παρά με εκείνα **άλλων ομάδων (clusters)**, τα συνδέει δηλαδή **μια τουλάχιστον κοινή ιδιότητα**.
- Η μέθοδος αποτελεί **κύριο εργαλείο της διερευνητικής εξόρυξης δεδομένων**, αλλά και **κοινή τεχνική ανάλυσης στατιστικών δεδομένων** και χρησιμοποιείται σε πολλούς τομείς, συμπεριλαμβανομένων των **Μηχανικής Μάθησης, Αναγνώρισης Μορφών (patterns, motives), Ανάλυσης Εικόνας, Ανάκτησης Πληροφοριών, Βιοπληροφορικής**.
- Η Ανάλυση Ομάδων είναι **δεν ένας συγκεκριμένος Αλγόριθμος**, αλλά αποτελεί **το γενικά προς επίλυση πρόβλημα**.
- Αυτό **μπορεί να επιτευχθεί με διάφορους αλγόριθμους**, που διαφέρουν σημαντικά μεταξύ τους, τόσο όσον αφορά **στην αντίληψη του τι συνιστά ένα σύμπλεγμα**, αλλά και **στη διαδικασία εντοπισμού και αποτελεσματικού διαχωρισμού τους**.

Ιδιότητες και αλγόριθμοι ομαδοποίησης

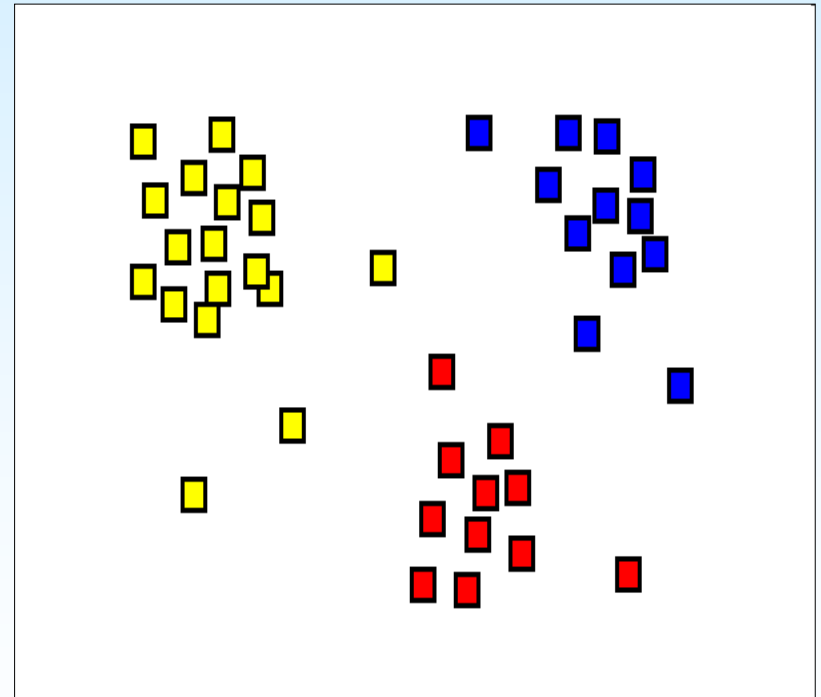
- «Δημοφιλείς» ομάδες είναι π.χ. αυτές που χαρακτηρίζονται από μικρές «αποστάσεις» μεταξύ των μελών τους, αυτές με περιοχές υψηλής πυκνότητας στο χώρο των δεδομένων, διαστήματα κλπ.
- Η «Ομαδοποίηση», επομένως, αποτελεί συχνά πρόβλημα πολυπαραμετρικής (multi-objective) βελτιστοποίησης.
- Οι κατάλληλοι αλγόριθμοι ομαδοποίησης και οι παράμετροι ρυθμίσεων (συμπεριλαμβανομένων των τιμών απόστασης, των κατωφλίων πυκνότητας ή αριθμού των αναμενόμενων συμπλεγμάτων), εξαρτώνται από το κάθε σύνολο δεδομένων και της σκοπούμενης χρήσης των αποτελεσμάτων.
- Η Ανάλυση Clusters συνεπώς, δεν είναι μίας «αυτόματη εργασία», αλλά μια επαναληπτική διαδικασία «εξόρυξης γνώσης» ή διαδραστικής πολυπαραμετρικής βελτιστοποίησης, που βασίζεται στη γενική προσέγγιση «δοκιμής και λάθους» (trial and error).
- Συχνά είναι αναγκαίο να τροποποιηθούν η προκαταρκτική επεξεργασία ή/και κάποιες ή όλες οι παράμετροι ή οι ιδιότητες, μέχρι να επιτευχθεί αναμενόμενο (ή το επιθυμητό) αποτέλεσμα.
- Ομαδοποιήσεις (συμπλέγματα) που βρέθηκαν με διαφορετικούς αλγόριθμους, μπορεί να διαφέρουν σημαντικά στις ιδιότητές τους, και τα μοντέλα αυτά ομαδοποίησης είναι το κλειδί για την κατανόηση των διαφορών μεταξύ των διαφόρων αλγορίθμων.

Τυπικά μοντέλα Ομαδοποιήσεων

- **Τα Κεντροειδή Μοντέλα:** Οι αλγόριθμοι αυτοί (k-means algorithms) αναπαριστούν κάθε ομάδα (σύμπλεγμα) με ένα Διάνυσμα.
- **Τα Μοντέλα Συνδεσιμότητας:** Είναι παραδείγματα Ιεραρχικής Ομαδοποίησης και δημιουργούν μοντέλα, που βασίζονται σε απόσταση συνδεσιμότητας.
- **Μοντέλα Πυκνότητας:** Για παράδειγμα τα μοντέλα DBSCAN και OPTICS ορίζουν συμπλέγματα ως συνδεδεμένες πυκνές περιοχές στο χώρο των δεδομένων.
- **Μοντέλα Κατανομής:** Είναι δυνατόν να μοντελοποιηθούν με τη χρήση στατιστικών κατανομών (όπως π.χ. πολύ-μεταβλητές κανονικές κατανομές, που χρησιμοποιούνται από τον αλγόριθμο μεγιστοποίησης προσδοκίας).
- **Μοντέλα Υποχώρων (Subspace):** Στο Biclustering ή Coclustering μοντελοποιούνται και με τα δύο μέλη του συμπλέγματος και πρόσθετες ενδεχόμενα σχετικές ιδιότητες.

