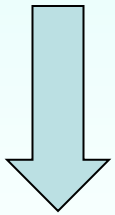
Where is the wisdom lost in knowledge?
G.Eliott

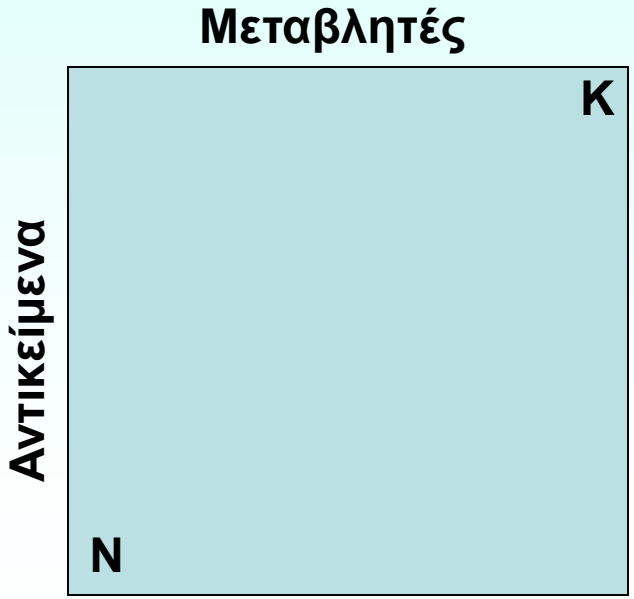
Where is the knowledge lost in information?

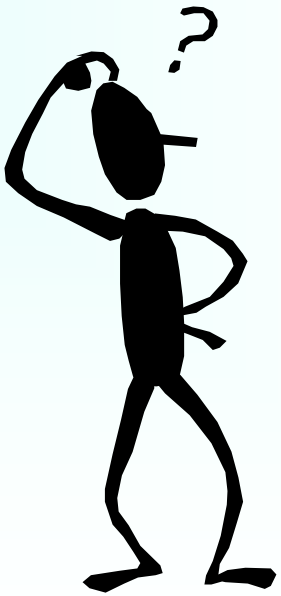


Ο κόσμος της πληροφορίας

Δεδομένα, δεδομένα, δεδομένα.....

Πολυμεταβλητές μετρήσεις- πληροφορίες, πληροφορίες, πληροφορίες....





Πολυπαραμετρικές Στατιστικές Μέθοδοι

Οι **Πολυπαραμετρικές Στατιστικές Μέθοδοι** επιτρέπουν:

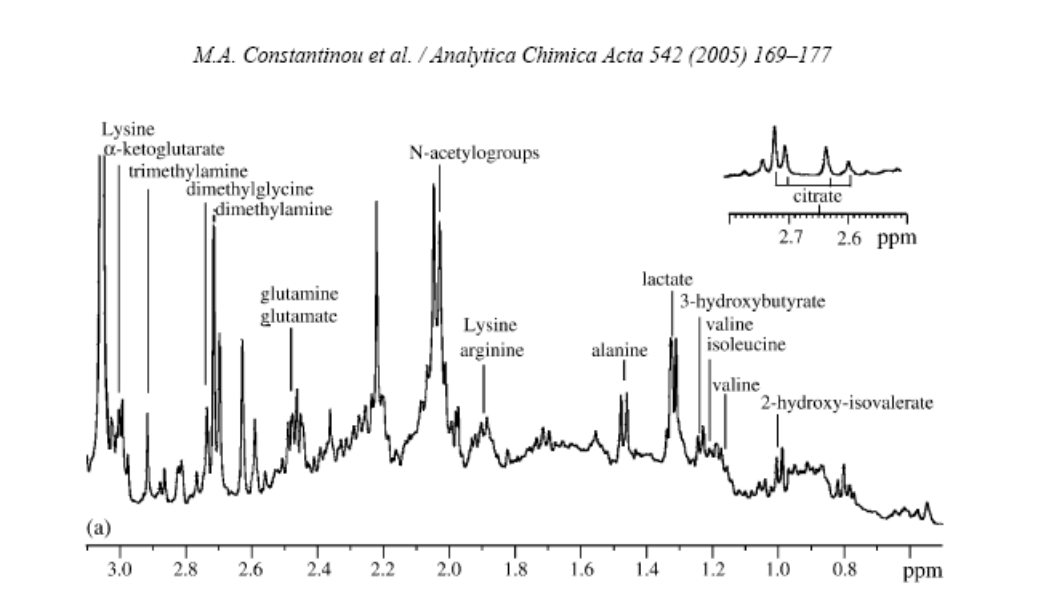
μέγιστη αξιοποίηση των πληροφοριών που περιέχονται σε μια σειρά (**πολυμεταβλητών**) δεδομένων

και ορισμένες από αυτές παράλληλα επιτρέπουν:

ελαχιστοποίηση του αριθμού των αντικειμένων που απαιτούνται την εξαγωγή ενός μοντέλου.

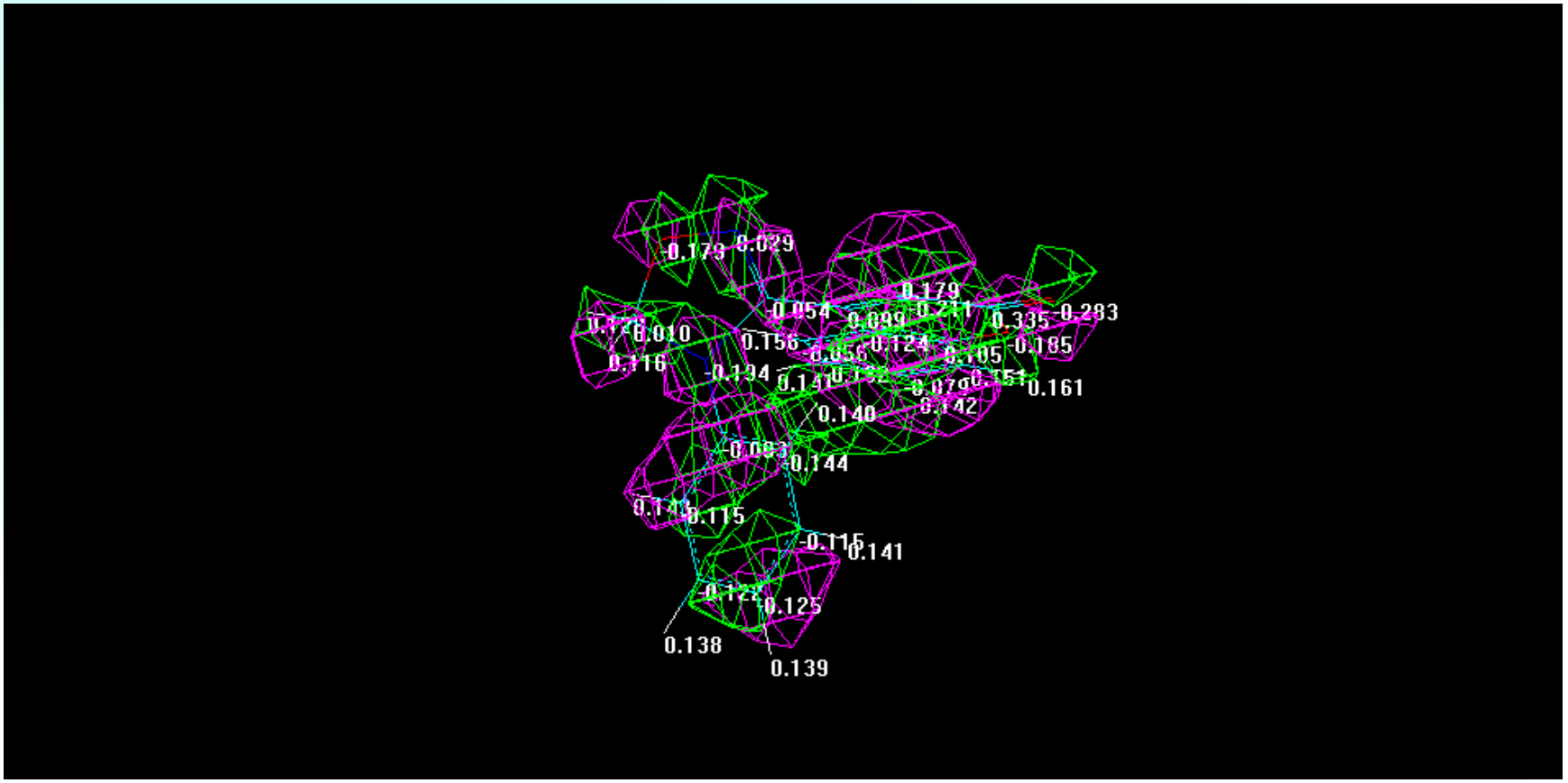
Πολυμεταβλητές μετρήσεις

Το φάσμα ^{1}H-NMR ενός δείγματος αποτελεί πολυμεταβλητή μέτρηση



Πολυμεταβλητές Ιδιότητες

Κατανομή ηλεκτρονιακού νέφους



Πολυπαραμετρικές Στατιστικές Μέθοδοι

Ταξινόμηση-Κατηγοριοποίηση
Classification methods, pattern recognition

Εξαγωγή μοντέλου για Ποσοτικές Προβλέψεις

Πολυπαραμετρικές Στατιστικές Μέθοδοι

Κλασικές πολυπαραμετρικές μέθοδοι

Μέθοδοι Προβολής- Projection methods
Πολυμεταβλητή Ανάλυση δεδομένων
Multivariate data analysis

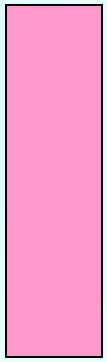
Γραμμικές μέθοδοι

Τεχνητά Νευρωνικά Δίκτυα
Artificial Neural Networks

Μηχανές Διανυσμάτων Στήριξης
Support Vector Machines

Μη γραμμικές μέθοδοι

Γραμμικές πολυπαραμετρικές στατιστικές μέθοδοι

	Πίνακας δεδομένων	Προϋποθέσεις
Κλασικές πολυπαραμετρικές στατιστικές μέθοδοι: *Linear Discriminant Analysis* *Canonical Correlation* *Cluster Analysis, ανάλυση σμηνών* ***Πολλαπλή γραμμική ανάλυση παλινδρόμησης***	Μακρύς  και στενός	•Μεταβλητές Χ ανεξάρτητες •Μεταβλητές Χ ακριβείς •Μεταβλητές Χ πλήρεις •Τυχαία κατανομή υπολοίπων
Πολυμεταβλητή Ανάλυση Δεδομένων(Ανάλυση πολυμεταβλητών δεδομένων) - Μέθοδοι Προβολής ***Ανάλυση Κυρίων Συνιστωσών, PCA*** ***Προβολή σε λανθάνουσες Δομές, PLS*** ***PLS-DA***	Κοντός και πλατύς 	•Μεταβλητές Χ όχι ανεξάρτητες •Μεταβλητές Χ όχι ακριβείς •Μεταβλητές Χ όχι πλήρεις •Τα υπόλοιπα μπορεί να έχουν δομημένη κατανομή

Ταξινόμηση

- Διατύπωση κανόνων για τη ταξινόμηση και υπαγωγή διαφόρων αντικειμένων σε συγκεκριμένες τάξεις με καθορισμένη συμπεριφορά (χημική βιολογική κλπ)

Ταξινόμηση και Ανάλυση σμηνών

Ταξινόμηση Classification, discriminant analysis	Ανάλυση σμηνών Cluster Analysis
• γνωστός αριθμός τάξεων	• άγνωστός αριθμός τάξεων
• βασίζεται σε σειρά εκμάθησης	• δεν υφίσταται προηγούμενη γνώση
• Αποτελεί επιβλεπόμενη τεχνική (supervised learning)	• εφαρμογή για την κατανόηση των δεδομένων
	• Αποτελεί μη επιβλεπόμενη τεχνική (unsupervised learning)
• Εφαρμογή στην ταξινόμηση αγνώστων δειγμάτων	•Εφαρμογή στην ταξινόμηση αγνώστων δειγμάτων

Ταξινόμηση

Τα δεδομένα διατάσσονται υπό μορφή πίνακα mxn,

- m οι τιμές μεταβλητών
- n τα αντικείμενα, οι παρατηρήσεις.
- Κάθε αντικείμενο συνιστά σημείο ενός m- διάστατου χώρου, στον οποίο κάθε μεταβλητή ορίζει έναν **ορθογώνιο** άξονα, ενώ το σύνολο των αρχικών δεδομένων απεικονίζεται υπό τη μορφή σμήνους n σημείων σ' έναν m-χώρο.
- Συχνά απαιτείται μια ομάδα εκμάθησης (training set) και μία ομάδα εξέτασης (test set)
- Εάν η τεχνική είναι επιβλεπόμενη υπάρχει κα η στήλη y που καθορίζει την τάξη

Ταξινόμηση

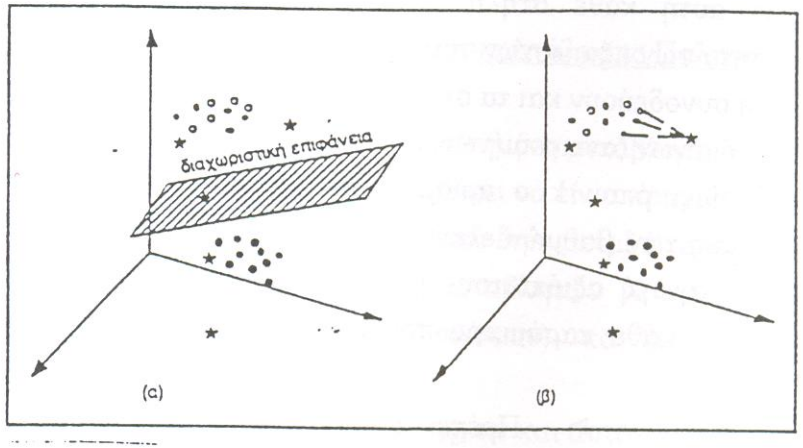
Οι μέθοδοι ταξινόμησης βασίζονται σε κριτήρια **ομοιότητας** στο σύνολο των πειραματικών σημείων.

Τα αντικείμενα των οποίων τα σημεία βρίσκονται πλησιέστερα στον χώρο m διαστάσεων ομαδοποιούνται με βάση την (Ευκλείδια) απόσταση (σμήνος, ομάδα).

- Στην LDA οι τάξεις διαχωρίζονται στον m- χώρο από μια επιφάνεια m-1 διαστάσεων ⟶ Δεν μπορούν οι μέθοδοι αυτές να χρησιμοποιηθούν στην ανάλυση πολλών μεταβλητών : m <n/3

- Στην KNN μετά τον διαχωρισμό σε τάξεις τα αντικείμενα της ομάδας ελέγχου ταξινομούνται σύμφωνα με τα k πλησιέστερα σημεία (k συνήθως = 3)

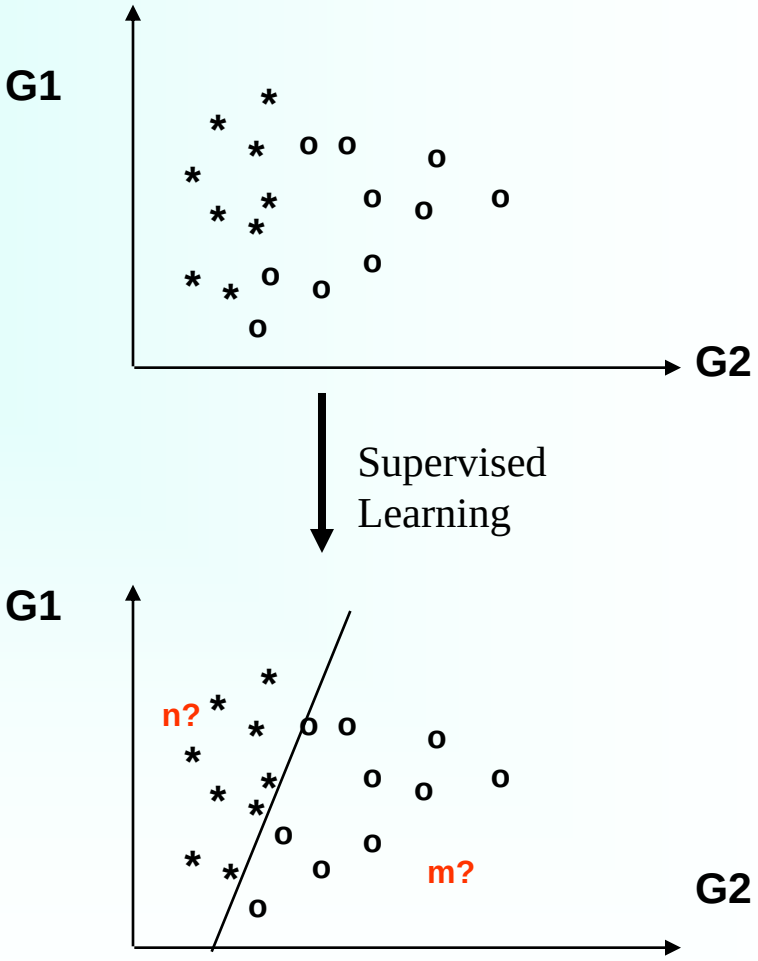
Αναγνώριση Προτύπων, Pattern Recognition



Σχήμα 1. Γραφικές απεικονίσεις των μεθόδων (α) LDA-LLM και (β) KNN. (Με αστερίσκο δηλώνονται τα αντικείμενα της 'ομάδας εξέτασης')

Ταξινόμηση

- Επισημαίνεται η τάξη (ο και x), σύμφωνα με τις τιμές στους άξονες G1 και G2 (G3..Gn).
- Αναζητείται μοντέλο που διαχωρίζει τα δεδομένα στις υφιστάμενες τάξεις
- Ακολούθως εφαρμόζεται το μοντέλο για να ταξινομηθούν άγνωστα δεδομένα n καιd m στη σωστή τάξη



Linear Discriminant Analysis

- Προτάθηκε απο τον Fisher (1936) για την ταξινόμηση μιας παρατήρησης σε μια από δυο δυνατές τάξεις

  στη βάση πολλών μετρήσεων $x_1, x_2, \ldots x_p$.

