La Cluster Analysis

OBIETTIVO

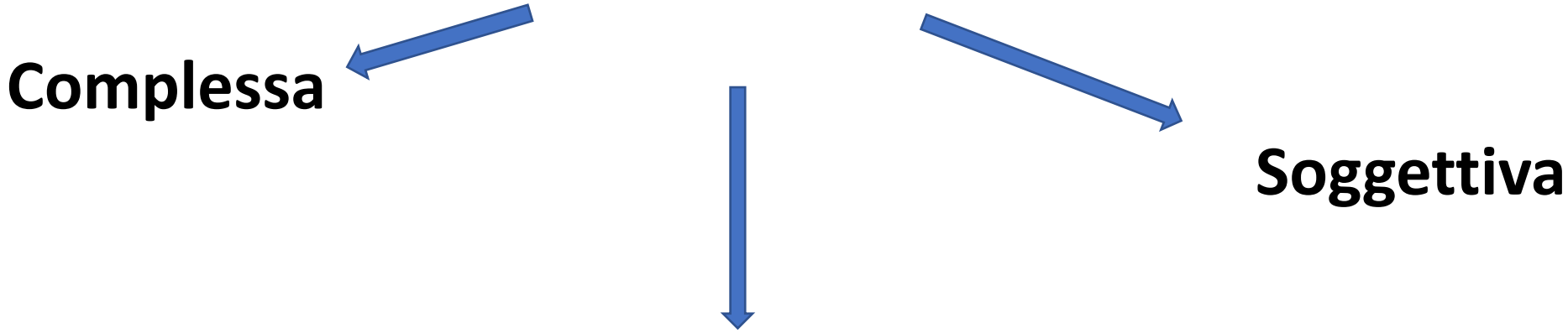
1) Individuazione clusters, clouds
   (oggetti, variabili simili)

2) Suggerire ‘categorie’ per studi di classificazione

La Cluster Analysis

PROBLEMATICHE

Similarità / distanza tra campioni



Complessa

Soggettiva

compresa tra 0 e 1

La Cluster Analysis

CLUSTER

Studio della somiglianza tra oggetti
(similarity)

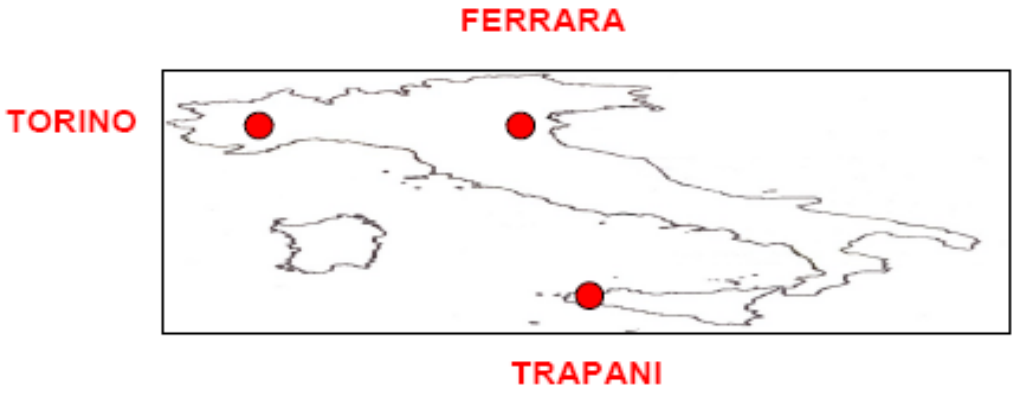
Attenzione alla scala!!! (le variabili possono avere natura e / o scala diversa
→ utilizzare una scalatura adatto
evitare che la vicinanza dipenda dalla scala delle variabili

Scelta variabili (descrittori)