Επιλογή αλγορίθμου ομαδοποίησης

- Υπάρχουν **αρκετές δεκάδες** σχετικών Αλγορίθμων (λιγότεροι από 100).
- Δεν υπάρχει **κανένας αντικειμενικά «σωστός» αλγόριθμος ομαδοποίησης**, αλλά συχνά **χρειάζεται να επιλέγεται πειραματικά**, εκτός και αν υπάρχει ένας **μαθηματικός λόγος προτίμησης ενός μοντέλου**.
- Πρέπει να σημειωθεί ότι ένας αλγόριθμος που έχει σχεδιαστεί **για ένα είδος μοντέλου**, δεν έχει **πιθανότητες επιτυχίας** σε ένα σύνολο δεδομένων που στηρίζεται σε **ριζικά διαφορετικού είδους μοντέλα**.



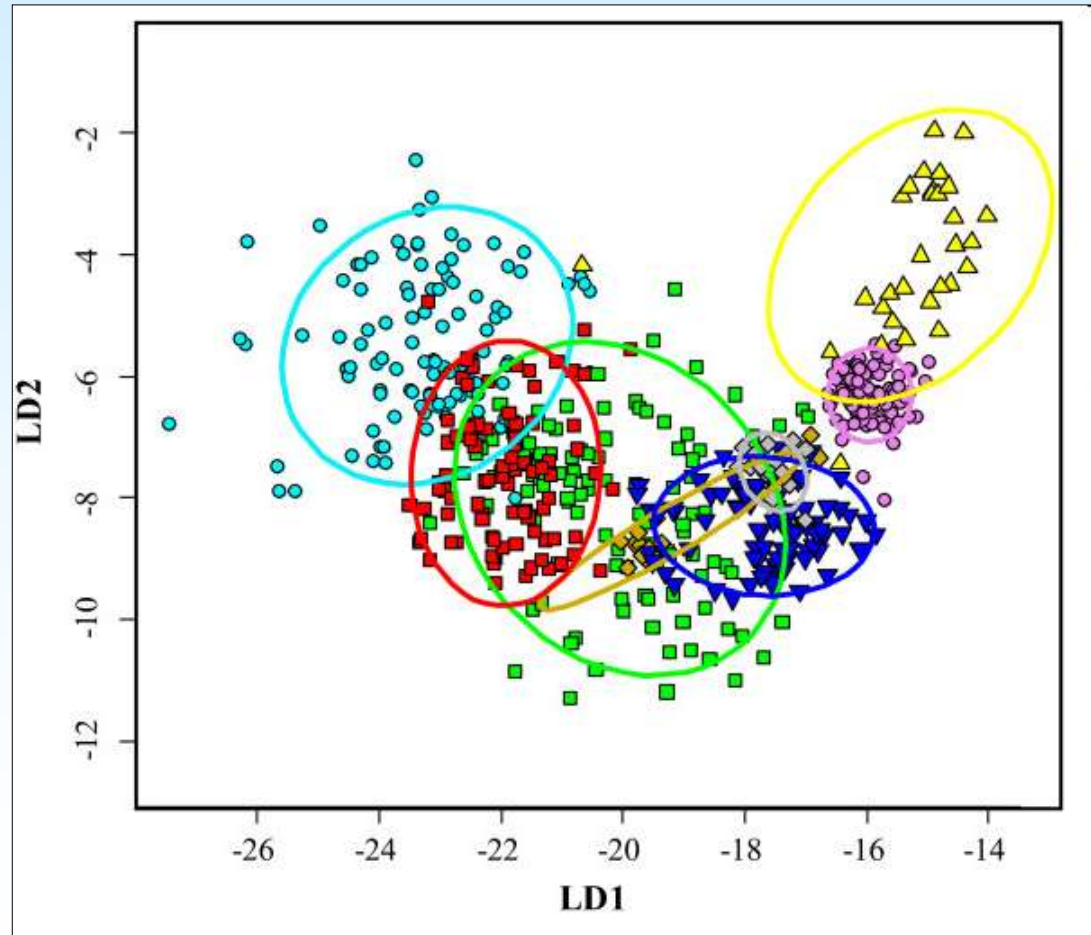
Γραμμική Διακριτή Ανάλυση (Linear Discriminant Analysis, LDA)

- Η Γραμμική Διακριτή Ανάλυση (Linear Discriminant Analysis, LDA) και η συναφής μέθοδος Fisher, είναι μέθοδοι που χρησιμοποιούνται για την επεξεργασία στατιστικών στοιχείων, αναγνώριση μορφών, διαδικασίες εκμάθησης μηχανών κλπ. με την εξεύρεση ενός Γραμμικού συνδυασμού χαρακτηριστικών, που χαρακτηρίζει ή διαχωρίζει δύο ή περισσότερες κατηγορίες, αντικειμένων ή γεγονότων.
- Ο συνδυασμός που προκύπτει μπορεί να χρησιμοποιηθεί ως ένας γραμμικός ταξινομητής ή πιο συχνά, για τη μείωση των διαστάσεων, πριν από την ταξινόμηση.
- Η LDA σχετίζεται στενά με την ανάλυση της διακύμανσης και με την ανάλυση της παλινδρόμησης, μεθόδους δηλαδή, που επίσης προσπαθούν να εκφράσουν μία εξαρτημένη μεταβλητή, ως ένα γραμμικό συνδυασμό άλλων χαρακτηριστικών ή μετρήσεων.

Ένα παράδειγμα εφαρμογής της Γραμμικής Διακριτής Ανάλυσης

Results of the linear discriminant analysis along the first and second discriminant axes (LD1 and LD2, respectively). LD1 and LD2 account for 80% and 18% discriminant power, respectively. 95% probability ellipses are indicated for each group.

Chiari et al. *Frontiers in Zoology* 2009 6:32



Η ειδοποιός διαφορά της LDA

- Στις άλλες δύο μεθόδους ωστόσο, η εξαρτημένη μεταβλητή είναι μια αριθμητική ποσότητα, ενώ για την LDA είναι μια κατηγορική μεταβλητής (δηλαδή η «ετικέτα» μιας κατηγορίας).
- Η λογιστική παλινδρόμηση και η probit παλινδρόμηση μοιάζουν περισσότερο με την LDA, καθώς εξηγούν επίσης μια κατηγορική μεταβλητή.
- Αυτές οι άλλες μέθοδοι είναι προτιμότερες σε εφαρμογές, όπου δεν είναι εύλογο να υποτεθεί ότι οι ανεξάρτητες μεταβλητές κατανέμονται κανονικά, πράγμα που αποτελεί μια θεμελιώδη παραδοχή της μεθόδου LDA.
- Η LDA έχει συνεχείς ανεξάρτητες μεταβλητές και μια κατηγορική εξαρτημένη μεταβλητή.
- Η LDA επιχειρεί ρητά να μοντελοποιήσει τη διαφορά μεταξύ των κατηγοριών των δεδομένων.