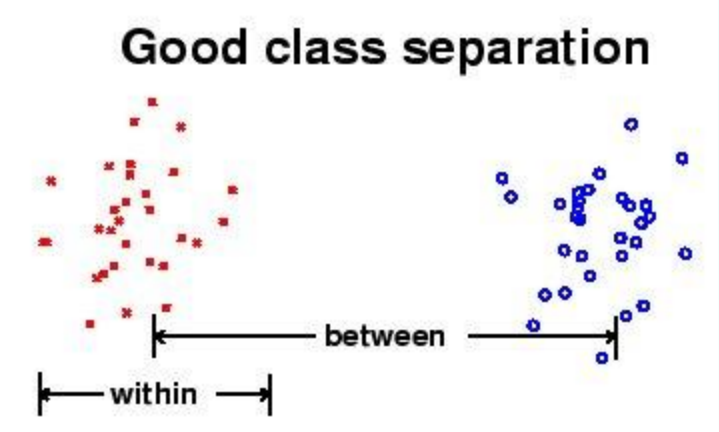
- Αναζητείται γραμμική μετατροπή των μεταβλητών

  $Y = a_1 x_1 + a_2 x_2 + .. + a_p x_p$

  ετσι ώστε ο διαχωρισμός μεταξύ των **μέσων όρων** των ομάδων στην μεταβληθείσα κλίμακα να είναι ο καλύτερος δυνατός

Πώς υπολογίζονται οι συντελεστές a_i ?

- Οι συντελεστές υπολογίζονται έτσι ώστε να **μεγιστοποιούν** το λόγο του αθροίσματος των τετραγώνων μεταξύ των ομάδων ως προς άθροισμα των τετραγώνων εντός της ομάδας, δηλ η απόσταση μεταξύ των ομάδων πρέπει να είναι μέγιστη και η απόσταση εντός των ομάδων ελάχιστη



Ευκλείδεια απόσταση

Είναι απλά η γεωμετρική απόσταση

$$\text{distance}(x,y) = \{ \Sigma_i (x_i - y_i)^2 \}^{1/2}$$

Η ευκλείδεια απόσταση επηρεάζεται απο το μέγεθος του πληθυσμού

Αποστάσεις

Απόσταση Manhattan

$$distance(x,y) = \Sigma_i |x_i - y_i|$$

Απόσταση Chebychev

$$distance(x,y) = Maximum|x_i - y_i|$$

Απόσταση σε δύναμη

$$distance(x,y) = (\Sigma_i |x_i - y_i|^p)^{1/r}$$

Ποσοστό ανομοιότητας

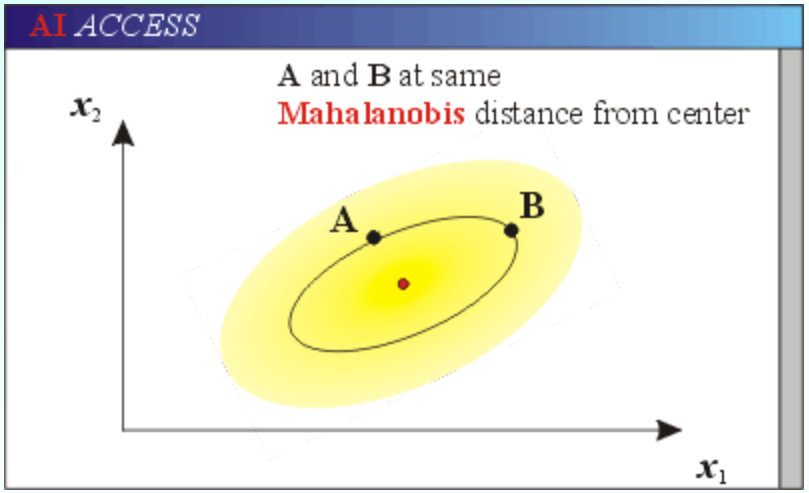
distance(x,y) = (Number of xi yi)/ i

Απόσταση Mahalanobis

$$(X_i - Y_i)' S_i^{-1} (X_i - Y_i)$$

X οι επί μέρους τιμές των μεταβλητών .
Y οι αντίστοιχοι μέσοι όροι, ή τα κέντρα βάρους
S-1 το αντίστροφο του πίνακα συνδιακύμανσης

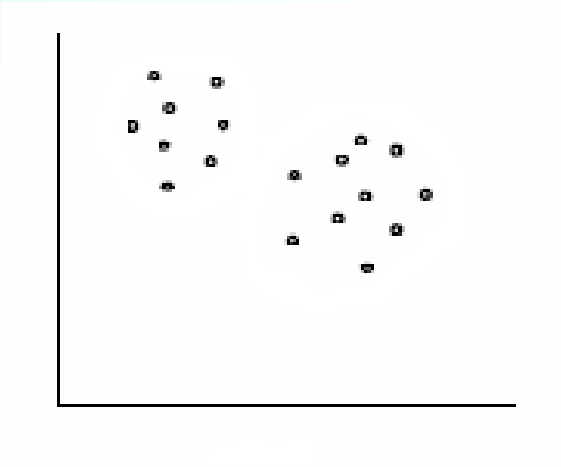
Απόσταση Mahalanobis



Αρχές Ομοιογένειας και Διαχωρισμού

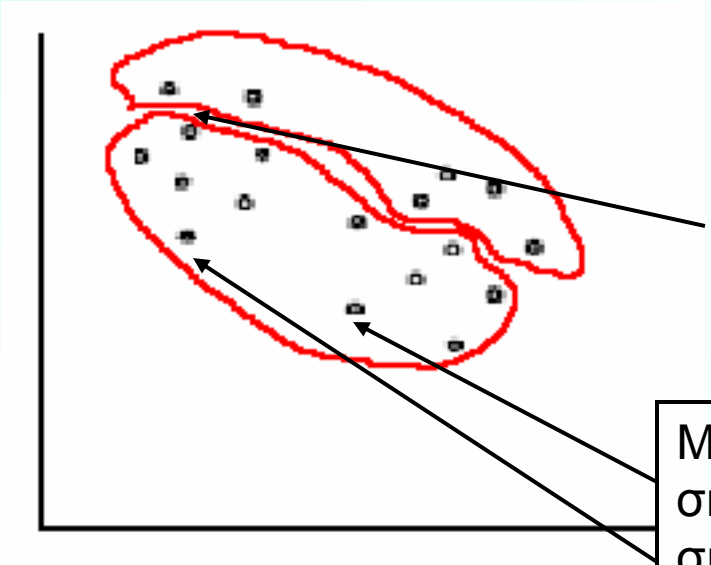
- **Ομοιογένεια:** Τα αντικείμενα εντός του σμήνους είναι κοντά μεταξύ τους
- **Διαχωρισμός:** Τα αντικείμενα σε διαφορετικά σμήνη ειναι απομακρυσμένα μεταξύ τους

Ενας δεδομένος αλγόριθμος θα μπορούσε να διαχωρίσει τα αντικείμενα σε σμήνη



Κακή ανάλυση σμηνών

Ο ανάλυση αυτή (βάσει κάποιου αλγορίθμου) παραβιάζει και την ομοιογένεια και το διαχωρισμό

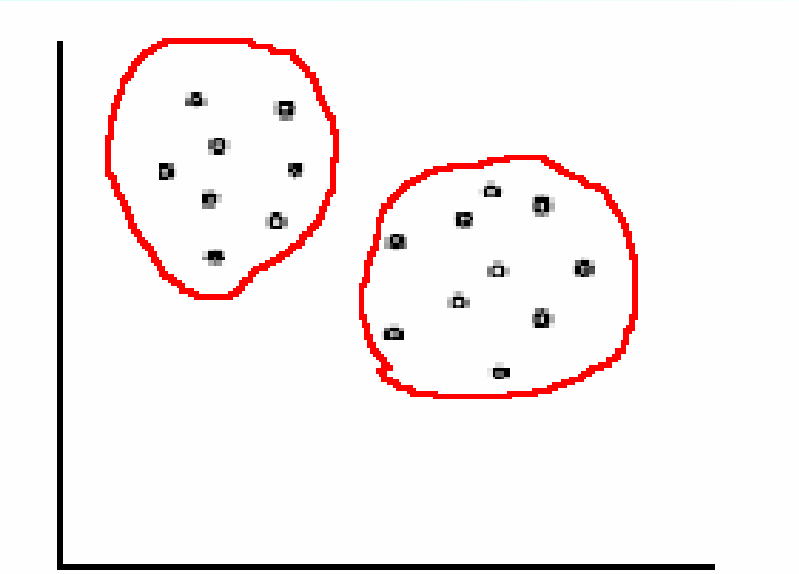


Μικρές αποστάσεις μεταξύ σημείων σε διαφορετικά σμήνη

Μεγάλες αποστάσεις σημείων εντός του ιδίου σμήνους.

Καλή ανάλυση σμηνών

Η ανάλυση αυτή εμφανίζει **καλή** ομοιογένεια και **καλό** διαχωρισμό



Μέθοδοι Ταξινόμησης-Αναγνώριση προτύπων

Ανάλυση Κυρίων Συνιστωσών,
Principal Component Analysis, PCA

Μέθοδος προβολής των σημείων από έναν πολυδιάστατο χώρο σε ένα χώρο λιγότερων διαστάσεων

Πολυπαραμετρική Ανάλυση

Εξαγωγή μοντέλου για ποσοτικές προβλέψεις

- **Πολλαπλή γραμμική ανάλυση παλινδρόμησης**
  Multiple Linear Regression Analysis (MLRA)

- **Μη γραμμική ανάλυση παλινδρόμησης**

- **Ανάλυση μερικών Ελαχίστων Τετραγώνων,**
  Προβολή σε Λανθάνουσες Δομές
  Partial Least Squares, Projection to Latent Structures

Εξαγωγή του κατάλληλου μοντέλου με χρήση στατιστικών μεθόδων

- Πολλαπλή γραμμική ανάλυση παλινδρόμησης
  Multiple Linear Regression Analysis (MLRA)

- **Ανάλυση μερικών Ελαχίστων Τετραγώνων,**
  Προβολή σε Λανθάνουσες Δομές
  Partial Least Squares, Projection to Latent Structures

Πολλαπλή γραμμική ανάλυση παλινδρόμησης

- Η πολλαπλή γραμμική ανάλυση παλινδρόμησης επιτρέπει την συσχέτιση μιας εξαρτημένης μεταβλητής με περισσότερες ανεξάρτητες μεταβλητές (παραμέτρους).

- Δημιουργείται ένας πίνακας δεδομένων, όπου οι σειρές αντιστοιχούν στις διαφορετικές ενώσεις και οι στήλες στις μεταβλητές. Η πρώτη στήλη αφορά στην εξαρτημένη μεταβλητή Υ και οι υπόλοιπες στις ανεξάρτητες μεταβλητές (παραμέτρους) Χ.

Πολλαπλή γραμμική ανάλυση παλινδρόμησης

- Εξάγεται γραμμική εξίσωση της μορφής:

Ιδιότητα = $\alpha_o + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots + \alpha_n X_n$

X_1 - X_n :παράμετροι που περιγράφουν την ιδιότητα

α_1-α_n : συντελεστές που εξάγονται με τη μέθοδο των ελαχίστων τετραγώνων

Πολλαπλή γραμμική ανάλυση παλινδρόμησης

Η Πολλαπλή γραμμική ανάλυση παλινδρόμησης στηρίζεται στη μέθοδο των ελαχίστων τετραγώνων

Πολλαπλή **γραμμική** ανάλυση παλινδρόμησης

Εάν η αναμενόμενη σχέση δεν είναι γραμμική, γίνονται μαθηματικές μετατροπές στην εξαρτημένη ή τις ανεξάρτητες μεταβλητές

$y \longrightarrow \log y$

$\longrightarrow 1/y$

$x \longrightarrow x^n$

$\longrightarrow 1/x$

$\longrightarrow \sqrt{x}$

Προϋποθέσεις εφαρμογής Πολλαπλής γραμμικής ανάλυσης παλινδρόμησης

- Τιμές επαναλήψιμες και καλά κατανεμημένες μιας συγκεκριμένης ιδιότητας
- Επιλογή διαφορετικών, **μη αλληλοεξαρτώμενων** παραμέτρων (πειραματικών και/ή θεωρητικών) για την περιγραφή της δομής όλων των ενώσεων (για την περιγραφή του μελετώμενου φαινομένου)
- Μέθοδος που επιτρέπει τον έλεγχο αξιοπιστίας του μοντέλου που προέκυψε

Πολλαπλή γραμμική ανάλυση παλινδρόμησης- Στατιστικά Στοιχεία

- **n**: Αριθμός σειρών δεδομένων (αντικειμένων, ενώσεων)
- **n-k**: Βαθμοί ελευθερίας, όπου k ο αριθμός των μεταβλητών (ανεξαρτήτων + εξαρτημένης). Ο αριθμός των βαθμών ελευθερίας πρέπει να είναι όσο το δυνατόν μεγαλύτερος για την εξαγωγή αξιόπιστου μοντέλου QSAR (μεγάλο n, μικρό k).

Γενικά θεωρείται ότι σε κάθε παράμετρο πρέπει να αντιστοιχούν τουλάχιστον 5 ενώσεις.

•.

Πολλαπλή γραμμική ανάλυση παλινδρόμησης- Στατιστικά Στοιχεία

- **r**: Συντελεστής συσχετίσεως ⟶ | 1 |
- **r** ^{2}x100: Ποσοστό περιπτώσεων που ερμηνεύει η εξίσωση

Στην πολλαπλή γραμμική ανάλυση παλινδρόμησης ο συντελεστής συσχετίσεως προσαρμόζεται στους βαθμούς ελευθερίας δεδομένου ότι η εισαγωγή περισσότερων παραμέτρων οδηγεί σε πλασματική βελτίωση του r.

Πολλαπλή γραμμική ανάλυση παλινδρόμησης- Στατιστικά Στοιχεία

- **s**: Τυπική απόκλιση $\longrightarrow$ 0.

- **2s**: όριο ανοχής σφάλματος υπολογισμού της εξίσωσης. Η διαφορά Δ των υπολογιζόμενων τιμών $Y_{υπολ}$ από τις πειραματικές τιμές $Y_{πειρ}$ πρέπει είναι μικρότερη από 2s (Δ<2s), για να προσαρμόζονται τα αντικείμενα στο μοντέλο της εξίσωσης. Εάν Δ>2s το αντικείμενο αποτελεί εκροπη τιμή 'outlier'

Στην πολλαπλή γραμμική ανάλυση παλινδρόμησης η τυπική απόκλιση προσαρμόζεται στους βαθμούς ελευθερίας.

Πολλαπλή γραμμική ανάλυση παλινδρόμησης- Στατιστικά Στοιχεία

- Fischer test (F-test)**: καθορίζει το επίπεδο σημαντικότητας της εξίσωσης.**

- **Student test t**: Καθορίζει την σημαντικότητα κάθε παραμέτρου. Για να είναι σημαντική μια παράμετρος t>2

Για να θεωρείται στατιστικά σημαντική η εισαγωγή μιας επί πλέον παραμέτρου πρέπει εκτός των προϋποθέσεων που προκύπτουν από τα F- και t-test, να αυξάνει τουλάχιστον κατά 10% το ποσοστό περιπτώσεων που ερμηνεύει η εξίσωση (r^2 x100).