La statistica forense: fondamenti teorici

LE VARIABILI

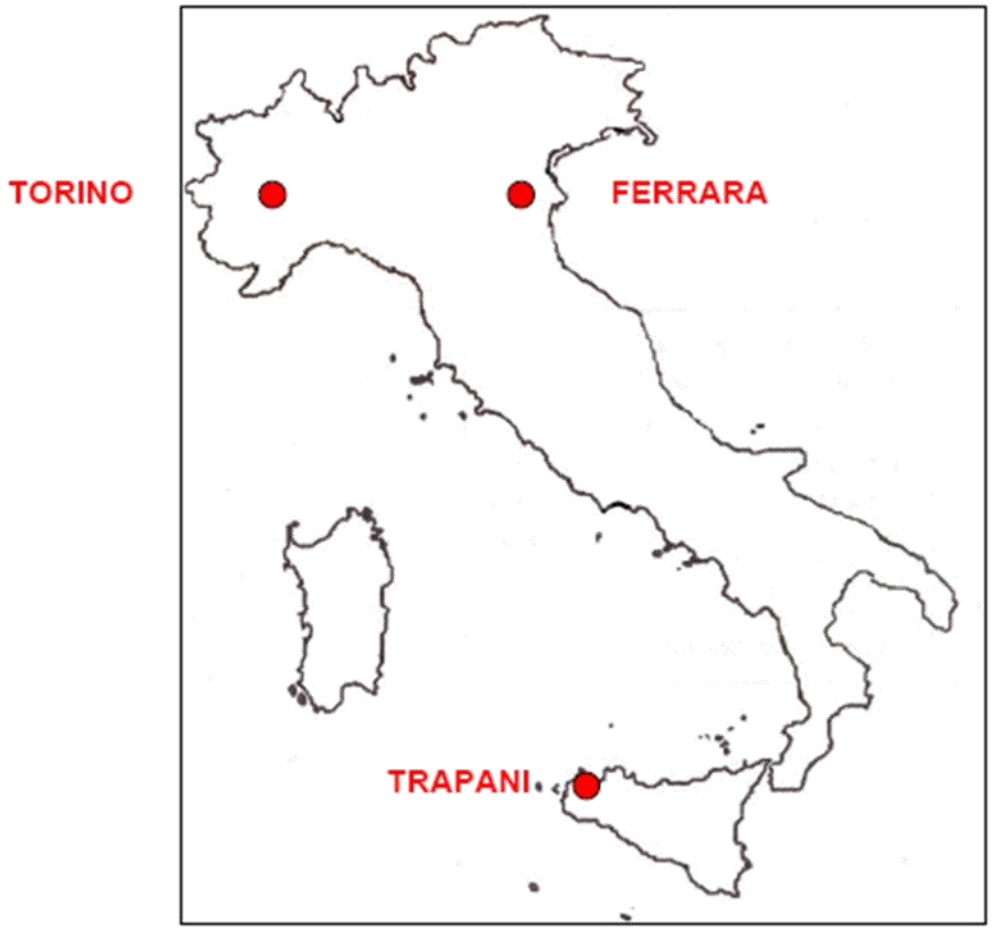
È importante fare attenzione alla scala dei valori (specie nel calcolo delle DISTANZE):



La statistica forense: fondamenti teorici

LE VARIABILI

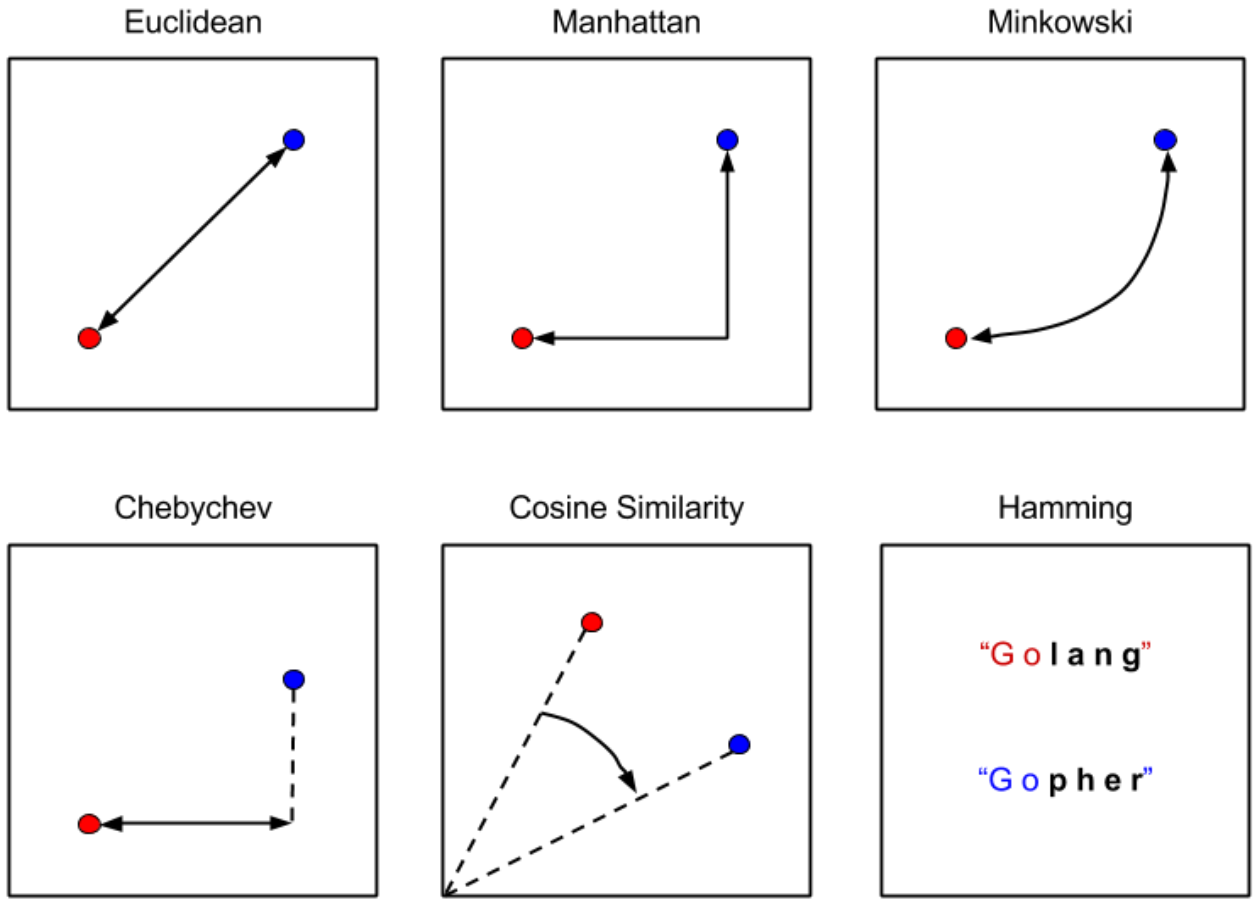
È importante fare attenzione alla scala dei valori (specie nel calcolo delle DISTANZE):