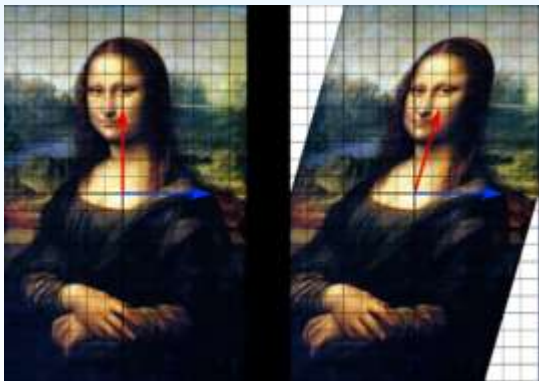
Ανάλυση κύριας συνιστώσας (Principal Component Analysis, PCA)

- Η Ανάλυση Κύριας Συνιστώσας (Principal Component Analysis, PCA) είναι μια μαθηματική διαδικασία η οποία χρησιμοποιεί έναν ορθογώνιο μετασχηματισμό, για να μετατρέψει ένα σύνολο παρατηρήσεων, πιθανώς συσχετιζόμενων μεταξύ τους μεταβλητών, σε ένα σύνολο τιμών, γραμμικά μη συσχετισμένων μεταβλητών.
- Ο αριθμός των κύριων συνιστωσών είναι μικρότερος ή ίσος προς τον αριθμό των αρχικών μεταβλητών.
- Η LDA είναι συγγενής με την PCA, διότι και οι δύο αναζητούν γραμμικούς συνδυασμούς μεταβλητών, οι οποίες εξηγούν καλύτερα τα δεδομένα
- Ο μετασχηματισμός αυτός ορίζεται κατά τέτοιο τρόπο ώστε, η πρώτη κύρια συνιστώσα να έχει τη μεγαλύτερη δυνατή διακύμανση.

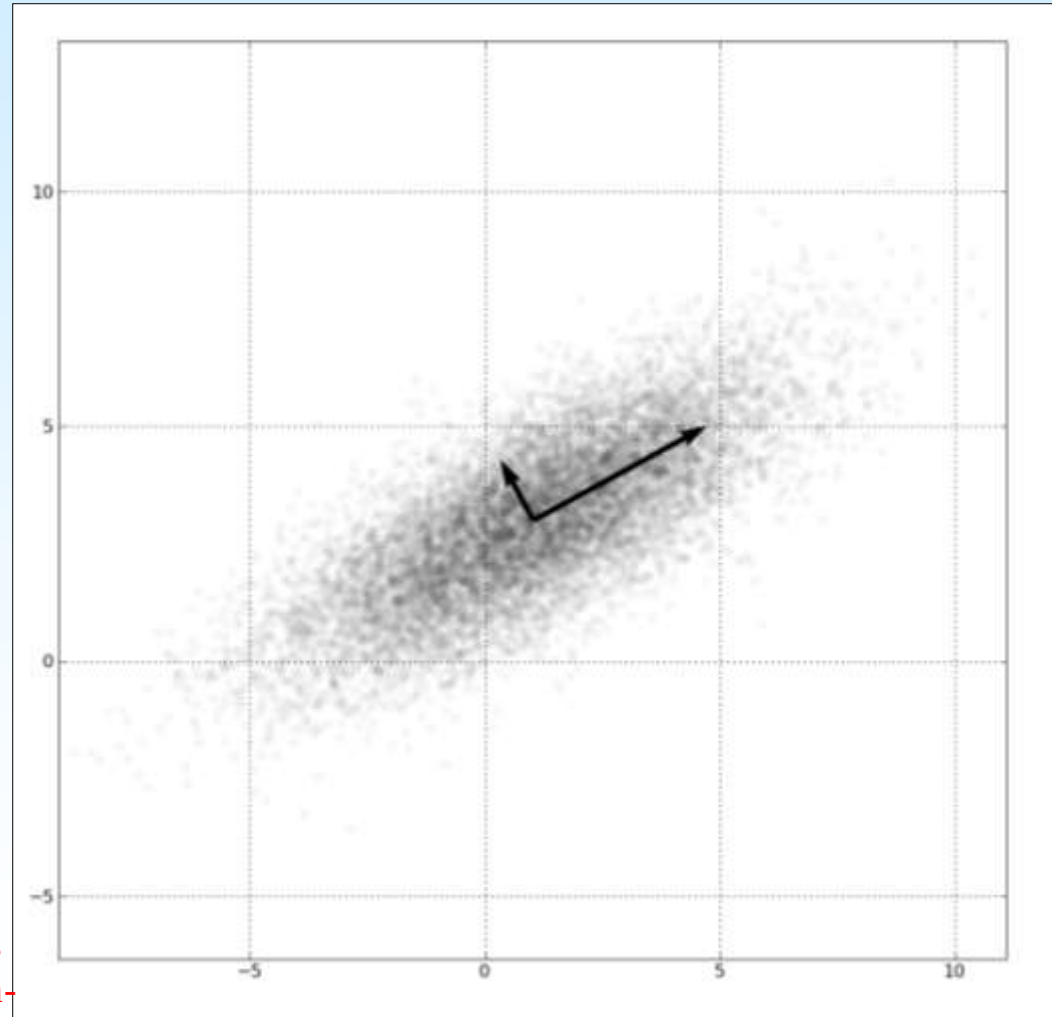
Παράδειγμα της PCA μιας κατανομής Gauss

PCA of a multivariate Gaussian distribution centered at $(1,3)$ with a standard deviation of 3 in roughly the $(0.878, 0.478)$ direction and of 1 in the orthogonal direction.