Επιλογή παραμέτρων

Εισαγωγή όλων των παραμέτρων (enter)

Σταδιακή εισαγωγή παραμέτρων με κριτήριο την σημαντικότητα της τιμής F ή την ίδια την τιμή F (stepwise)

Εισαγωγή παραμέτρων κατ' ακολουθία με κριτήριο την σημαντικότητα της τιμής F ή την ίδια την τιμή F (forward)

Εισαγωγή όλων των παραμέτρων και σταδιακό αποκλεισμό με κριτήριο την σημαντικότητα της τιμής F ή την ίδια την τιμή F (backward)

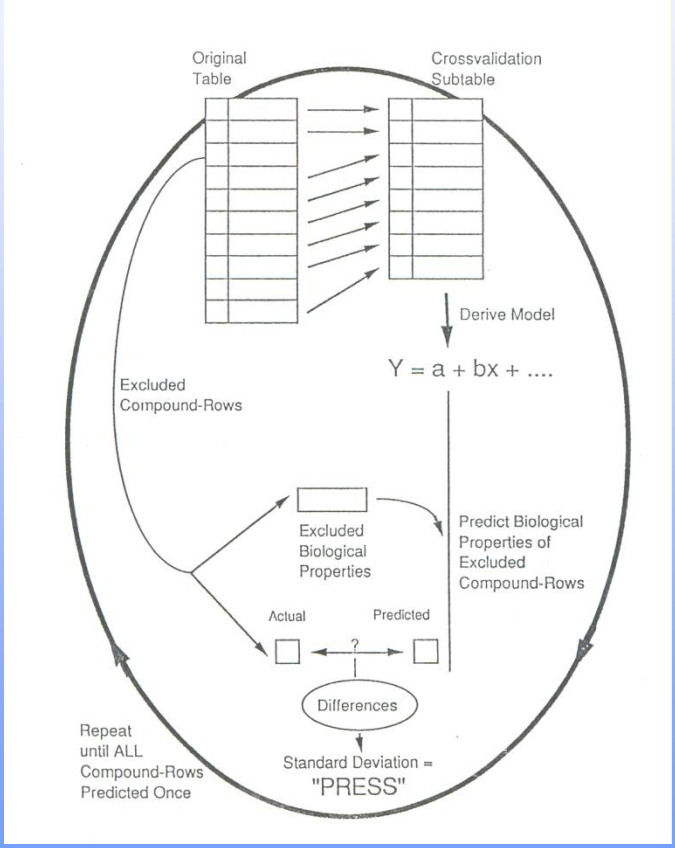
Tab. 12-10. Obere Signifikanzschranken der F-Verteilung für $S = 90\ \%$; ν_1 = Freiheitsgrade des Zählers; (s. Gl. (11-25), $\nu_1 = k$); ν_2 = Freiheitsgrade des Nenners; ($\nu_2 = n - k - 1$).

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,00	62,26	62,53
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85
16	3,05	2,67	2,46	2,33	2,24	2,18	2,14	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,75	1,72	1,69
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30

Πολλαπλή γραμμική ανάλυση παλινδρόμησης

Γενικά προτιμάται το μοντέλο να είναι όσον το δυνατόν απλούστερο, να περιέχει δηλαδή λίγες παραμέτρους.

Cross-Validation



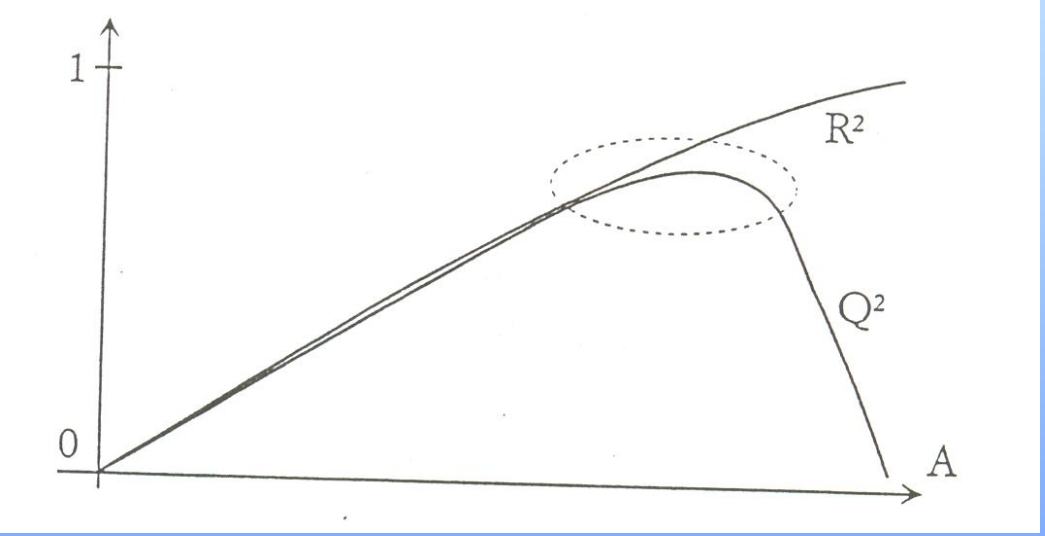
Cross-validation

“Crossvalidated r^2”

$$r^2_{CV} = 1.0 - \frac{\sum_{Y} (Y_{prédit} - Y_{mesuré})^2}{\sum_{Y} (Y_{mesuré} - Y_{moyen})^2}$$

<== Press

Σύγκριση R^2 και R^2_{cv} (Q^2)



Πολλαπλή γραμμική ανάλυση παλινδρόμησης

- Κάθε στήλη του πίνακα των X των ανεξάρτητων μεταβλητών υφίσταται ανεξάρτητη επεξεργασία
- Η μέθοδος αναλύει μια μόνο εξαρτημένη μεταβλητή Y
- Ο πίνακας των X πρέπει να έχει μεγάλο ύψος (αυξημένος αριθμός ενώσεων, δειγμάτων = αριθμός σειρών) και μικρό πλάτος (μικρός αριθμός μεταβλητών X= αριθμός στηλών)
- Οι μεταβλητές X πρέπει να είναι πραγματικά ανεξάρτητες (ορθογώνιες)
- Κίνδυνος: Τυχαία συσχέτιση

Πολυμεταβλητή Ανάλυση Δεδομένων
Multivariate Data Analysis

- Principal Component Analysis Ανάλυση Κυρίων Συνιστωσών
- Projections to Latent Structures, Partial Least Squares (PLS) Προβολές Μερικών Ελαχίστων Τετραγώνων σε Λανθάνουσες Δομές
- Projections to Latent Structures- Partial Least Squares Discriminant Analysis (PLS-DA)
- Multivariate Design Πολυμεταβλητός Σχεδιασμός

Πολυμεταβλητή Ανάλυση Δεδομένων
Multivariate Data Analysis

Εφαρμογές

- Έλεγχος Ποιότητας
- Βελτιστοποίηση Ποιότητας
- Διαδικασία Βελτιστοποίησης και Ελέγχου
- Ανάπτυξη και Βελτιστοποίηση Μεθόδων Προσδιορισμού
- Ταξινόμηση βακτηρίων, ιών, ιστών, κ.λ.π
- Metabonomics
- Ανάλυση οικονομικών μεγεθών
- Σχεδιασμός νέων φαρμάκων
- Ανάπτυξη νέων υλικών
- Προσδιορισμός επικινδυνότητας στην Τοξικολογία (Risk assessment)

Γιατί χρειάζεται η MVDA ?

- Συσσώρευση πλήθους πληροφοριών και δεδομένων
- Αύξηση του κόστους των πειραμάτων με ταυτόχρονη μείωση του κόστους πολλαπλών μετρήσεων στα εν εξελίξει πειράματα λόγω της σύγχρονης τεχνολογίας.
- Περιβαλλοντικοί περιορισμοί στη διεξαγωγή πειραμάτων
- Ηθικοί περιορισμοί στη διεξαγωγή πειραμάτων

Πότε χρειάζεται η MVDA ?

- Όταν πρόκειται να αναλύσουμε περισσότερες από 5 μεταβλητές για ένα σύνολο δεδομένων

- Όταν πρόκειται να αναλύσουμε περισσότερες εξαρτημένες μεταβλητές ταυτόχρονα

Πλεονεκτήματα MVDA

- Ο πίνακας των δεδομένων είναι στενός και μακρύς (Λίγα αντικείμενα, πολλές μεταβλητές
- Δεν επηρεάζεται από αλληλοσυσχέτιση (collinearity) μεταξύ των μεταβλητών
- Επιτρέπει την ανάλυση πολλών εξαρτημένων μεταβλητών ταυτόχρονα.

Πλεονεκτήματα MVDA

- Επιτρέπει προκαταρκτικό πειραματικό σχεδιασμό
- Οι μεταβλητές δεν απαιτείται να είναι ακριβείς
- Είναι δυνατόν να λείπουν τιμές στον πίνακα των δεδομένων
- Δεν ακολουθείται η φιλοσοφία της αλλαγής μιας μεταβλητής τη φορά

Μειονεκτήματα MVDA

- Απαιτείται σχετική εμπειρία στην ερμηνεία των αποτελεσμάτων
- Μια απλή συσχέτιση μπορεί να ‘κρυφτεί’

MVDA

Η πολυμεταβλητή αναλυση δεδομένων στηρίζεται στη μέθοδο των προβολών των σημείων (αντικειμένων) από ένα πολυδιάστατο χώρο σε ένα χώρο μικρότερων διαστάσεων.

Ανάλυση κυρίων συνιστωσών (PCA)

- Η PCA εφαρμόζεται σε ενιαίο πίνακα μεταβλητών X

$$N \bullet K$$

N: σειρές -αντικείμενα (παρατηρήσεις)
K: στήλες-μεταβλητές.

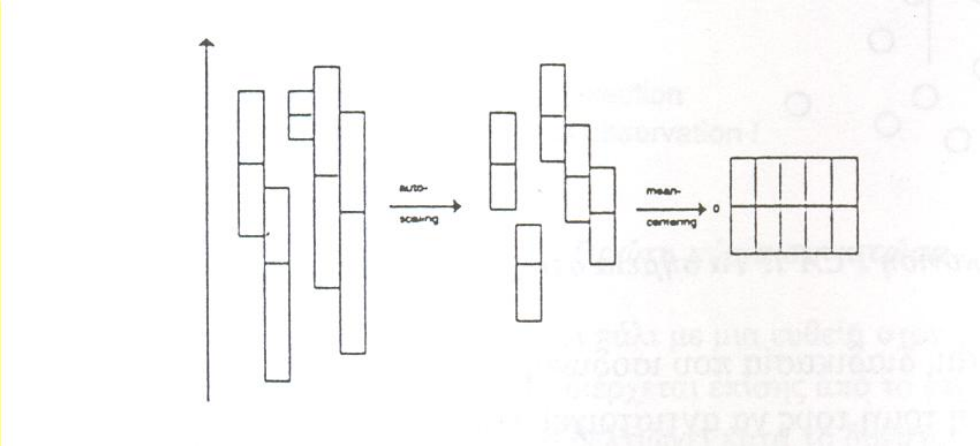
Γεωμετρικά μπορούν να απεικονιστούν οι παρατηρήσεις ως σημεία σε ένα πολυδιάστατο χώρο, όπου οι μεταβλητές προσδιορίζουν τους άξονες. Το μήκος των αξόνων προσδιορίζεται από την κλίμακα των μεταβλητών

Ανάλυση κυρίων συνιστωσών (PCA)

- Τα αποτελέσματα της πολυμεταβλητής ανάλυσης δεδομένων επηρεάζονται από την μέθοδο που εφαρμόζεται για την κανονικοποίηση της κλίμακας των μεταβλητών.
- Συνήθως τίθεται το μήκος κάθε άξονα ίσο με 1. (Scaling to unit variance).
- 

  Όλες οι μεταβλητές έχουν το ίδιο μήκος και θεωρούνται κατ' αυτόν τον τρόπο ίσης σημασίας.

Ανάλυση κυρίων συνιστωσών (PCA)- Scaling



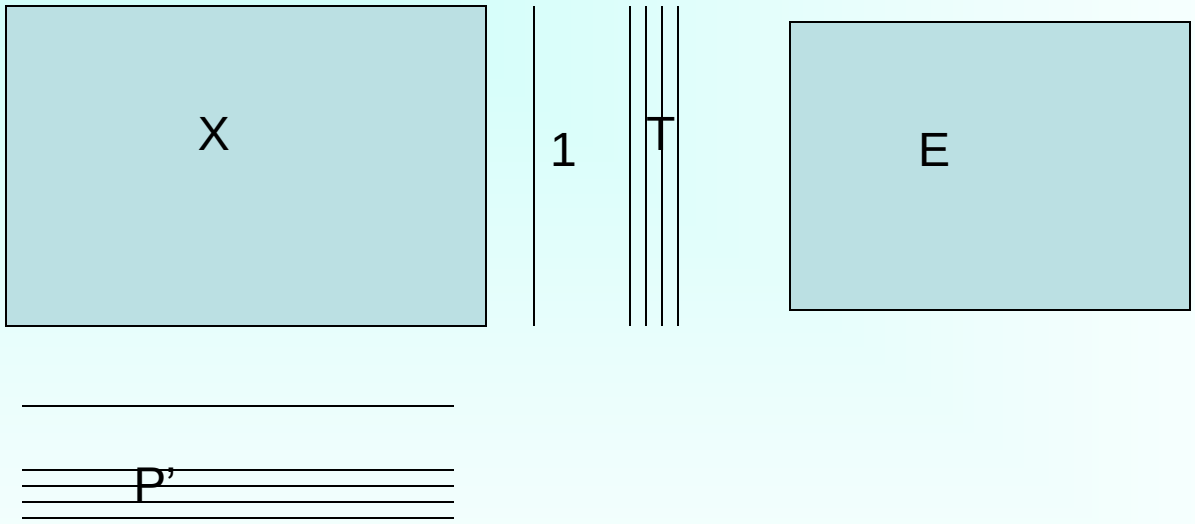
Ανάλυση κυρίων συνιστωσών (PCA)

- Στην Ανάλυση Κυρίων Συνιστωσών ο αρχικός πίνακας Χ περιγράφεται από το γινόμενο ενός πίνακα μικρότερων διαστάσεωνΤΡ’ συν τον πίνακα των υπολοίπων Ε

$$X = NK$$

$$X = 1 * \bar{X} + TP' + E$$

Τ: ο πίνακας των νέων συντεταγμένων (scores)
Ρ΄ : ο πίνακας των φορτίων-διευθύνσεων των νέων μεταβλητών ως προς τις αρχικές (loadings)
$\bar{X}$ ο μέσος όρος των μεταβλητών


X
1
T
E
P’

Ανάλυση κυρίων συνιστωσών (PCA)

- Η τυπική απόκλιση των υπολοίπων residual standard deviation (RSD) υπολογίζεται τόσο για τα αντικείμενα όσο και για τις μεταβλητες

- Η RSD για τα αντικείμενα (σειρές στον πίνακα E) δείχνει την απόσταση των αντικειμένων από το μοντέλο (DModX) .

- Η RSD για τις μεταβλητές αντιστοιχεί στην σημασία της μεταβλητής στο μοντέλο

Ανάλυση κυρίων συνιστωσών (PCA)

- Μια συνιστώσα θεωρείται σημαντική όταν η κανονικοποιημένη ιδιοτιμή της είναι μεγαλύτερη του 2.

prediction error sum of squares (PRESS) αφορά στα αντικείμενα που έχουν βγει από την ανάλυση

$PRESS = \Sigma(Y_{υπολ} - Y_{πειρ})^2$

SS το άθροισμα τετραγώνων των υπολοίπων της προηγούμενης συνιστώσας

- Μια συνιστώσα θεωρείται σημαντική όταν $PRESS/SS < 1$

Ανάλυση κυρίων συνιστωσών (PCA)- Στατιστικά στοιχεία

- **R^2 X:** Κλάσμα του αθροίσματος των τετραγώνων (SS) όλων των μεταβλητών X που ερμηνεύεται από τη συγκεκριμένη κύρια συνιστώσα

- **R^2 Xadj :** Κλάσμα της διακύμανσης όλων των μεταβλητών X που ερμηνεύεται από τη συγκεκριμένη κύρια συνιστώσα - κλάσμα προσαρμοσμένο ως προς τους βαθμούς ελευθερίας

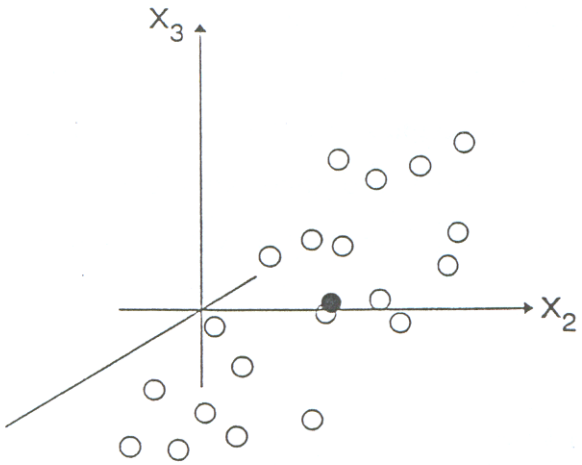
- **R^2 X(cum):** Συνολικό SS όλων των μεταβλητών X που ερμηνεύεται από όλες τις κύριες συνιστώσες

- **R^2Xadj(cum):** Η συνολική διακύμανση όλων των μεταβλητών X που ερμηνεύεται από όλες τις κύριες συνιστώσες- Συνολικό SS προσαρμοσμένο ως προς τους βαθμούς ελευθερίας

Ανάλυση κυρίων συνιστωσών (PCA)- Στατιστικά στοιχεία

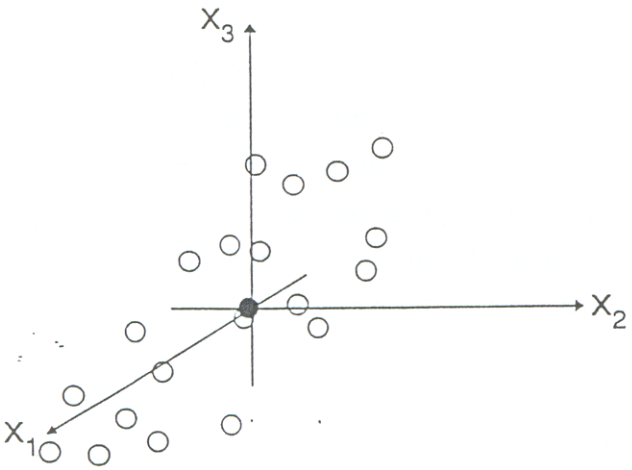
- **Q^2:** Το κλάσμα της συνολικής διακύμανσης των X που μπορεί να προβλεφτεί σύμφωνα με τη διαδικασία cross-validation.
- $Q^2 = (1.0 - PRESS/SS)$
- **Q^2_v :** Το κλάσμα της διακύμανσης μιας μεταβλητής $_XK$ που μπορεί να προβλεφτεί σύμφωνα με τη διαδικασία cross-validation.
- $Q^2 = (1.0 - PRESS/SS)_K$
- **Q^2(cum):** Το συνολικό Q^2 για όλες τις συνιστώσες.
- **Q^2(cum)**$= (1.0 - \prod PRESS/SS)_a$
- [a = 1, ...A]