La Cluster Analysis

LA SCELTA DELLE VARIABILI (DESCRITTORI)

1. $d_{st} = \sqrt{\sum_{j=1}^{p} (x_{sj} - x_{tj})^2}$ distanza Euclidea
2. $d_{st} = \sum_j |x_{sj} - x_{tj}|$ distanza di Manhattan
3. $d_{st} = \max_j |x_{sj} - x_{tj}|$ distanza di Chebyshev o Lagrange
4. $d_{st} = \sum_{j=1}^{p} \frac{|x_{sj} - x_{tj}|}{(x_{sj} + x_{tj})}$ distanza di Camberra
5. $d_{st} = \frac{\sum_j |x_{sj} - x_{tj}|}{\sum_j (x_{sj} + x_{tj})}$ distanza di Lance-Williams
6. $d_{st} = \sqrt[r]{\sum_j |x_{sj} - x_{tj}|^r}$ distanza di Minkowski
7. $d_{st} = \sqrt{(\mathbf{x_s} - \mathbf{x_t})^{\mathbf{T}} \mathbf{S}^{-1} (\mathbf{x_s} - \mathbf{x_t})}$ distanza di Mahalanobis
8. $d_{st} = \sqrt{\sum_j \frac{(x_{sj} - x_{tj})^2}{s_j^2}}$ distanza di Pearson



La Cluster Analysis

LA SCELTA DELLE VARIABILI (DESCRITTORI)

1. $d_{st} = \sqrt{\sum_{j=1}^{p} (x_{sj} - x_{tj})^2}$ distanza Euclidea

2. $d_{st} = \sum_j |x_{sj} - x_{tj}|$ distanza di Manhattan

3. $d_{st} = \max_j |x_{sj} - x_{tj}|$ distanza di Chebyshev o Lagrange

4. $d_{st} = \sum_{j=1}^{p} \frac{|x_{sj} - x_{tj}|}{(x_{sj} + x_{tj})}$ distanza di Camberra

5. $d_{st} = \frac{\sum_j |x_{sj} - x_{tj}|}{\sum_j (x_{sj} + x_{tj})}$ distanza di Lance-Williams

6. $d_{st} = \sqrt[r]{\sum_j |x_{sj} - x_{tj}|^r}$ distanza di Minkowski

7. $d_{st} = \sqrt{(\mathbf{x_s} - \mathbf{x_t})^{\mathbf{T}} \mathbf{S}^{-1} (\mathbf{x_s} - \mathbf{x_t})}$ distanza di Mahalanobis

8. $d_{st} = \sqrt{\sum_j \frac{(x_{sj} - x_{tj})^2}{s_j^2}}$ distanza di Pearson

Le distanze di Kendall e Spearman sono non-parametriche (rango)

La Cluster Analysis

IL CONCETTO DI SIMILARITÀ (DISTANZA)