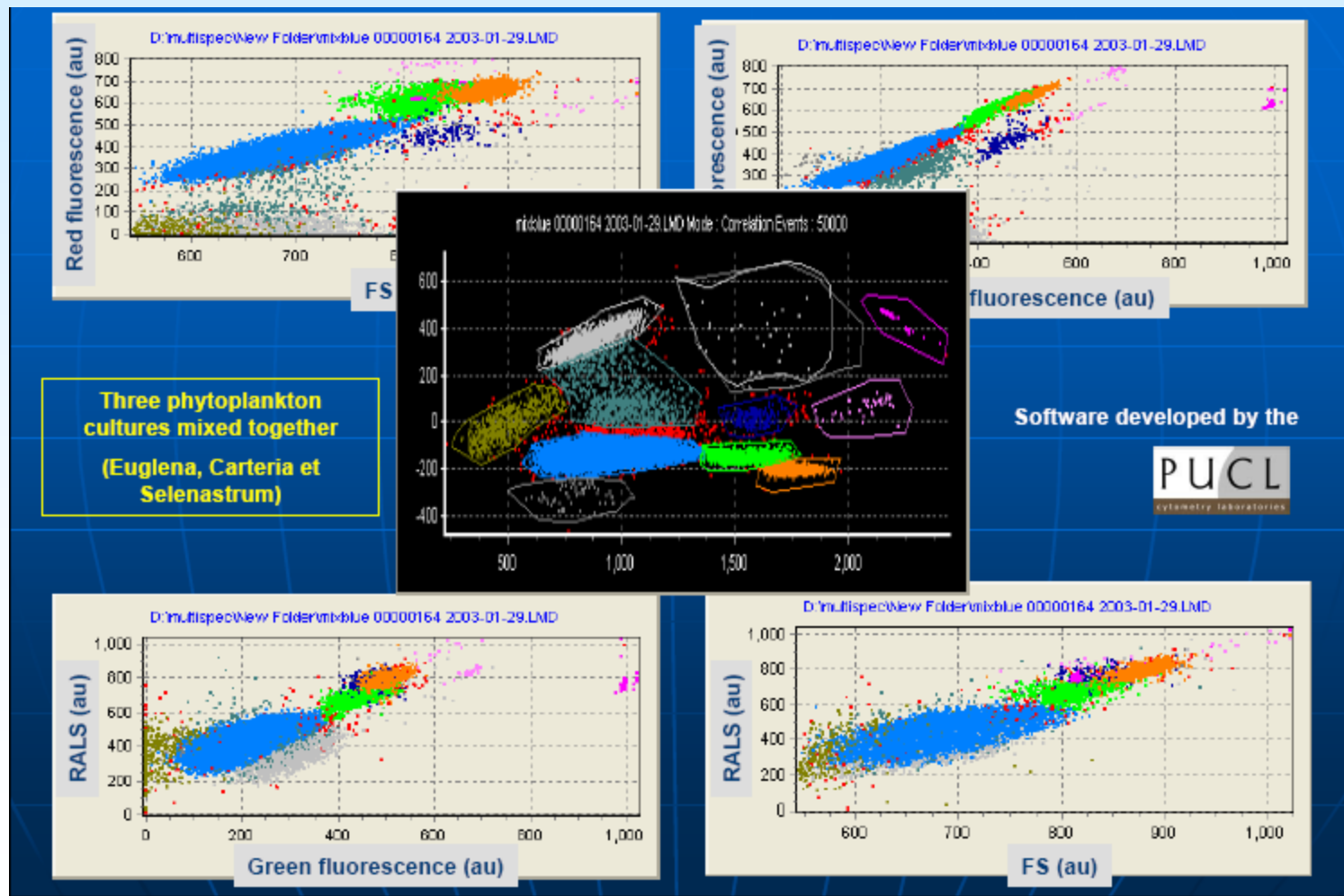
The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean.



The blue arrow is an eigenvector of this shear mapping, and since its length is unchanged its Eigenvalue is 1.



Άλλα παραδείγματα της PCA (G. Gregori)



Η κύρια συνιστώσα συνεισφέρει περισσότερο στη διακύμανση των δεδομένων

- Δηλαδή, η πρώτη κύρια συνιστώσα συνεισφέρει όσο το δυνατόν περισσότερο στη διακύμανση των δεδομένων, και κάθε επόμενο στοιχείο έχει την υψηλότερη δυνατή συνεισφορά.
- Ο περιορισμός είναι να είναι ορθογώνιος προς την διακύμανση, δηλαδή μη συσχετιζόμενος με τα παραπάνω στοιχεία.
- Η ανεξαρτησία είναι εγγυημένη, εφόσον τα Δεδομένα ακολουθούν κανονική κατανομή.
- Η PCA είναι η απλούστερη από τις μεθόδους πολλών μεταβλητών, που βασίζονται στα Ιδιοδιανύσματα.
- Συχνά, η λειτουργία μπορεί να θεωρηθεί, ότι αποκαλύπτει την εσωτερική δομή των δεδομένων, κατά τρόπο που να εξηγεί καλύτερα τη διακύμανση των δεδομένων.



E. Lugli et al. "Subject Classification Obtained by Cluster Analysis and Principal Component Analysis Applied to Flow Cytometric Data
Cytometry Part A 71A:334–344 (2007)

First, T cells were identified by gating on lymphocytes on the basis of FSC and SSC, then by gating on CD31, CD41 or CD31, CD81 cells.

Multiple gates were subsequently drawn on the basis of positivity and negativity of each indicated antigen.

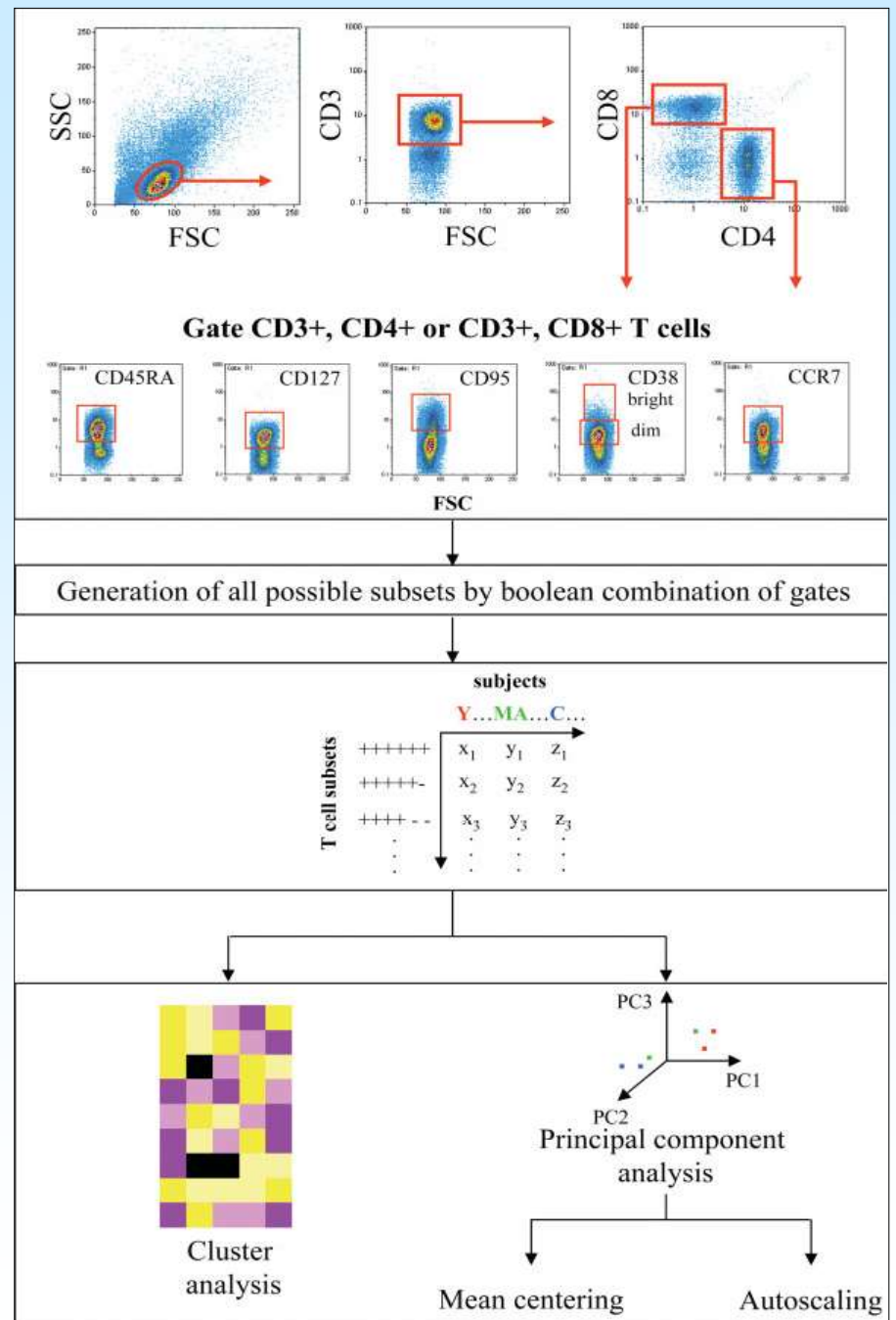
In the case of CD38, the expression was further distinguished between dim and bright.