Ανάλυση κυρίων συνιστωσών (PCA)



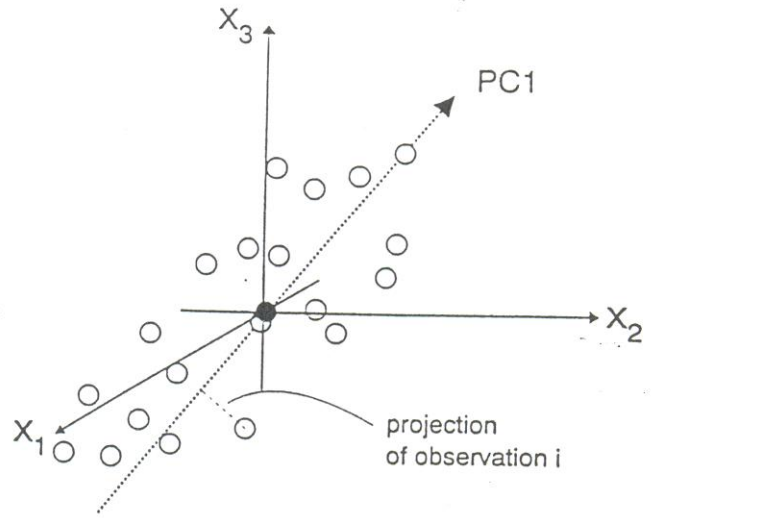
Σχήμα 3. Γραφική απεικόνιση PCA 1. Τα σημεία στο χώρο Κ διαστάσεων

Ανάλυση κυρίων συνιστωσών (PCA)



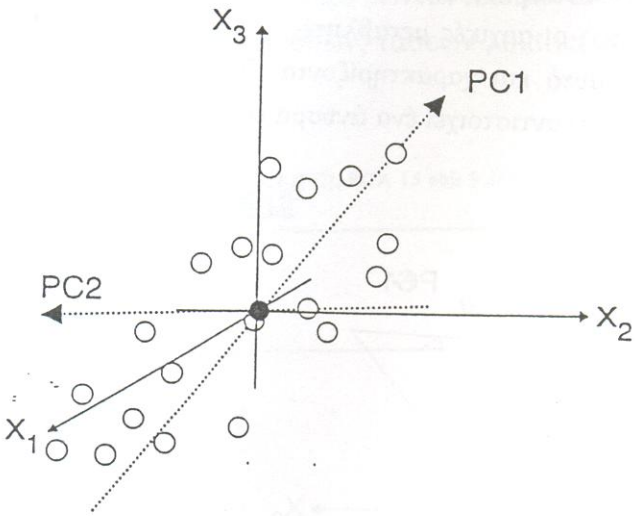
Σχήμα 4. Γραφική απεικόνιση PCA 2. Μετακίνηση των συντεταγμένων

Ανάλυση κυρίων συνιστωσών (PCA)



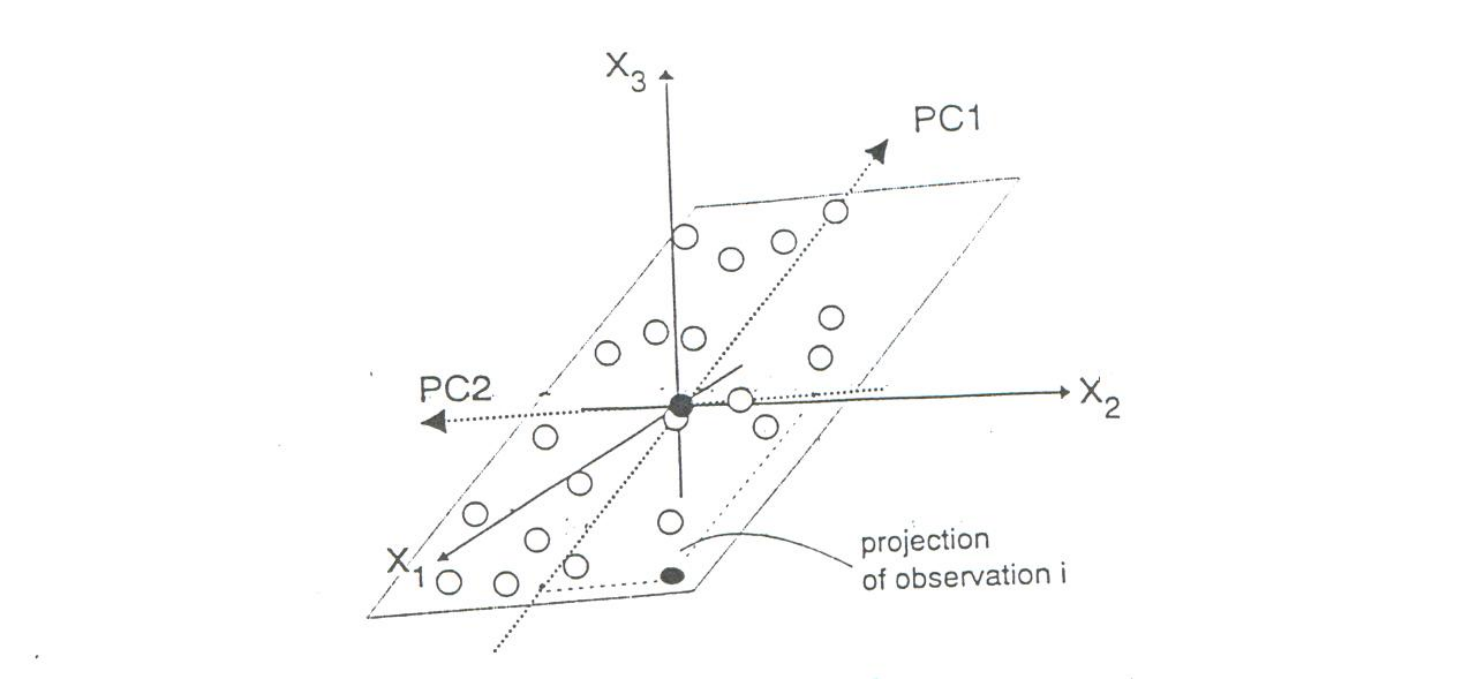
Σχήμα 5. Γραφική απεικόνιση PCA 3. Πρώτη κύρια συνιστώσα

Ανάλυση κυρίων συνιστωσών (PCA)

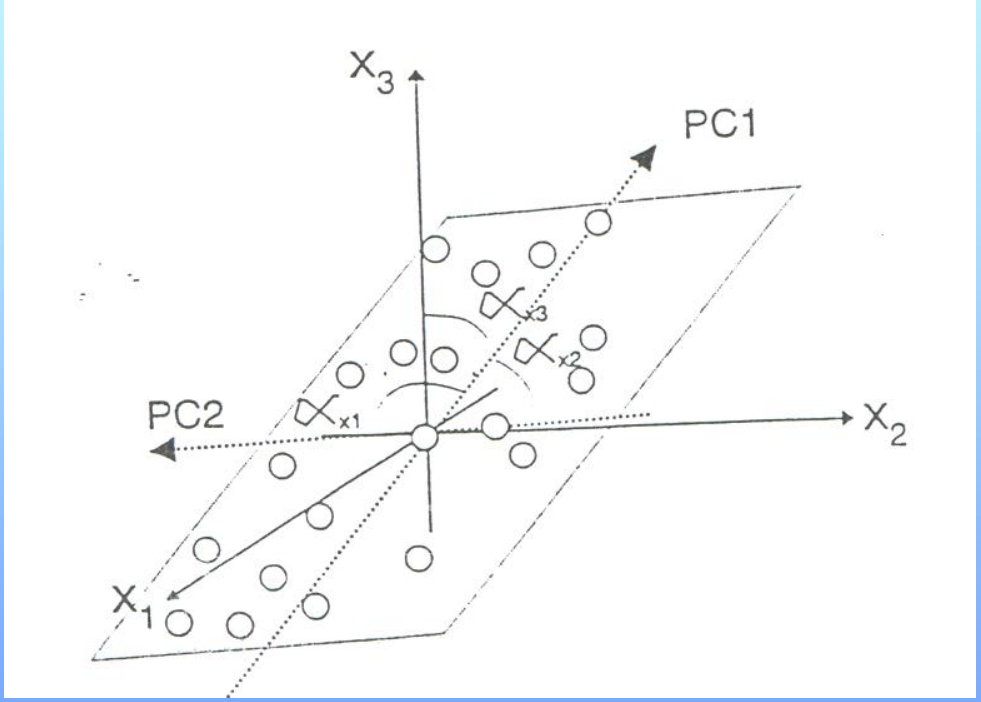


Σχήμα 6. Γραφική απεικόνιση PCA 4. Δεύτερη κύρια συνιστώσα κάθετη στην πρώτη

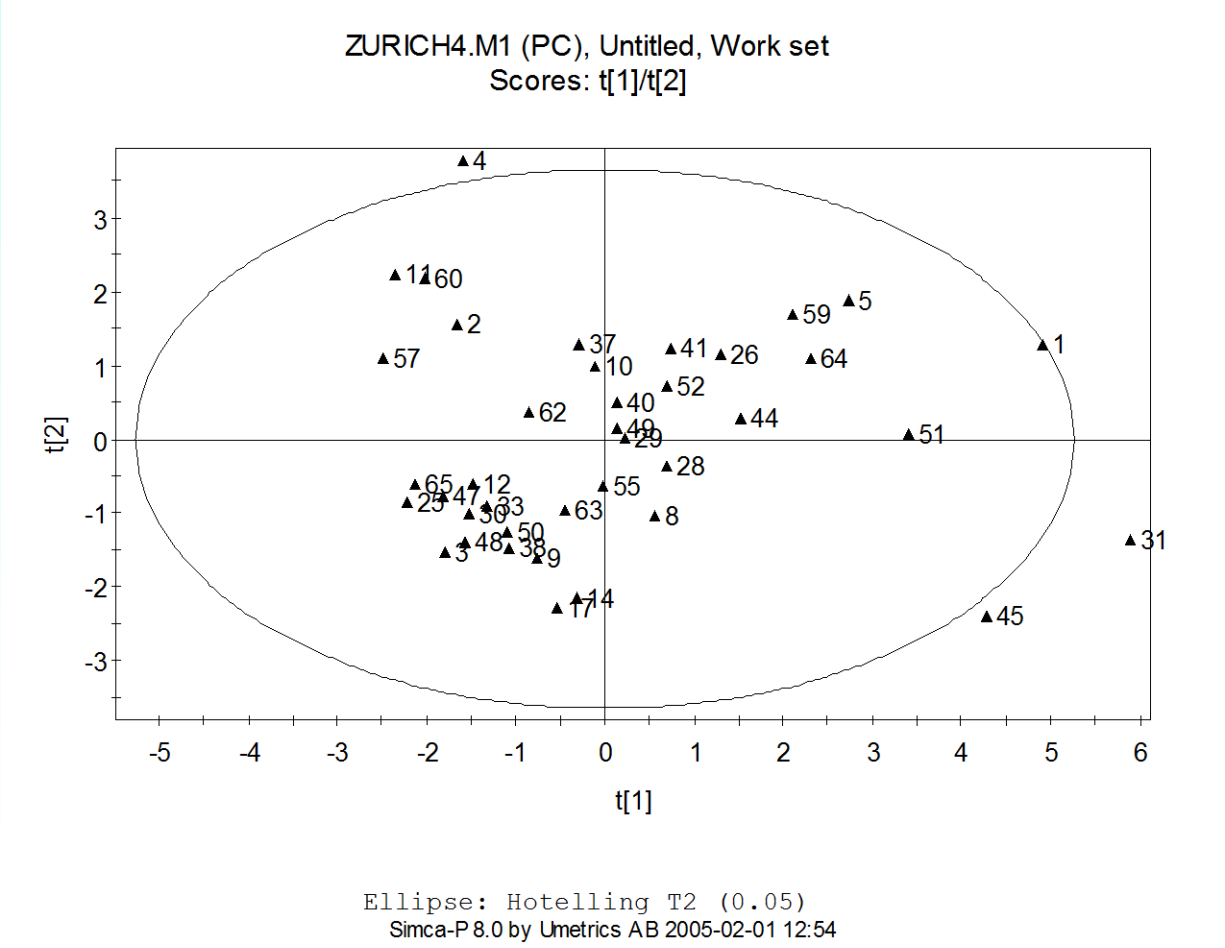
Ανάλυση κυρίων συνιστωσών (PCA)



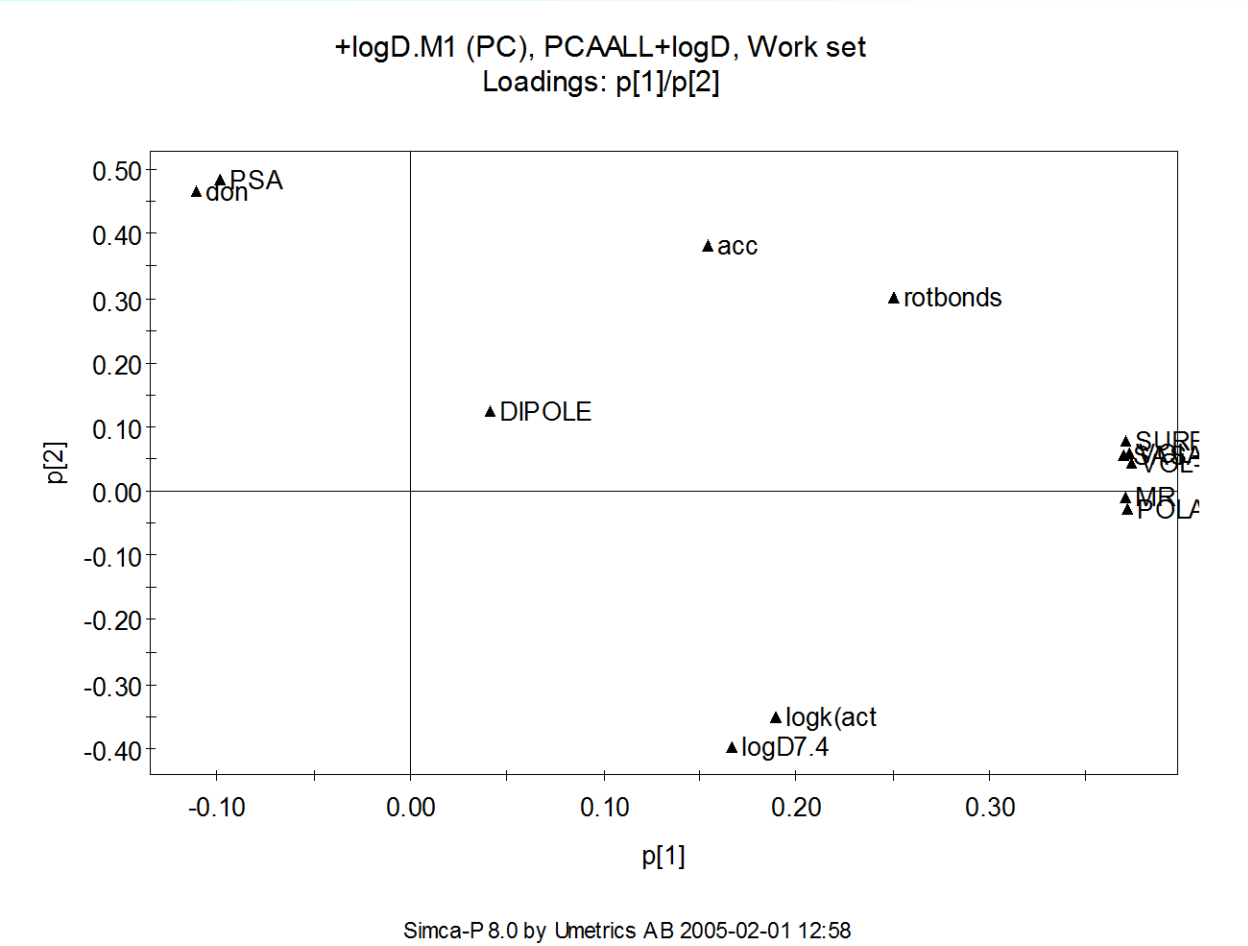
Ανάλυση κυρίων συνιστωσών (PCA)