La similarità tra due oggetti (i e j) o tra due variabili è così definita:

$$s_{ij} = 1 - \frac{d_{ij}}{d_{MAX}}$$

d_{ij} è la distanza (Euclidea, Mahalanobis, Minkowski…) tra due oggetti

d_{MAX} è la massima distanza che esiste tra due oggetti del dataset

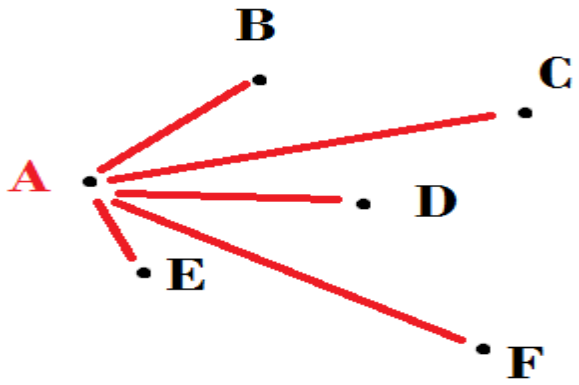
Due oggetti aventi la massima distanza avranno similarità = 0!!!

Dunque la similarità dipende dal dataset…

La Cluster Analysis

IL CONCETTO DI SIMILARITÀ (DISTANZA)

CALCOLO DELLE DISTANZE TRA GLI OGGETTI



COSTRUZIONE MATRICE DELLE DISTANZE

MATRICE DELLE DISTANZE

	A	B	C	D	E	F
A	0	…	…	…	…	…
B	…	0	…	…	…	…
C	…	…	0	…	…	…
D	…	…	…	0	…	…
E	…	…	…	…	0	…
F	…	…	…	…	…	0

MATRICE DI SIMILARITA’

	A	B	C	D	E	F
A	1	…	…	…	…	…
B	…	1	…	…	…	…
C	…	…	1	…	…	…
D	…	…	…	1	…	…
E	…	…	…	…	1	…
F	…	…	…	…	…	1

La Cluster Analysis

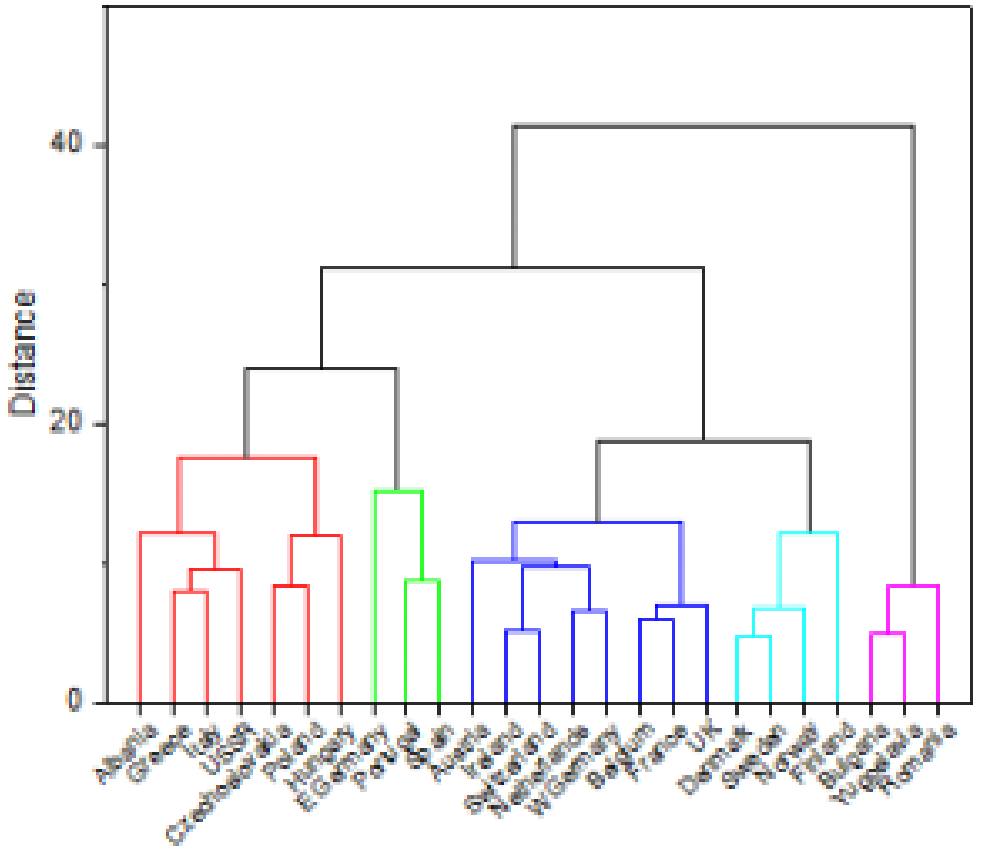
TECNICHE DI CLUSTERING

Suddivise in:

a) Gerarchiche

- a_1) Agglomerative
- a_2) Divisive

b) Non gerarchiche

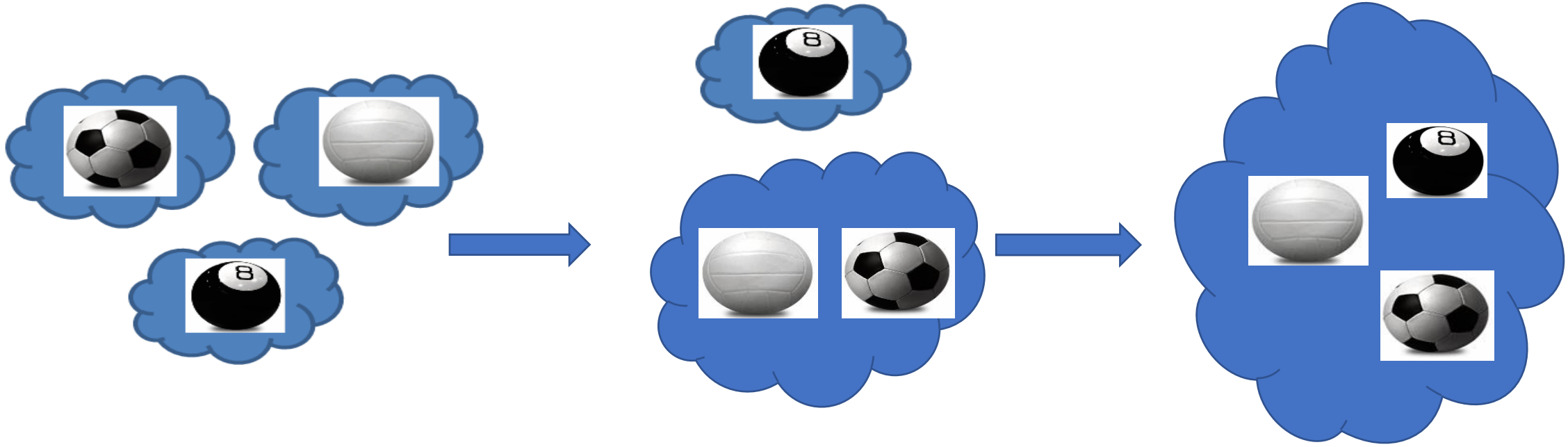


La Cluster Analysis

TECNICHE DI CLUSTERING: GERARCHICHE AGGLOMERATIVE

- ***Inizialmente ogni oggetto viene considerato un cluster***
- ***Gradualmente gli oggetti vengono riuniti in clusters sempre più grandi***

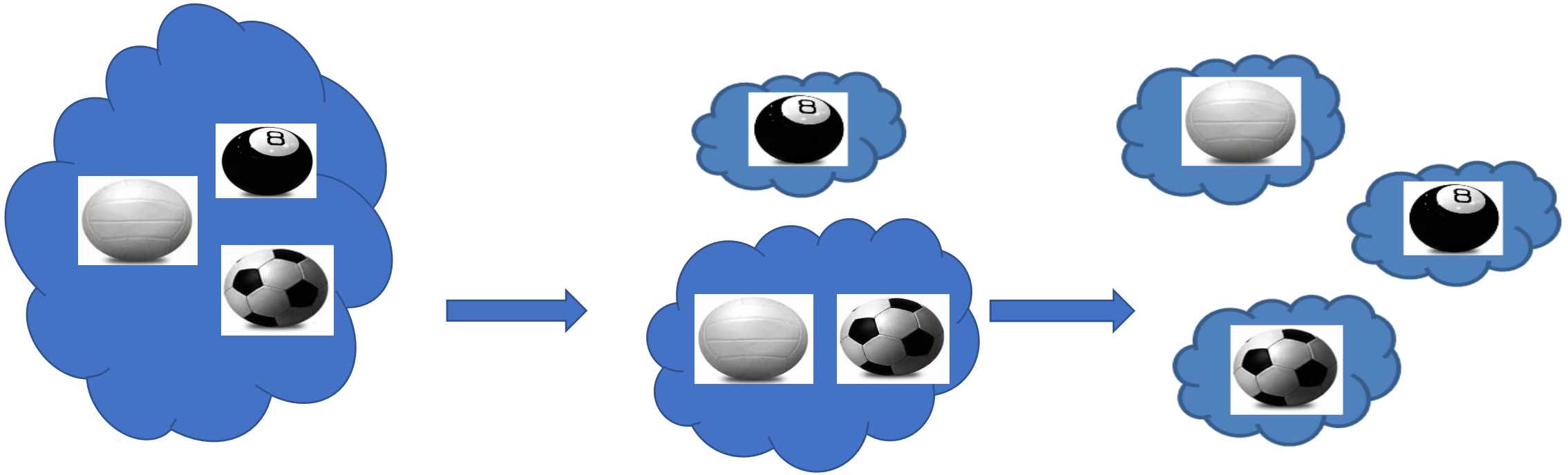
La gerarchia è una conseguenza del fatto che i clusters più grandi sono sempre ottenuti dalla fusione di quelli più piccoli (con tutti i loro oggetti).