All possible phenotypes were next generated by performing automatic Boolean combinations.

The data obtained were put in a matrix whose rows indicate T cell subsets and columns indicate subjects.

The entire data sets were analyzed by CA and PCA.

In the latter case, both mean centering and auto-scaling were tested.



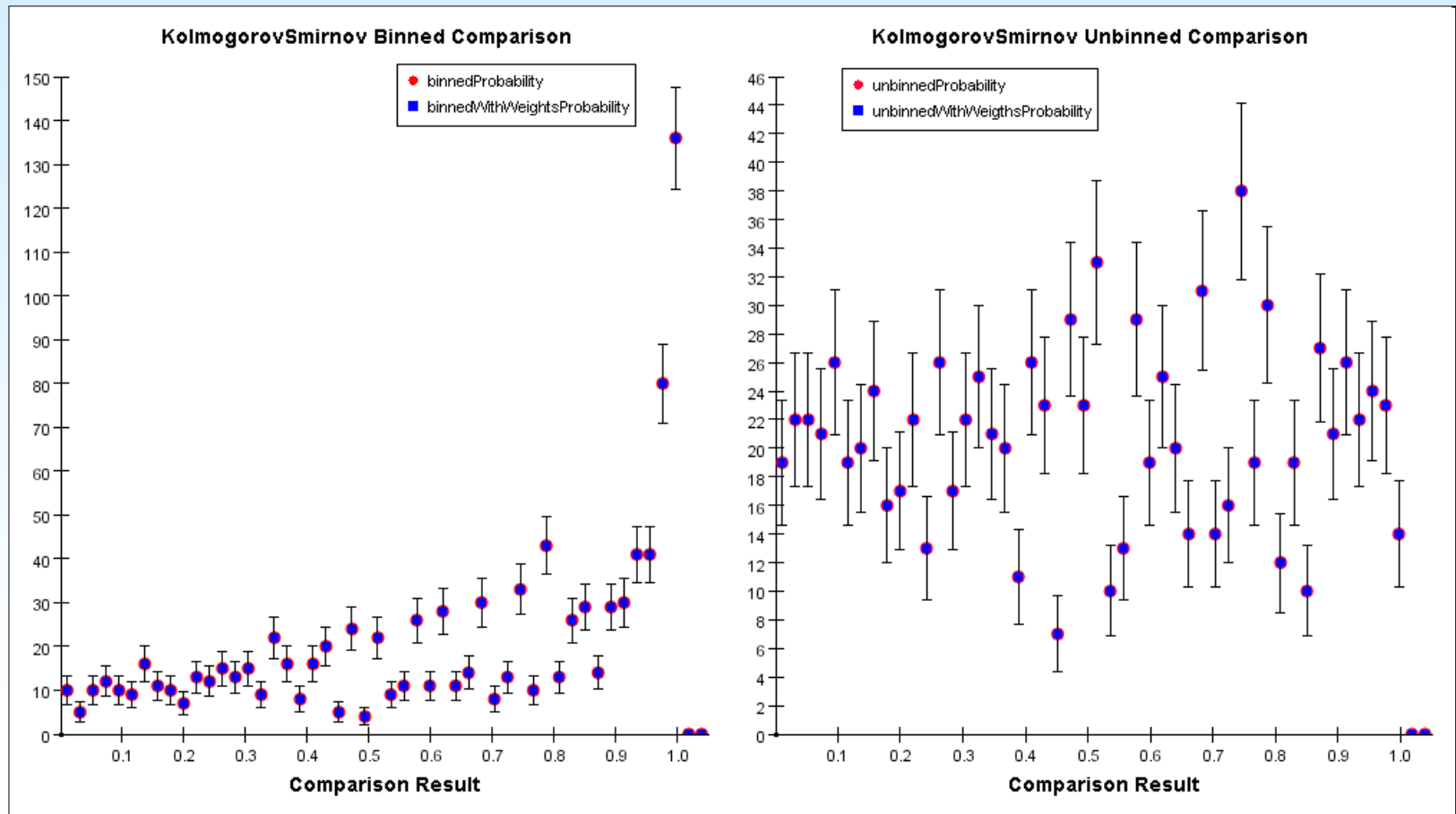
Η δοκιμασία Kolmogorov–Smirnov (K–S test)

- Η δοκιμασία Kolmogorov–Smirnov (K–S test) είναι μια μη παραμετρική δοκιμή, για την ισότητα συνεχών, μονοδιάστατων κατανομών πιθανότητας.
- Μπορεί να χρησιμοποιηθεί για να συγκριθεί ένα δείγμα με μια κατανομή πιθανότητας αναφοράς (δοκιμή K–S ενός δείγματος), ή για να συγκριθούν δύο δείγματα μεταξύ τους (δοκιμή K–S δύο δειγμάτων).
- Η στατιστική Kolmogorov–Smirnov ποσοτικοποιεί την απόσταση μεταξύ της εμπειρικής συνάρτησης κατανομής του δείγματος και της αθροιστικής συνάρτησης κατανομής της κατανομής αναφοράς, ή μεταξύ των συναρτήσεων εμπειρικής κατανομής των δύο δειγμάτων.