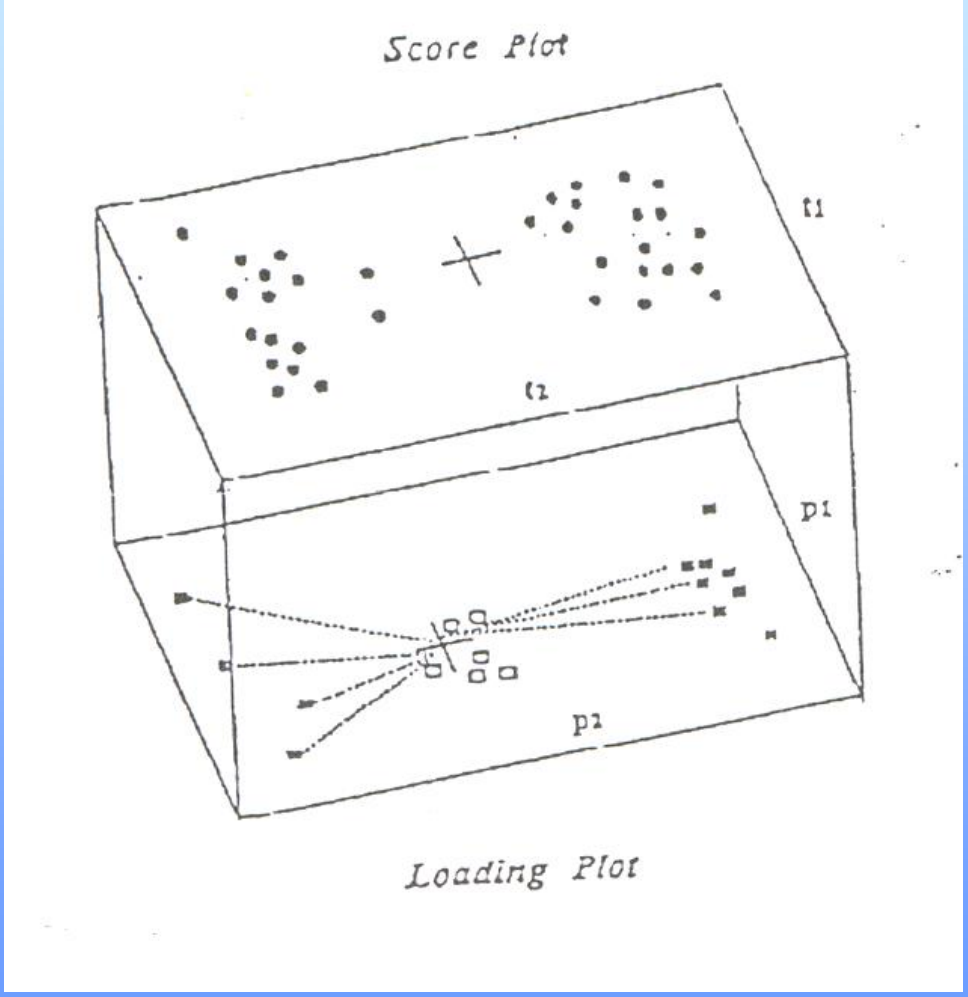
Ανάλυση κυρίων συνιστωσών



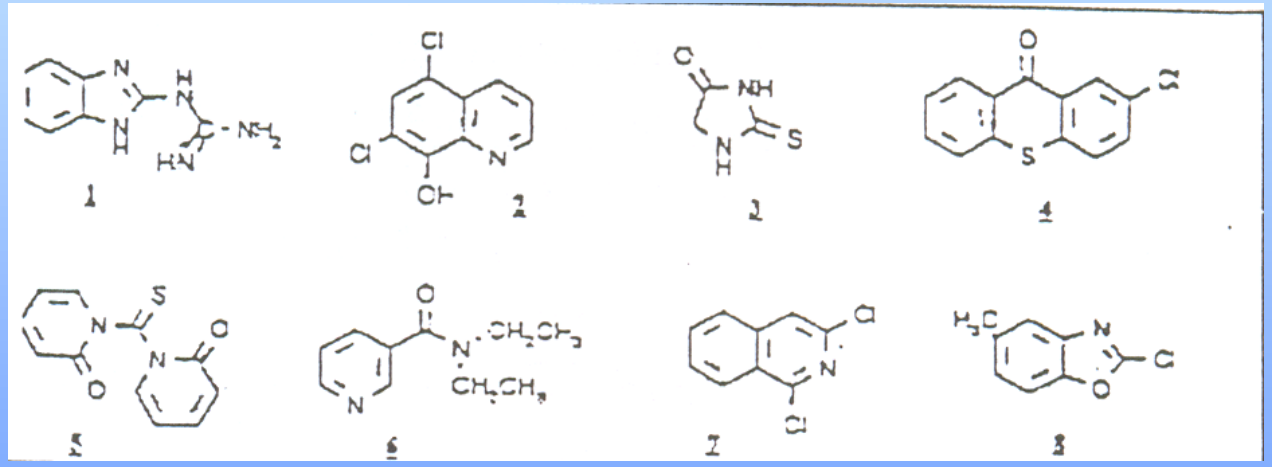
Ανάλυση κυρίων συνιστωσών



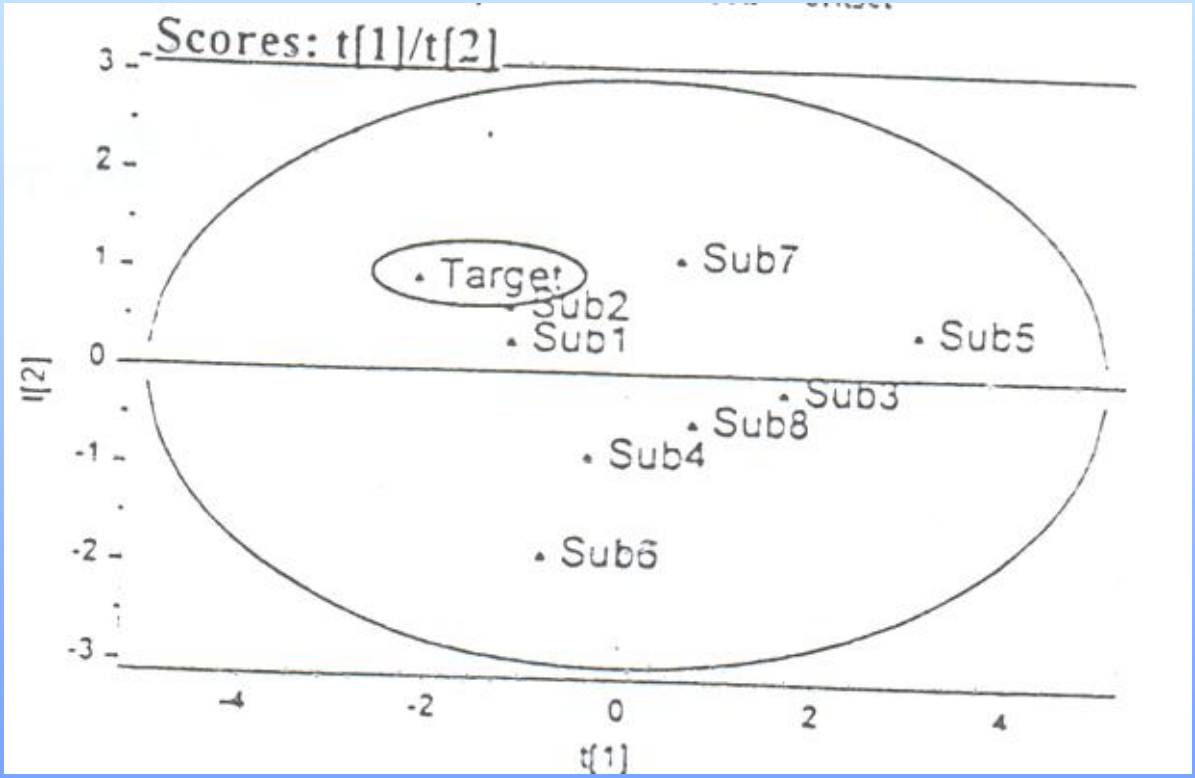
Ανάλυση κυρίων συνιστωσών



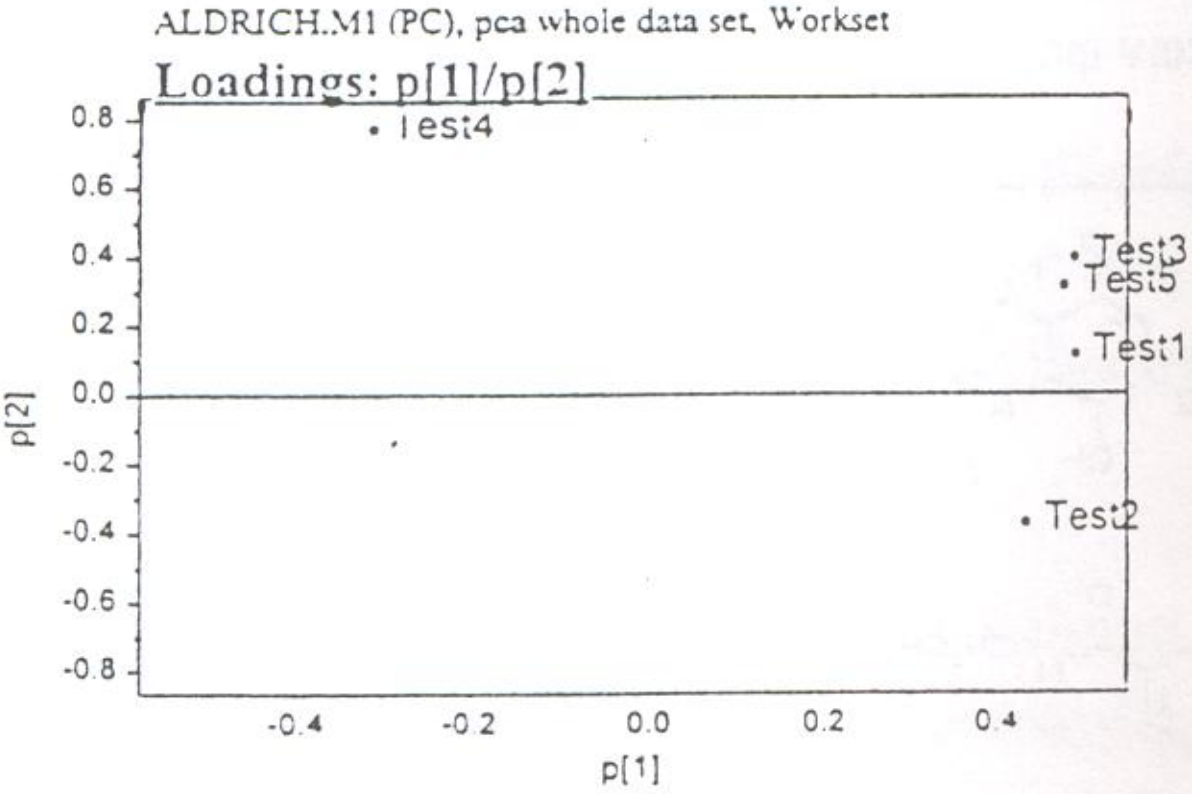
Ανάλυση κυρίων συνιστωσών



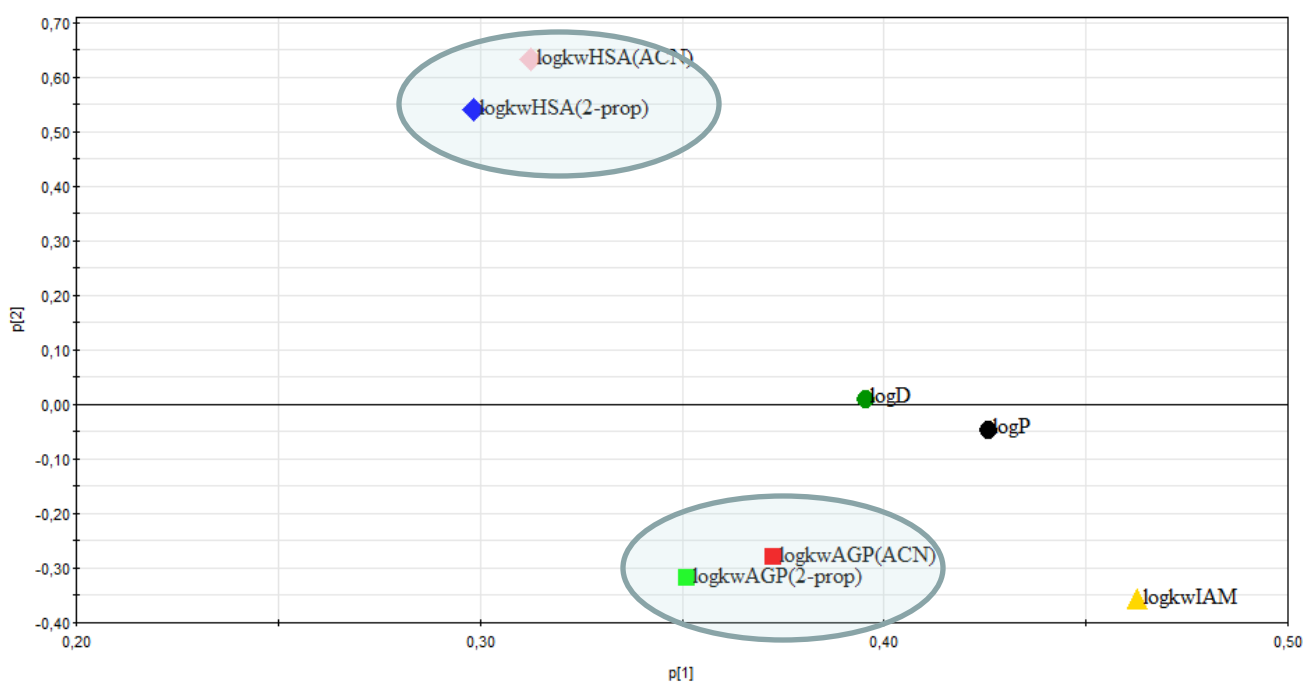
Ανάλυση κυρίων συνιστωσών



Ανάλυση κυρίων συνιστωσών



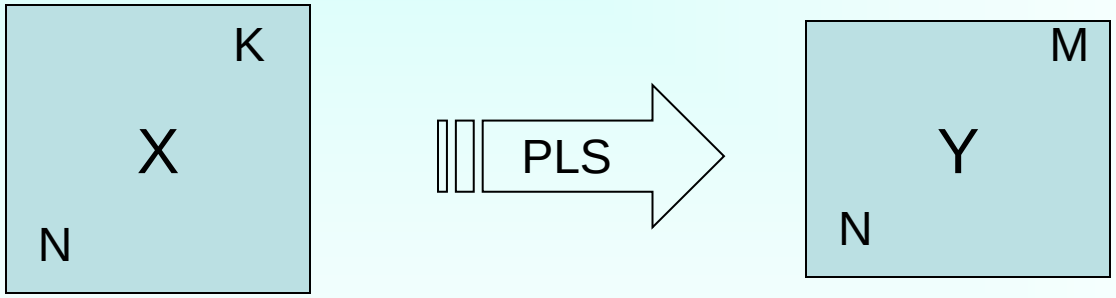
Σύγκριση χρωματογραφικών στηλών



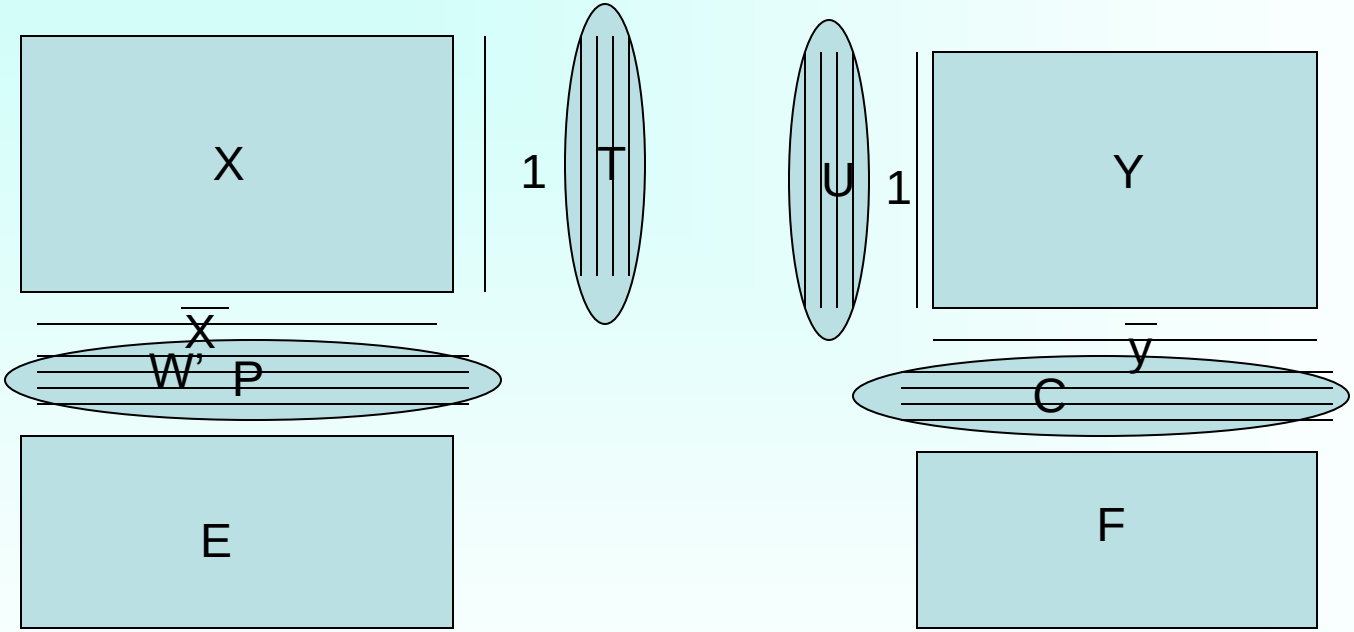
Partial Least Squares Projections to Latent Structures (PLS)
Προβολές Μερικών Ελαχίστων Τετραγώνων σε Λανθάνουσες Δομές