La Cluster Analysis

TECNICHE DI CLUSTERING: GERARCHICHE DIVISIVE

- Si parte da un grande cluster contenente tutti gli oggetti
- Gradualmente gli oggetti vengono inseriti in piccoli clusters ottenuti come sottoinsiemi di quello iniziale



La Cluster Analysis

TECNICHE DI CLUSTERING: GERARCHICHE AGGLOMERATIVE

1) Metodo del legame medio non pesato

2) Metodo del legame medio pesato

3) Metodo del legame completo

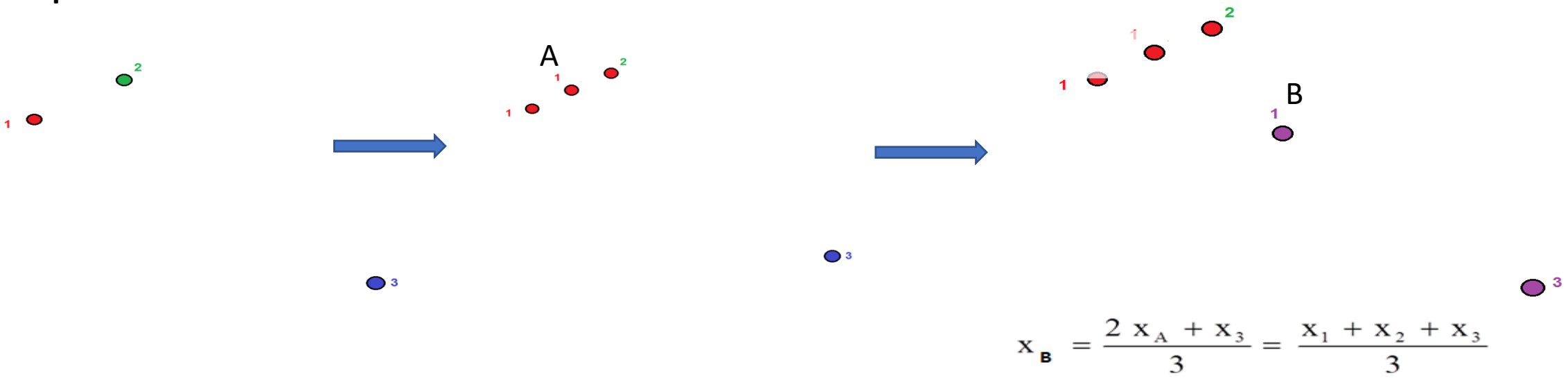
4) Metodo del legame singolo

5) Altri…

La Cluster Analysis

TECNICHE DI CLUSTERING: Metodo del legame medio non pesato

La posizione del centroide del nuovo cluster è influenzata del numero di oggetti presenti nei due clusters di partenza.