Η δοκιμή K–S για τη σύγκριση δύο δειγμάτων

- Η δοκιμή K–S δύο δειγμάτων είναι μία από τις πιο χρήσιμες μη παραμετρικές μεθόδους, **για τη σύγκριση δύο δειγμάτων**, καθώς είναι ευαίσθητη σε διαφορές **τόσο στη θέση, όσο και στο σχήμα**, των εμπειρικών συναρτήσεων αθροιστικής κατανομής των δύο δειγμάτων.
- Η δοκιμή Kolmogorov–Smirnov μπορεί να τροποποιηθεί για να χρησιμεύσει **ως ένα κριτήριο καλής προσαρμογής** (fitting) μιας κατανομής.
- Στην ειδική περίπτωση των **δοκιμών για την κανονικότητα** της κατανομής, τα δείγματα είναι **τυποποιημένα** και συγκρίνονται με μία **τυπική κανονική κατανομή**.

Παράδειγμα εφαρμογής της δοκιμασίας K-S

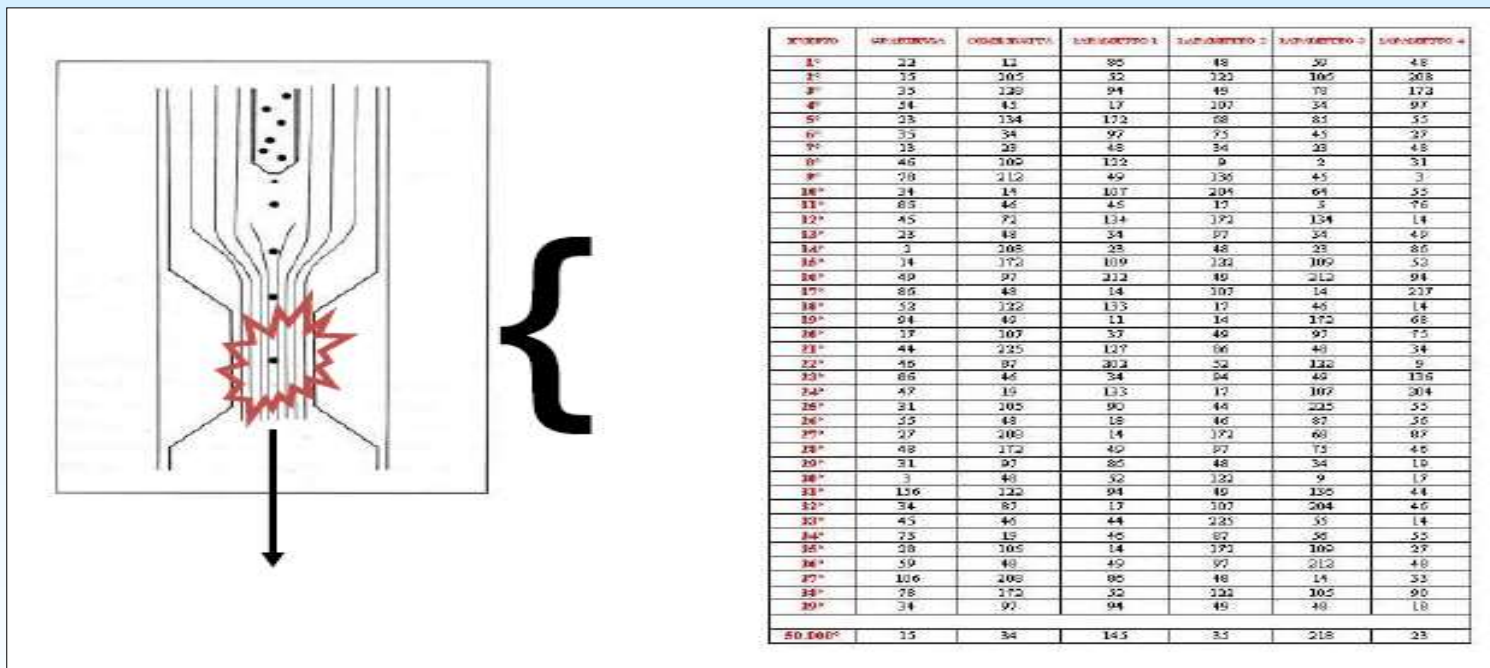


The outcome of a Kolmogorov-Smirnov test is a probability, whose distribution is shown.

Ενδεικτικά εμπορικά ή δωρεάν διαθέσιμα λογισμικά Συλλογής ή/και Ανάλυσης Δεδομένων

Λογισμικό Συλλογής & Ανάλυσης Δεδομένων	Λογισμικό Ανάλυσης Δεδομένων	Βοηθητικό Λογισμικό
Consort 30	Flow-Jo	Flow Explorer
Diva	Infinicyt	Cluster, Trees κλπ.
Expo 32	Kaluza	
	Paint-a-Gate	
	Winlist	
	Win-MDI	