K: Περιγραφικές μεταβλητές

M:Αποκρίσεις



N:αντικείμενα, παρατηρήσεις


X
1
T
U
1
Y
X
W'
P
y
C
E
F

Partial Least Squares Projections to Latent Structures (PLS)

- **X=1*$\bar{x}$+TP'+E**
- **Y=1*$\bar{y}$+UC'+F**
- **U=T+H (εσωτερική σχέση)**

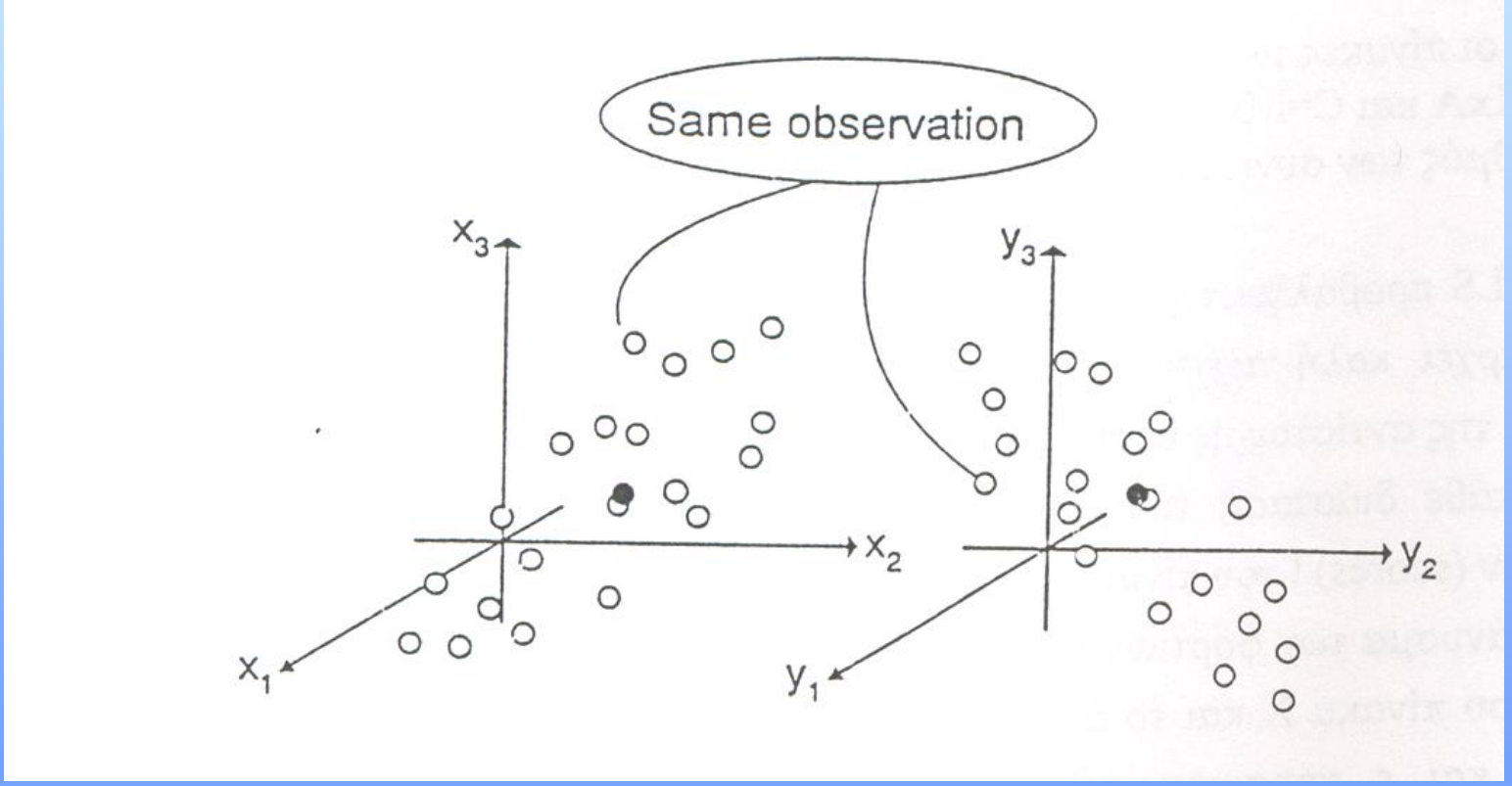
Κατά τη διαδικασία **PLS** τα δεδομένα προσαρμόζονται ταυτόχρονα σε δύο μοντέλα PCA

W' : Πίνακας βαρών που εκφράζουν τη συσχέσρτιση μεταξύ X και **U (Y)**

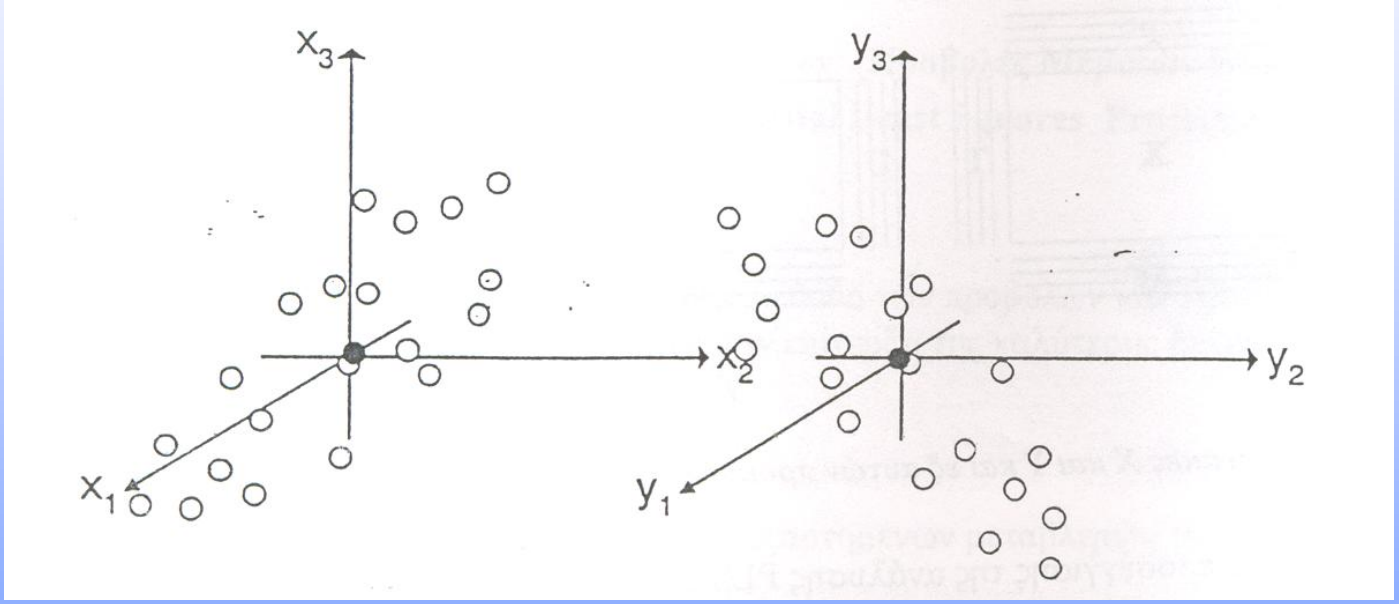
Partial Least Squares Projections to Latent Structures (PLS)

T	Πίνακας των scores που συνοψίζει τις μεταβλητές X
U	Πίνακας των scores που συνοψίζει τις μεταβλητές Y
C	Πίνακας των φορτίων που εκφράζει τη συσχέτιση μεταξύ Y και T (X)
E,F, H, G	Πίνακας υπολοίπων

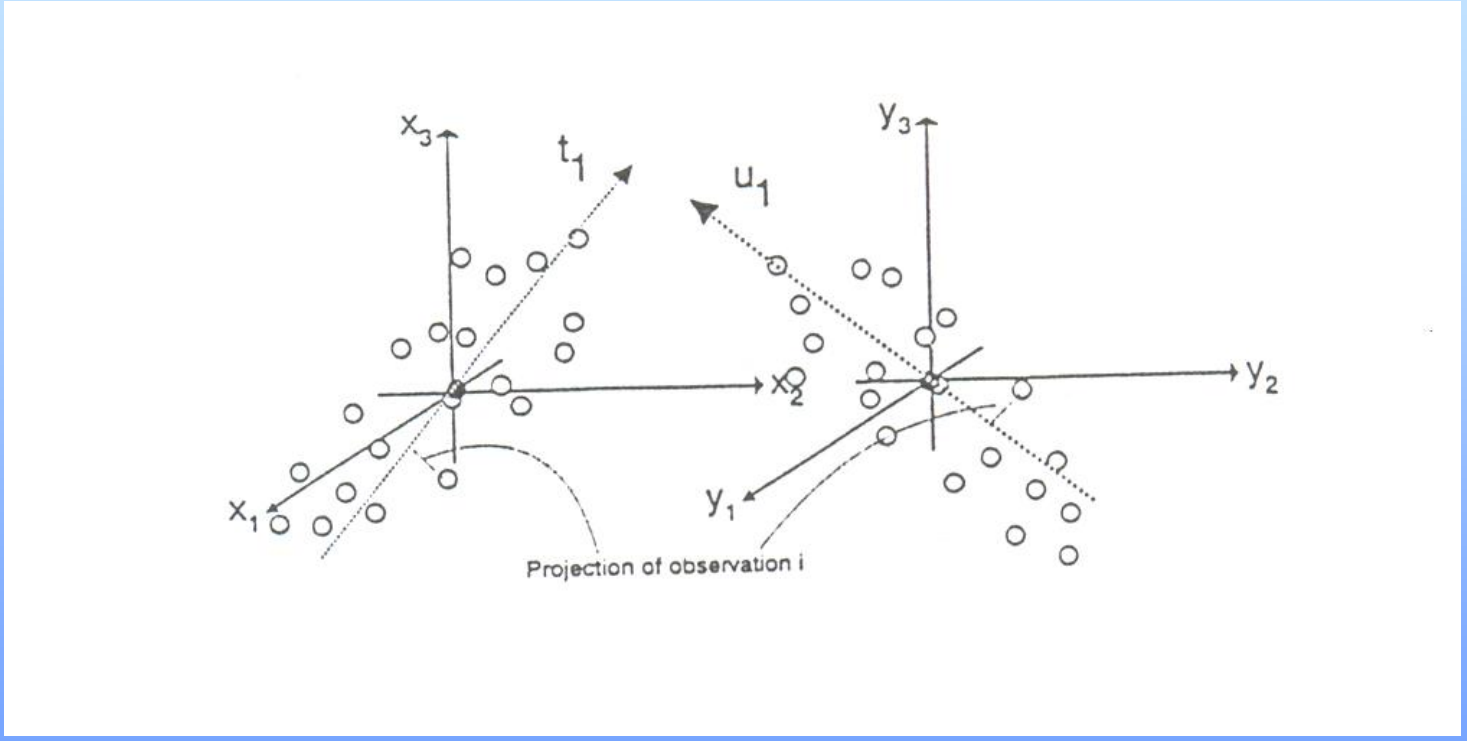
Partial Least Squares Projections to Latent Structures (PLS)



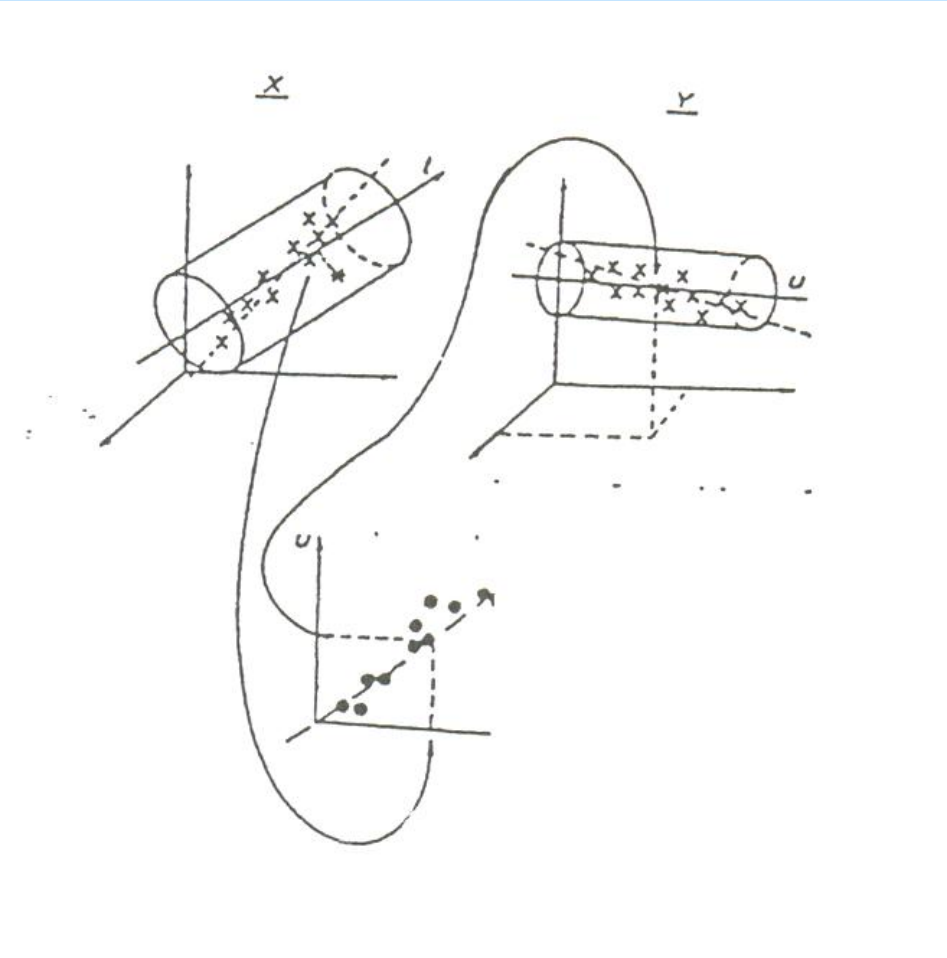
Partial Least Squares Projections to Latent Structures (PLS)



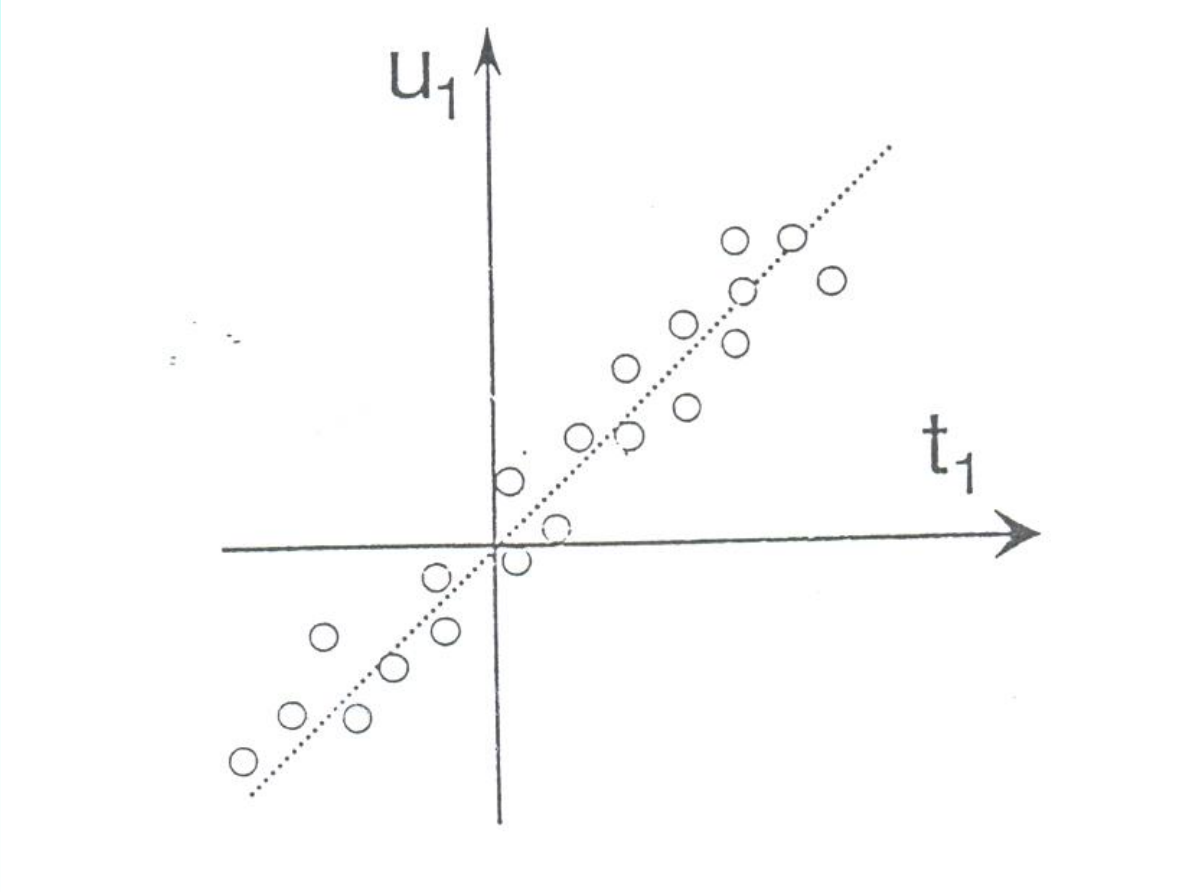
Partial Least Squares Projections to Latent Structures (PLS)