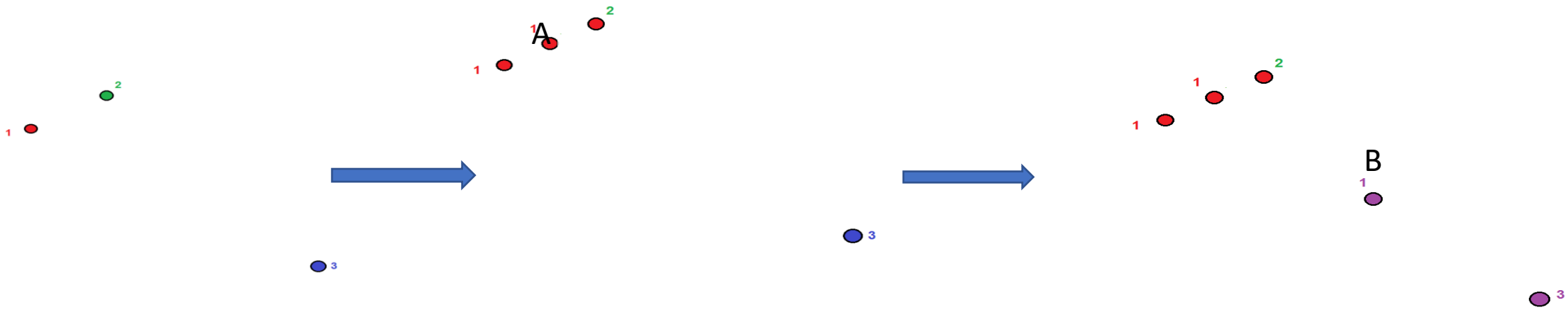


$$x_B = \frac{2\,x_A + x_3}{3} = \frac{x_1 + x_2 + x_3}{3}$$

La Cluster Analysis

TECNICHE DI CLUSTERING: Metodo del legame medio pesato

La posizione del centroide del nuovo cluster non risente del numero di oggetti presenti nei due clusters di partenza.



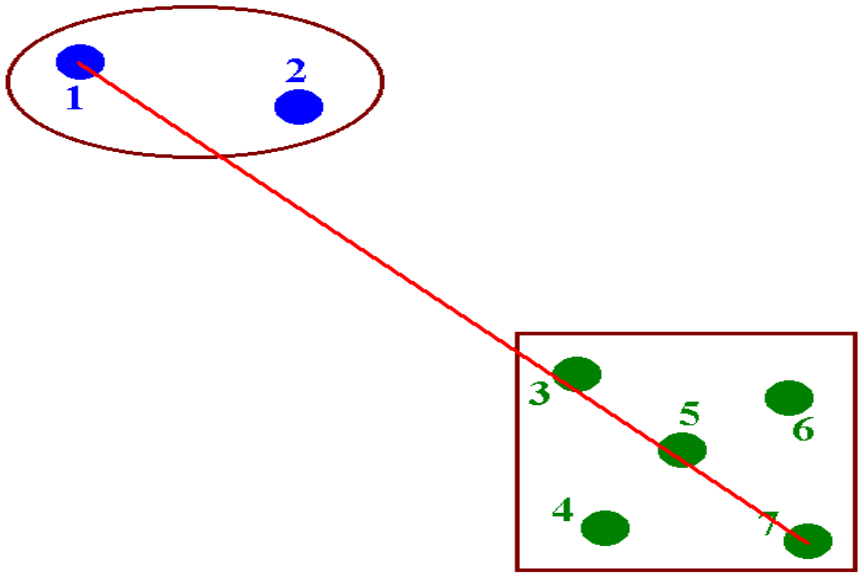
$$x_A = \frac{x_1 + x_2}{2} \qquad x_B = \frac{x_A + x_3}{2} = .25\,x_1 + .25\,x_2 + .5\,x_3$$

La Cluster Analysis

TECNICHE DI CLUSTERING: Metodi del legame completo e singolo

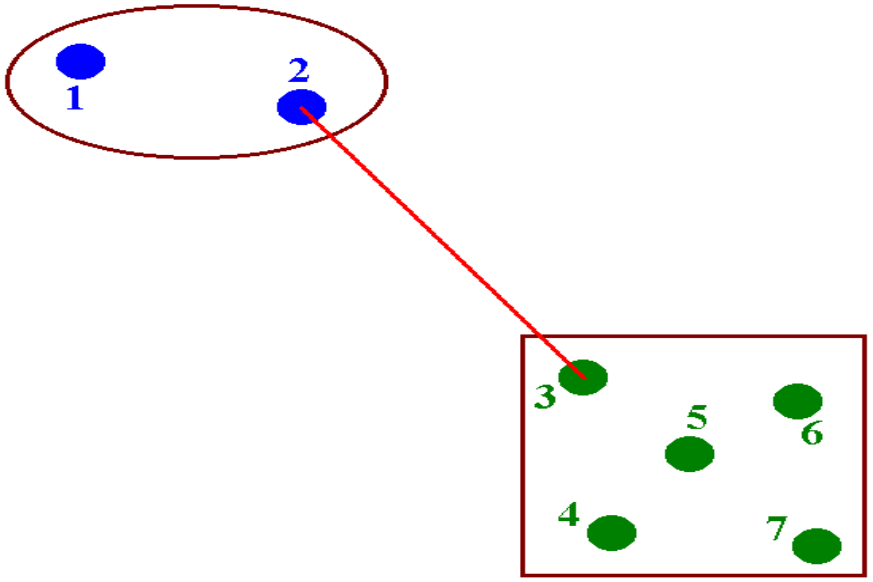
Metodo del legame completo

I clusters si uniscono considerando la massima distanza tra due oggetti individuali, uno di ogni cluster.