Σύγχρονες τάσεις: Ενιαίο Λογισμικό Κυτταρομετρίας Ροής για όλες τις χρήσεις



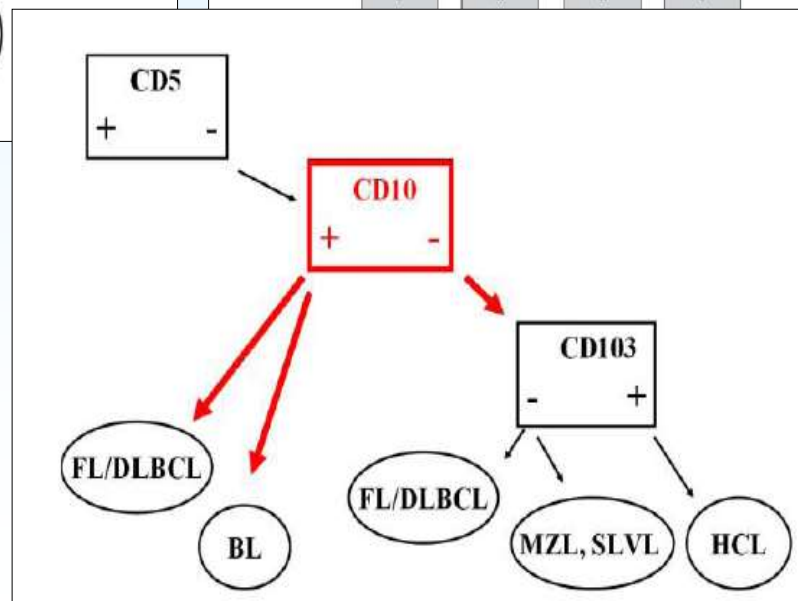
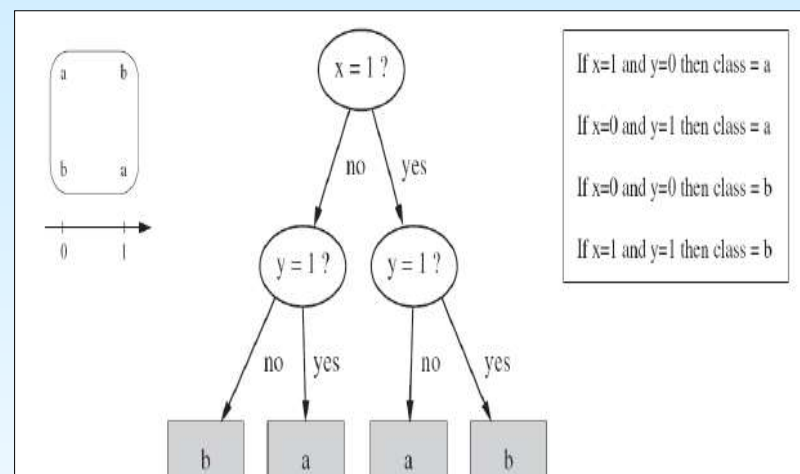
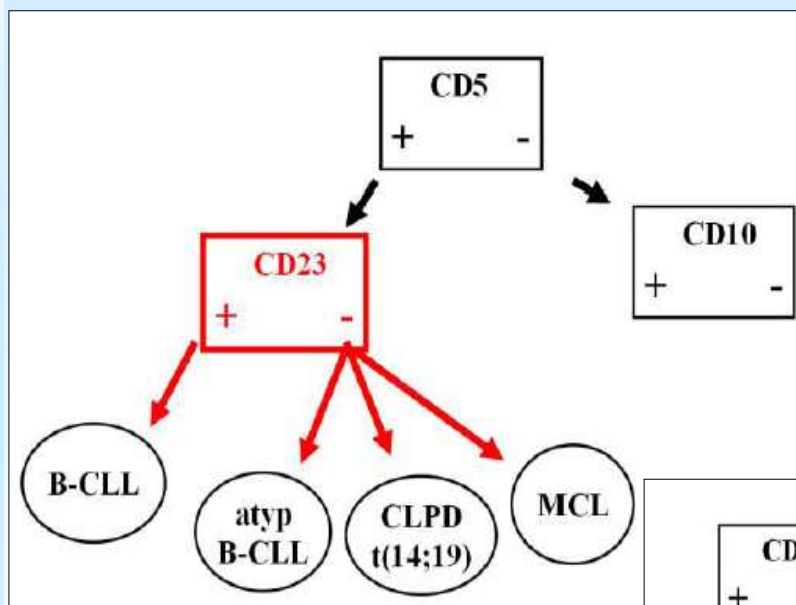
- Συλλογή δεδομένων μέσω του **data-bus** του K/P.
- Έλεγχος του K/P.
- Γραφικά αναπαράστασης των δεδομένων.
- Ανάλυση (επεξεργασία) των δεδομένων.
- Βοηθητικές λειτουργίες.

Βοηθητικές λειτουργίες

- Προχωρημένες αναπαραστάσεις και αναλύσεις δεδομένων μέσω περίπλοκων Μαθηματικών συναρτήσεων (Biexpro, Hyperlog κλπ.).
- Πρόσβαση και επεξεργασία του Κυτταρομετρικού Αρχείου.
- Αξιολόγηση και έλεγχος της ποιότητας των μετρήσεων.
- Διαχείριση των Αρχείων Δεδομένων.
- Δραστηριότητα «Εξόρυξης Δεδομένων», δηλαδή εξαγωγής λανθανόντων, προηγούμενων αγνώστων και ενδεχομένως χρήσιμων πληροφοριών από αυτά.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	AG RFI	4	5	6	7	11	12	13	14	15	16	17	18	19	22	23	24	25	26
2	CD5	13,6	3,8	5,8	2,0	21,7	1,0	2,7	62,2	3,1	75,8	93,7	3,7	16,8	16,1	19,5	66,3	2,36	
3	CD79b	1,8	1,7	10,1	14,8	4,0		0,9		21,9		2,5	3,5	5,1	4,1	2,4	3,1	4,62	
4	CD85j	7,5	4,3	8,6	12,6	4,5	10,5	4,9	5,5	11,6	14,3	7,4	14,2	9,7		3,8	9,3	5,97	18,39
5	CD11c	1,3	1,1	1,3	1,1	1,1	1,5	0,4	1,6	1,7		1,1	0,9	1,1	2,7	5,4	1,1	26,27	3,93
6	CD22	1,8	3,8	1,8	4,9	1,1	34,4	2,3	3,9	8,1		1,6	2,3	3,0	19,2	16,8	31,4	111,82	
7	CD23	2,1	1,7	1,2	1,5	2,2	1,1	2,0	57,0	22,4	3,3	1,9	2,4	3,7	34,8	27,5	64,8	238,21	
8	CD24	921,8	890,8	334,1	545,0	146,1	42,0	1516,7	50,2	686,0		38,0		149,2	467,9	119,7	212,8	643,72	59,96
9	LAIR	43,9	5,8	2,8	4,1	79,0	14,0	83,7	38,1	58,4	5,3	6,8	1,6	1,0	80,7	0,7	0,4	41,94	40,93
10	CD20	6,0	2,5	25,3	5,3	1,3	69,7	1,3	4,0		4,5	15,5		3,8	6,8	5,4	9,6	6,17	3,56
11	LC	17,9		33,5	49,1	64,8	349,0	0,9	12,3	74,4	13,7	33,9	52,8	22,2	190,8		25,0		
12																			
13																			
14																			
15																			

Εξόρυξη Δεδομένων μέσω Δενδροδιαγραμμάτων

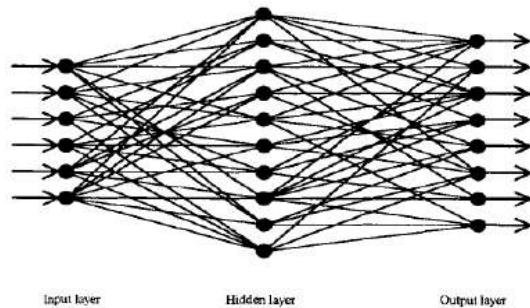


Witten & Frank

Εξόρυξη Δεδομένων μέσω Νευρωνικών Δικτύων

Neural Network Analysis of Flow Cytometric Data for 40 Marine Phytoplankton Species