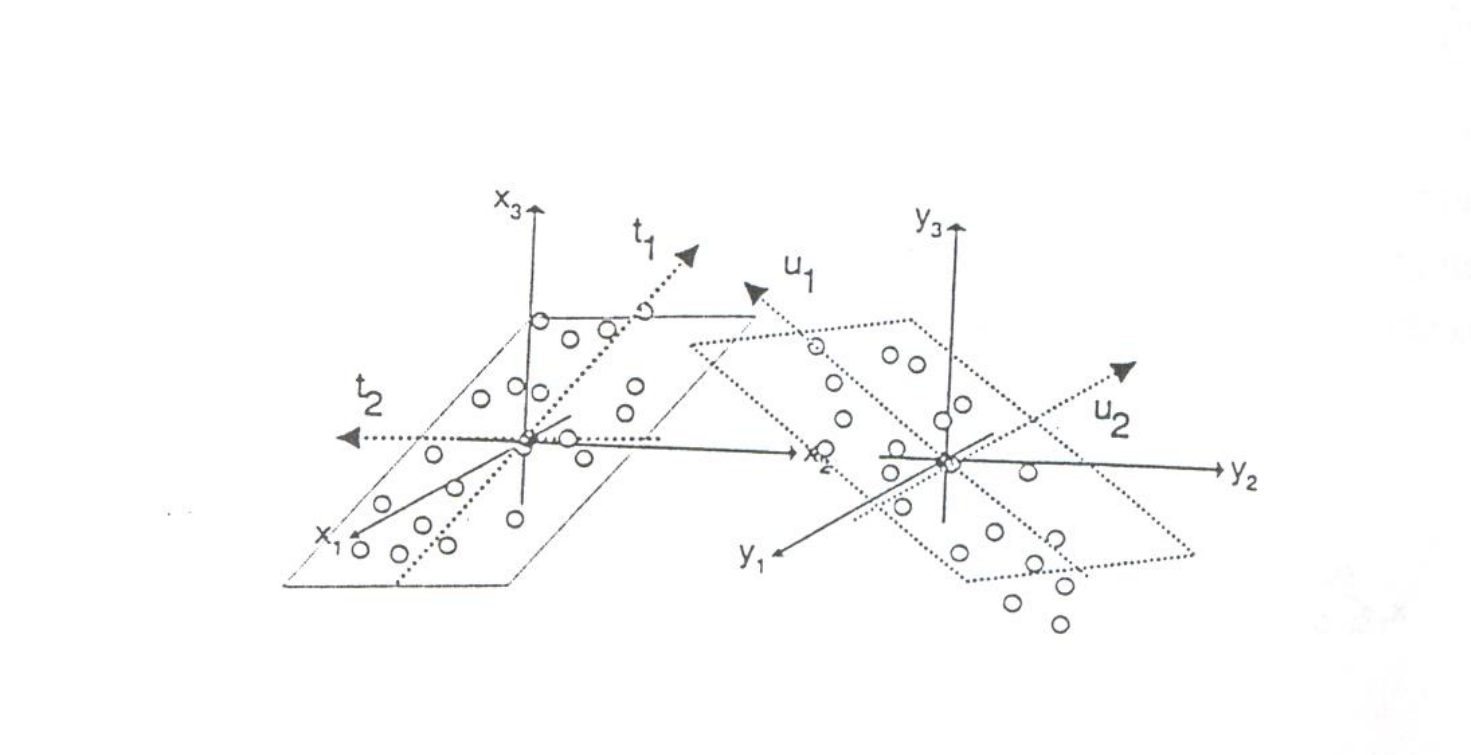
Partial Least Squares Projections to Latent Structures (PLS)



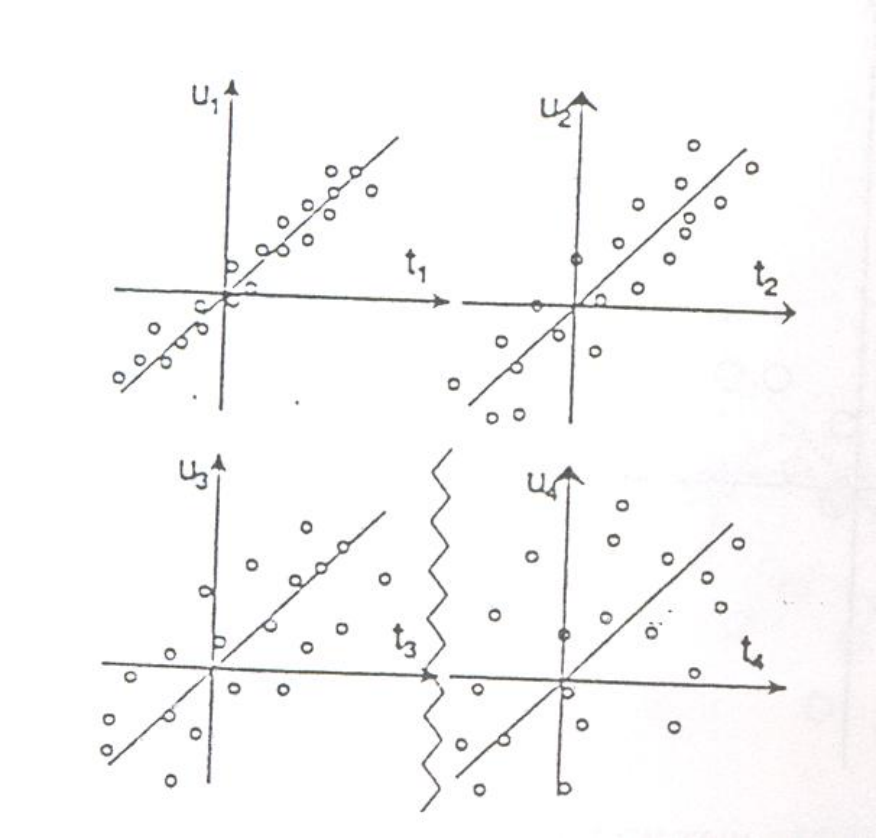
Partial Least Squares Projections to Latent Structures (PLS)



Partial Least Squares Projections to Latent Structures (PLS)



Partial Least Squares Projections to Latent Structures (PLS)



Partial Least Squares Projections to Latent Structures (PLS)

- **Στον αλγόριθμο PLS περιλαμβάνονται επι πλέον τα βάρη W.**
- **Τα βάρη W εκφράζουν τη συσχέτιση μεταξύ U(Y) και X (πρώτη διάσταση, και ακολούθως των υπολοίπων των μεταβλητών X) και χρησιμοποιούνται για τον υπολογισμό των T**
- **Τα βάρη W επιλέγονται έτσι ώστε να μεγιστοποιείται η συσχέτιση μεταξύ T και U**
- **Μεταβλητές X με μεγάλες (θετικές ή αρνητικές) τιμές W συσχετίζονται ισχυρά με τις συντεταγμένες U (μεταβλητές Y)**

W* είναι τα βάρη που συσχετίζουν τις αρχικές μεταβλητές X (όχι τα υπόλοιπα) ώστε να προκύψουν οι συντεταγμένες t . Ια την πρώτη διάσταση W= W*

Partial Least Squares Projections to Latent Structures (PLS)

- **Συντελεστές παλινδρόμησης**
  (Regression Coefficients)

B=w*C’

Y=Xw*C’

Μεγάλες τιμές w* σχετίζονται με τη σπουδαιότητα των μεταβλητών X στο μοντέλο

Partial Least Squares Projections to Latent Structures (PLS)

- **Σπουδαιότητα μεταβλητών- Επίδραση μεταβλητών**
  Variable Importance – Variable Influence

$$VIP_k = \sum_a (VIN)_k^2)$$

$$(VIN)_{ak}^2 = (w_{ak})^2/SS$$

Μια μεταβλητή είναι ιδιαίτερα σημαντική εάν $VIP_k > 1$

Συνήθως το όριο για τη σπουδαιότητα της μεταβλητής τίθεται ~ 0.7-0.8

Partial Least Squares Projections to Latent Structures (PLS)

Σημαντικές απεικονίσεις:

- **Score Plots**

t1 vs. t2,...αποτελούν παράθυρο στο χώρο των X, δείχνουν τη θέση των αντικειμένων στα επίπεδα ή υπερεπίπεδα των προβολών

Οι απεικονίσεις αυτές αποκαλύπτουν:

- ομάδες
- τάσεις
- Outliers
- ομοιότητες
- κλπ.

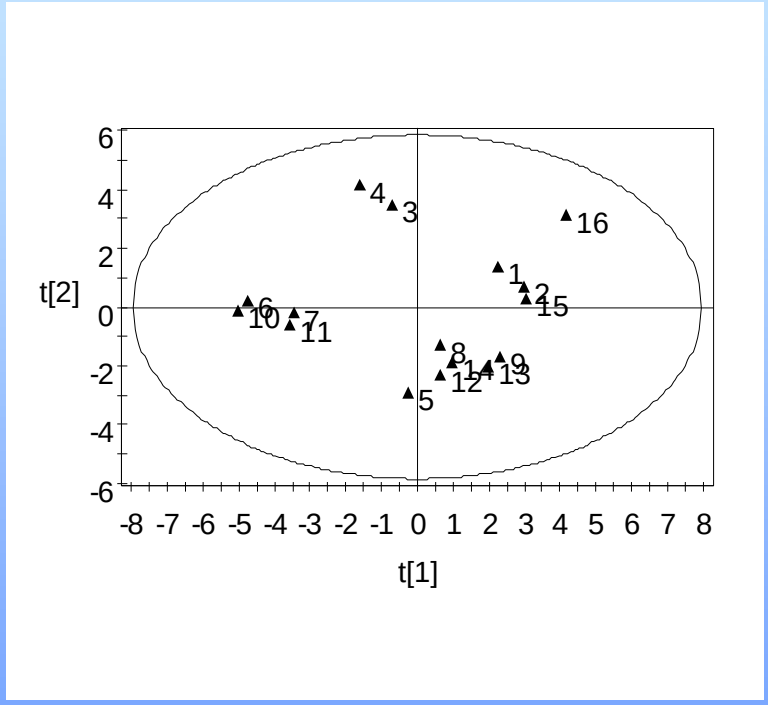
Partial Least Squares Projections to Latent Structures (PLS)

Σημαντικές απεικονίσεις:

- u_1 vs. u_2, ...αποτελούν παράθυρα στο χώρο των Y, δείχνουν τη θέση των παρατηρήσεων στο επίπεδο ή υπερεπίπεδο προβολής

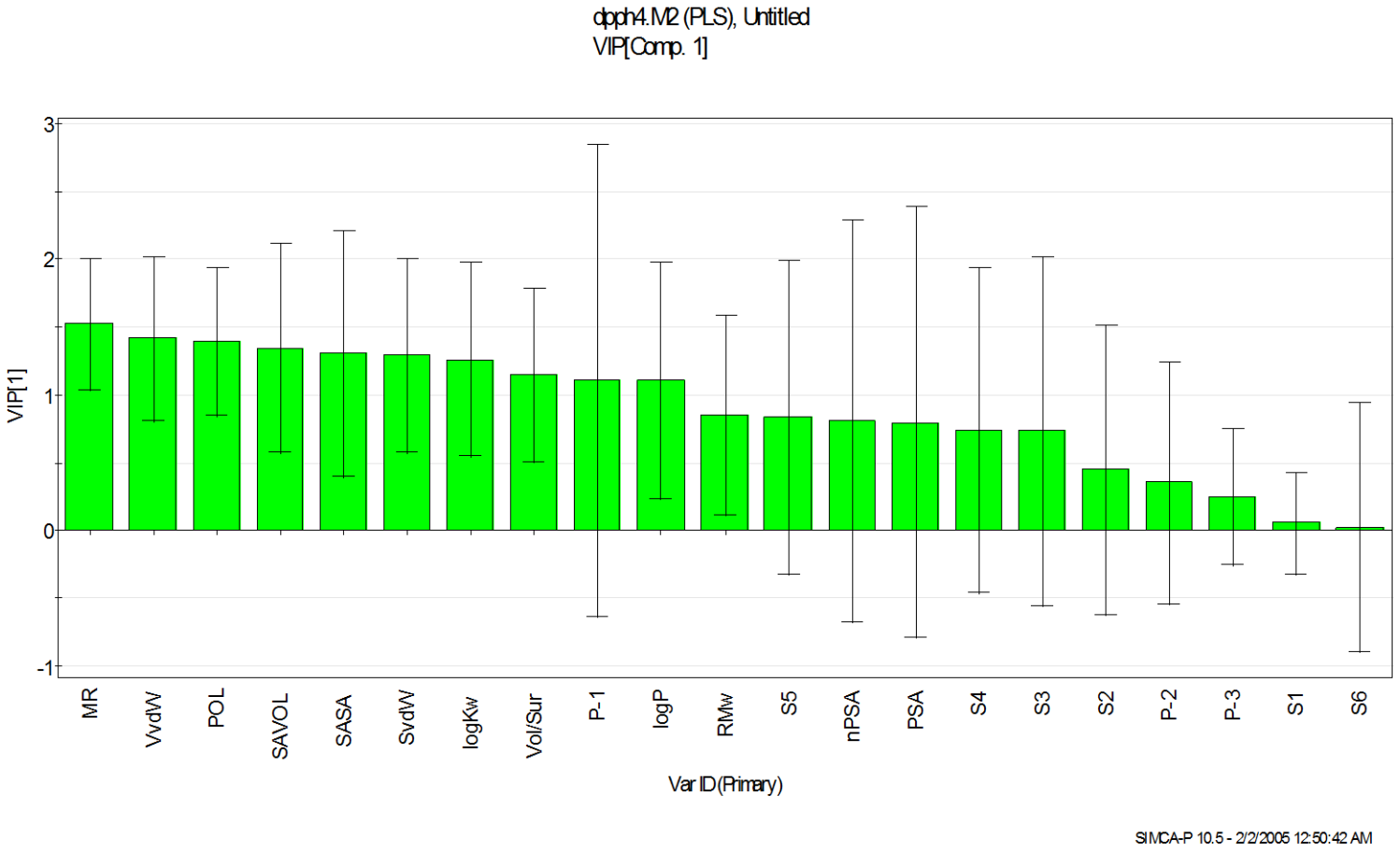
- u_1 vs. t_1, ...απεικονίζουν τις παρατηρήσεις όπως προβάλλονται στο χώρο X(T) και Y(U) και δείχνουν πόσο καλά συσχετίζεται ο χώρος Y με το χώρο X