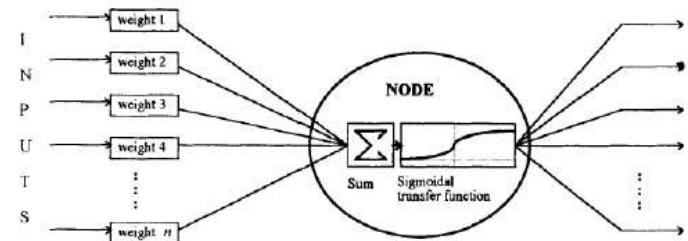
Lynne Boddy, C.W. Morris, M.F. Wilkins, G.A. Tarran, and P.H. Burkill



Cytometry 15:283-293 (1994)

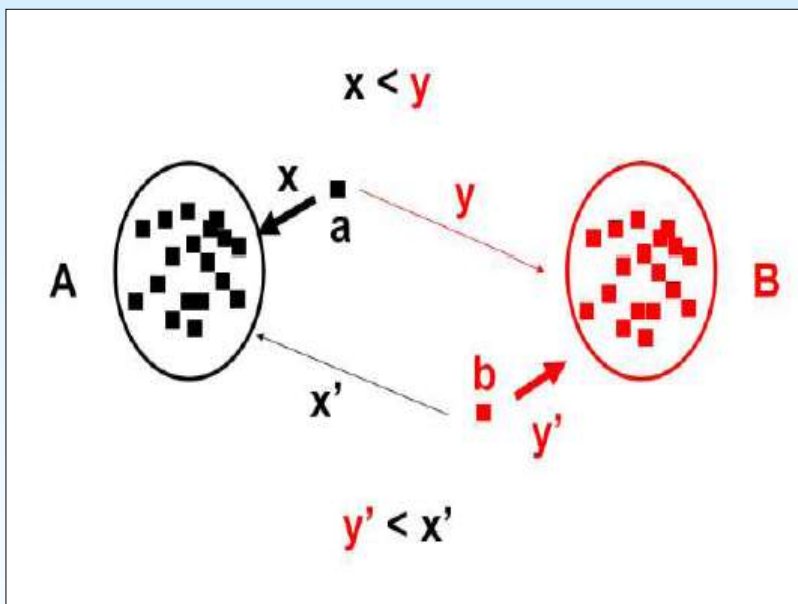
Neural Network Analysis of Flow Cytometric Data for 40 Marine Phytoplankton Species

Lynne Boddy, C.W. Morris, M.F. Wilkins, G.A. Tarran, and P.H. Burkill



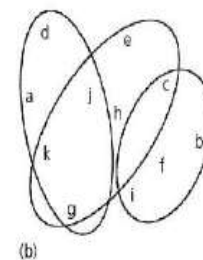
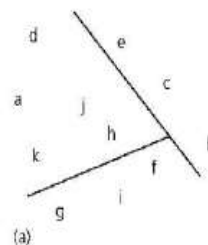
Cytometry 15:283-293 (1994)

Εξόρυξη Δεδομένων μέσω Ανάλυσης Ομάδων



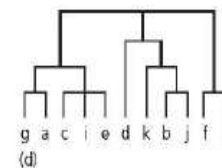
Αρχή της Απόστασης

Μορφές Αναπαράστασης Clusters



	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

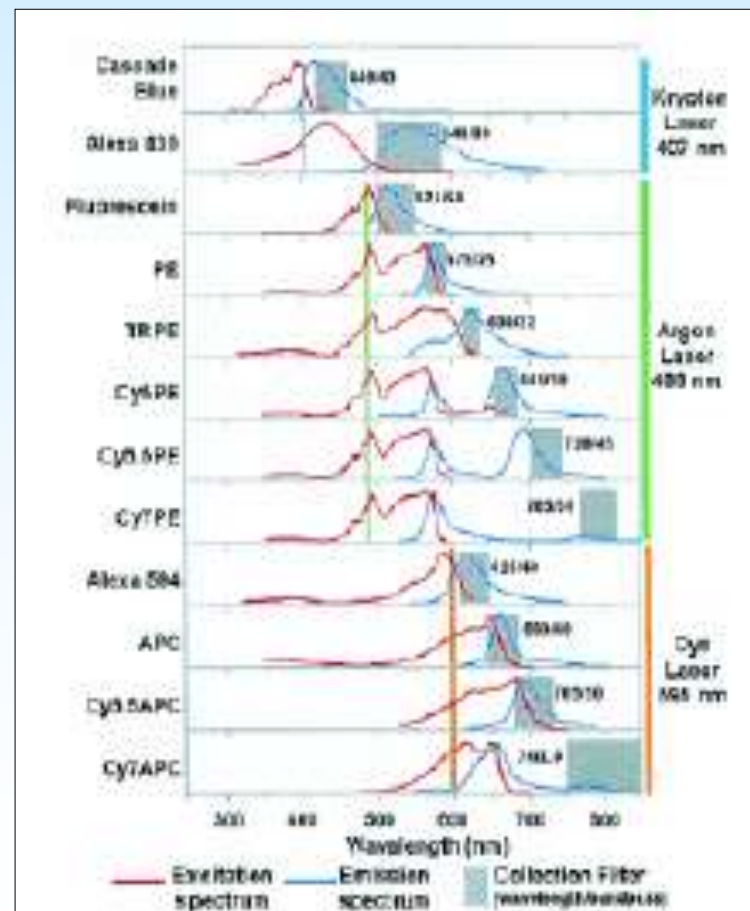
(c)



Witten & Frank

Συμπεράσματα

- Η τεχνολογική εξέλιξη φέρνει ήδη:
 - ◆ *Νέες πηγές (LEDs, solid state LASERS).*
 - ◆ *MC-PMTs, Avalanche Diodes.*
 - ◆ *Ταχύτητα Ηλεκτρονικά & H/Y.*
 - ◆ *Miniaturization.*
 - ◆ *Lower costs.*
 - ◆ *New fluorescent compounds as organic dyes & nanocrystals.*
- Αυτό σημαίνει νέα δεδομένα:
 - ◆ *Ανά σωματίδιο-κύτταρο.*
 - ◆ *Πολυπαραμετρικότητα ↑↑↑*
 - ◆ *Νέα Φάσματα.*
 - ◆ *Νέα δεδομένα όγκου.*
- Η Μαθηματική επεξεργασία γίνεται ακόμα πιο σημαντική στο μέλλον!



Και μη ξεχνάτε, η **Στατιστική** μοιάζει με το **bikini**:
Δείχνει πολλά αλλά πρέπει να μαντεύετε τα ουσιώδη...