Αλληλεπίδραση υποκατεστημένων κουμαρινών με DPPH

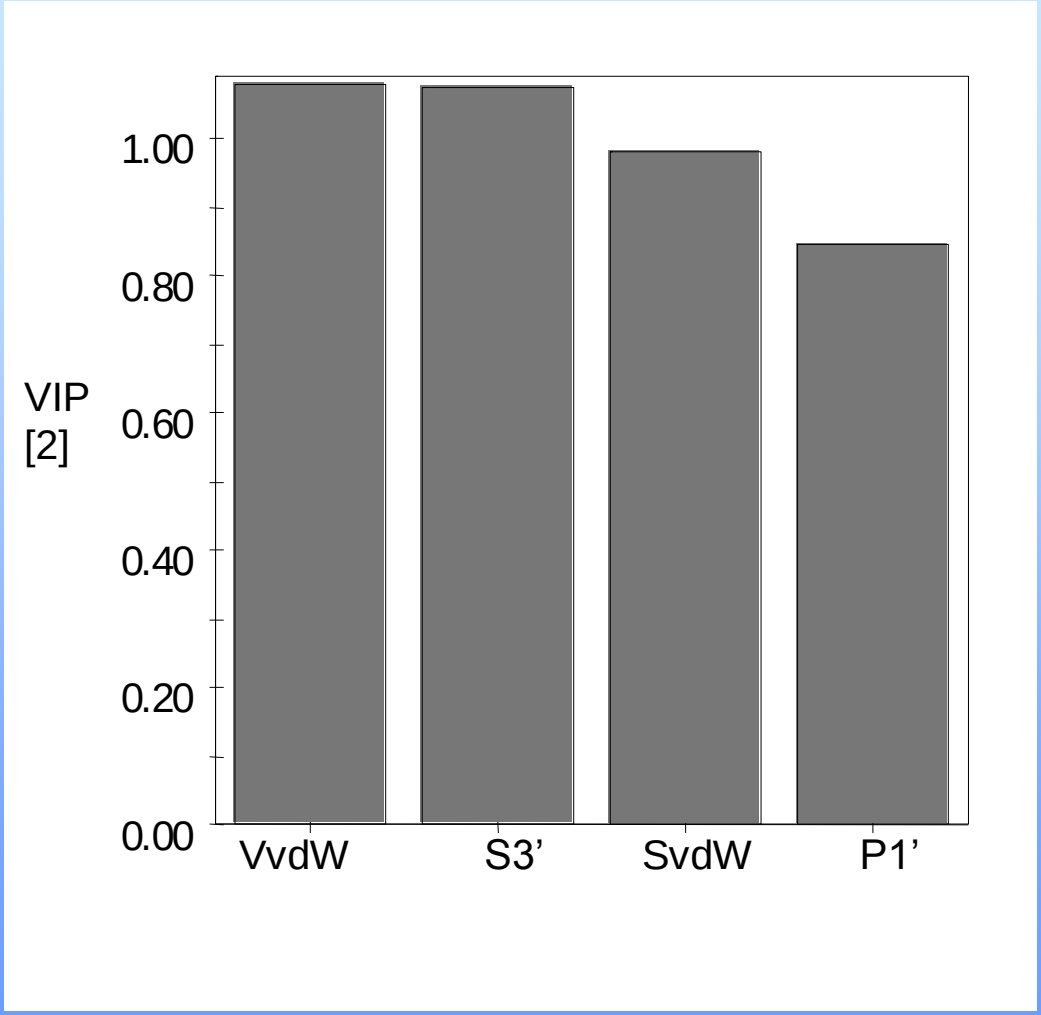


A=4 r^2=0.899 and q^2=0.5

Αλληλεπίδραση υποκατεστημένων κουμαρινών με DPPH

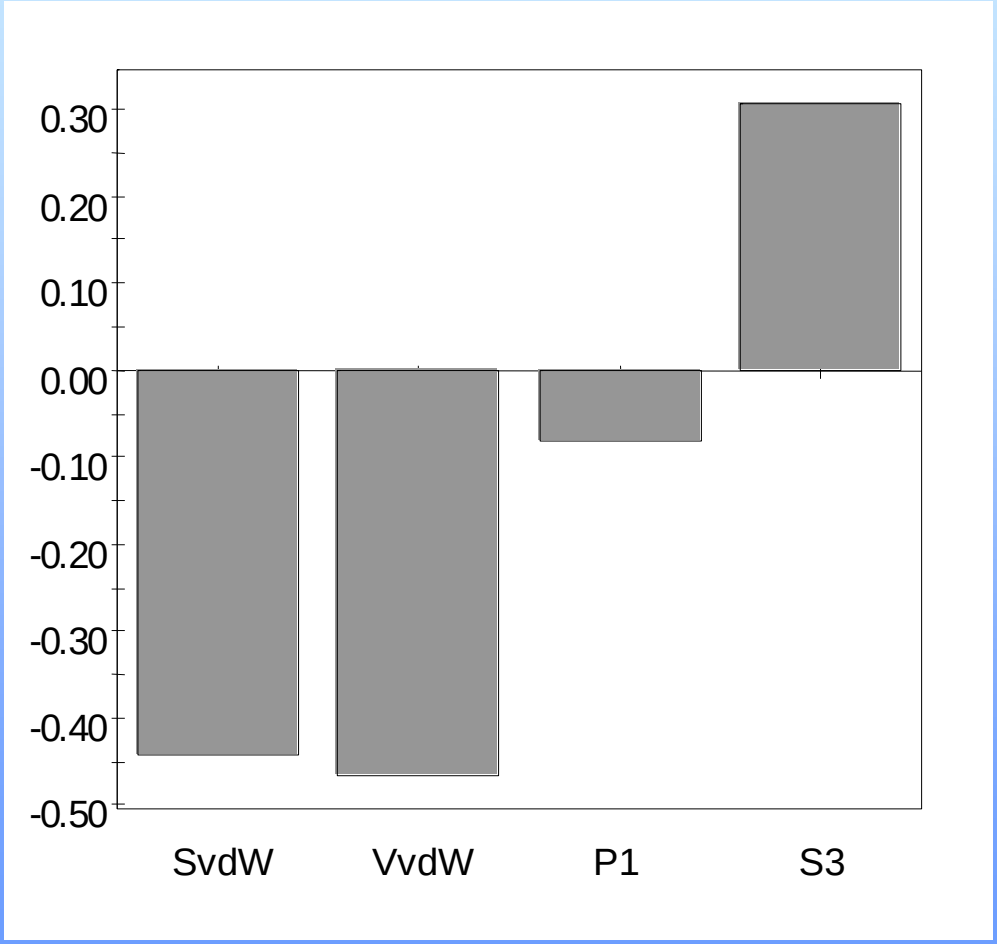


Αλληλεπίδραση υποκατεστημένων κουμαρινών με DPPH

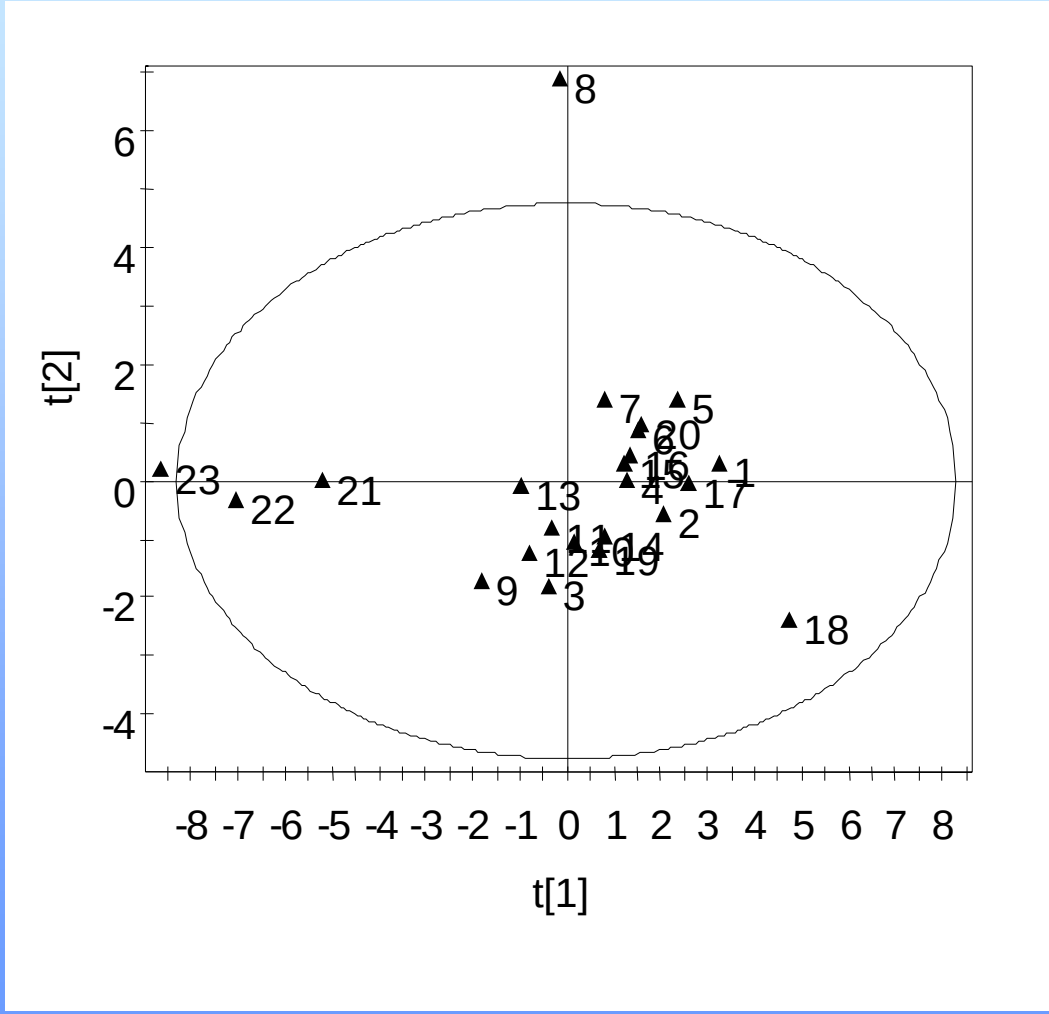


A=2, r^2=0.864 q^2= 0.728

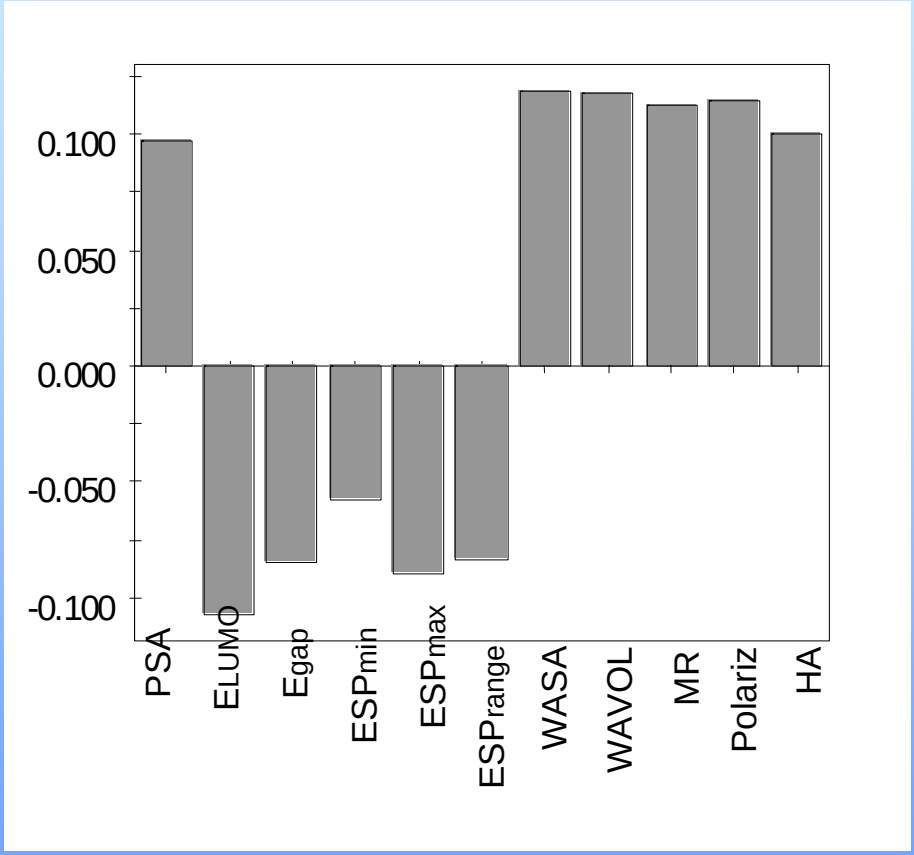
Αλληλεπίδραση υποκατεστημένων κουμαρινών με DPPH



Αναστολτική δράση στην αναγωγάση της αλδόζης παραγώγων του πυρρολυλ-οξικού οξέος



Ανασταλτική δράση στην αναγωγάση της αλδόζης



A=1: R2= 0.838, Q2=0.810, RMS=0.222.

A=1, R2=0.818, Q2=0.777, RMS=0.238

Ανασταλτική δράση στην αναγωγάση της αλδόζης

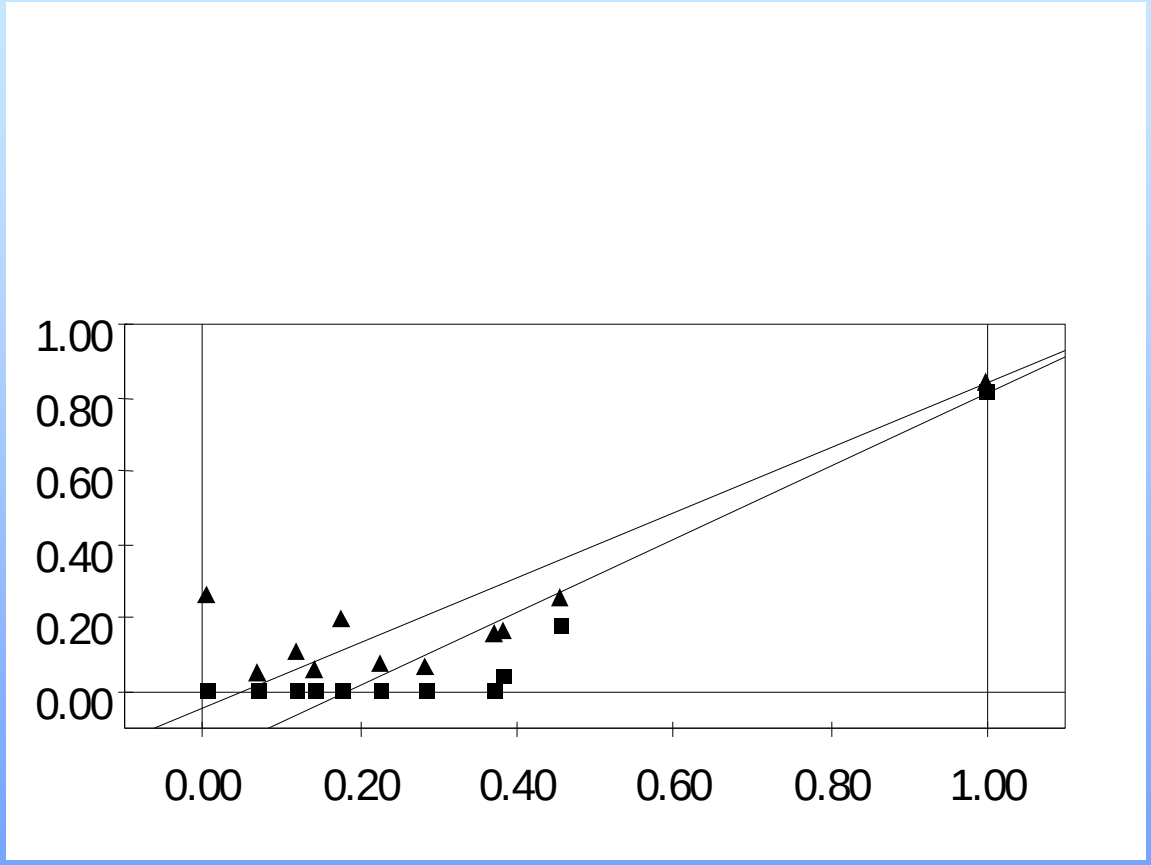
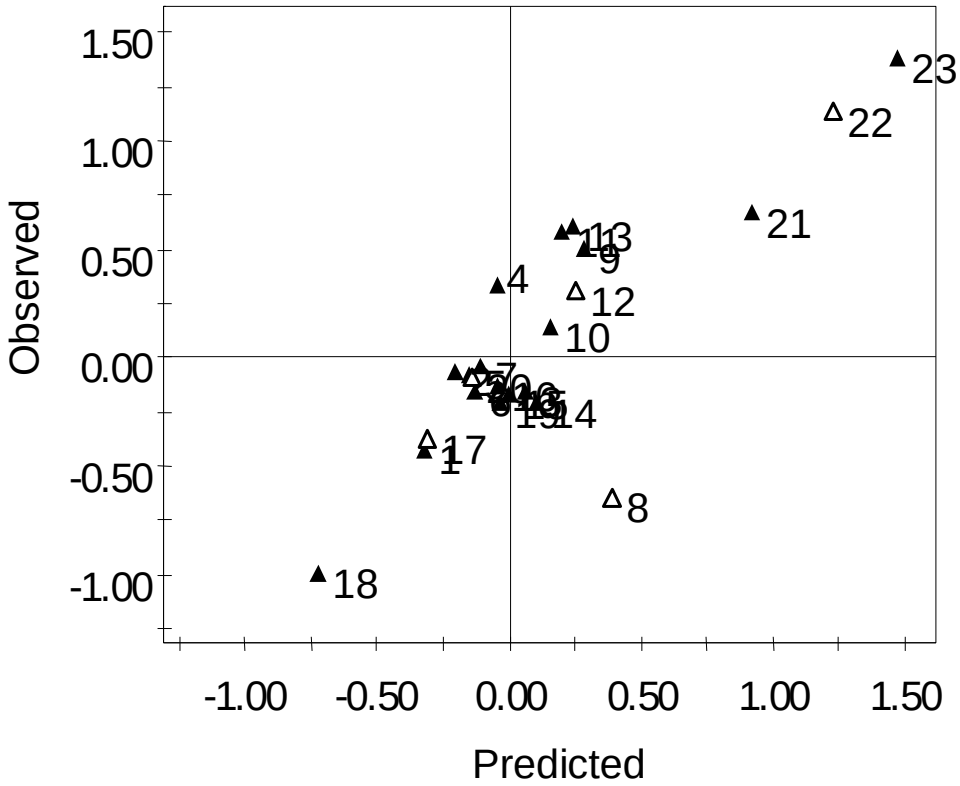


Fig. (5). Permutation test: R2: intercept = -0.04, Q2: intercept = -0.18.

Ανασταλτική δράση στην αναγωγάση της αλδόζης



PLS-DA

- **Μέθοδος ταξινόμησης supervised**
- **Τα αντικείμενα χωρίζονται σε σειρά εκμάθησης και σειρά ελέγχου.**
- **Ακολουθεί κατάταξη σε τάξεις ανάλογα με την ιδιότητα στη βάση της οποίας θα γίνει η κατάταξη και ακολουθεί ανάλυση PLS.**
- **Ενδιαφέρει το διάγραμμα των scores όπως και στην ανάλυση PCA.**
- **Επι πλέον δίνονται οι πιθανότητες κάθε αντικειμένου της σειράς εκμάθησης** και της σειράς ελέγχου να ανήκει στη μία ή την άλλη τάξη.