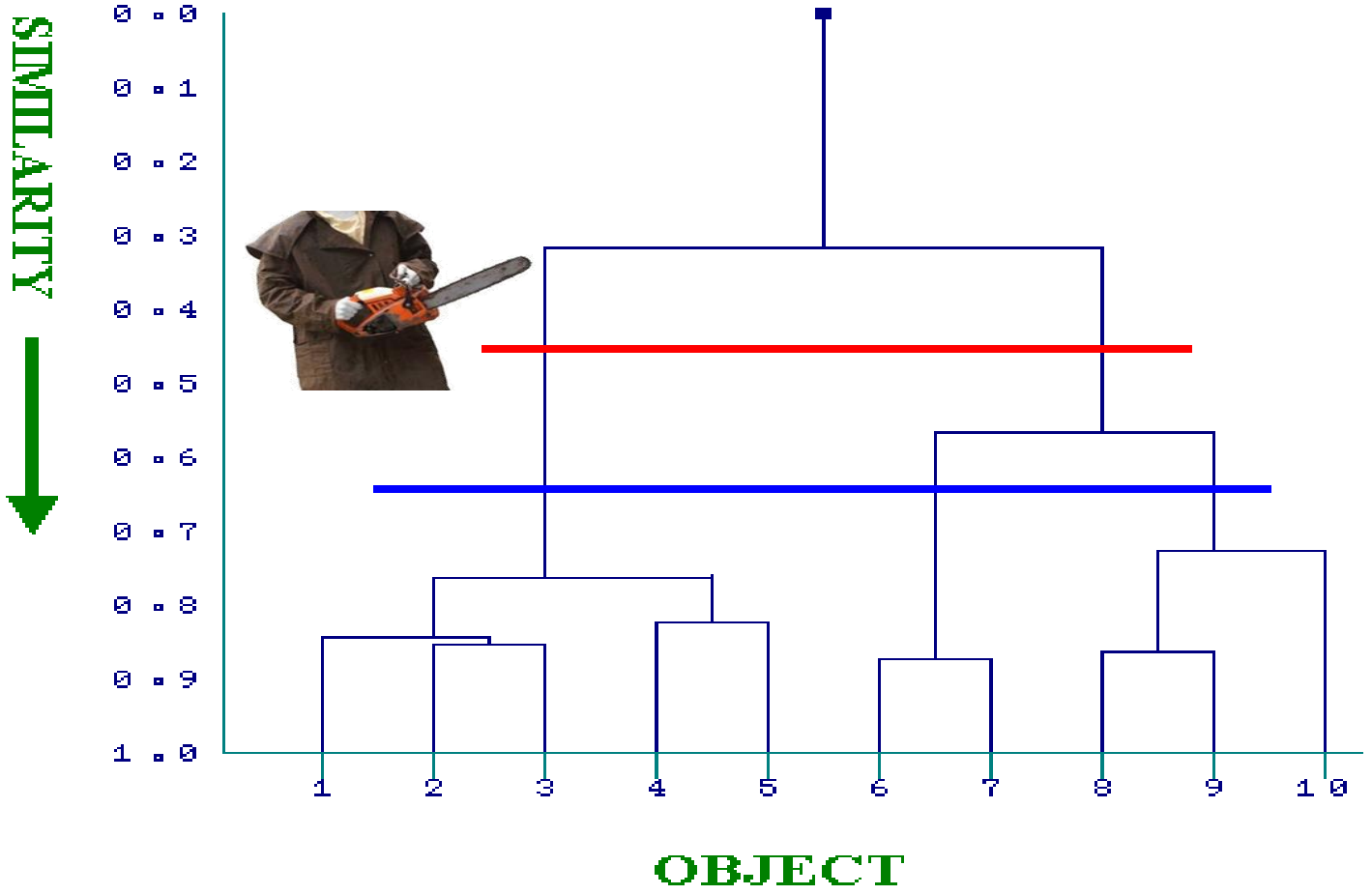
Metodo del legame singolo

I cluster si uniscono considerando la minima distanza tra gli oggetti individuali di ogni cluster. Tale tecnica è detta anche “natural clustering”….



La Cluster Analysis

IL DENDROGRAMMA



Normalmente i rami più lunghi indicano i clusters separati meglio

La FASE INTERPRETATIVA è più importante rispetto al test statistico

La Cluster Analysis

IL DENDROGRAMMA

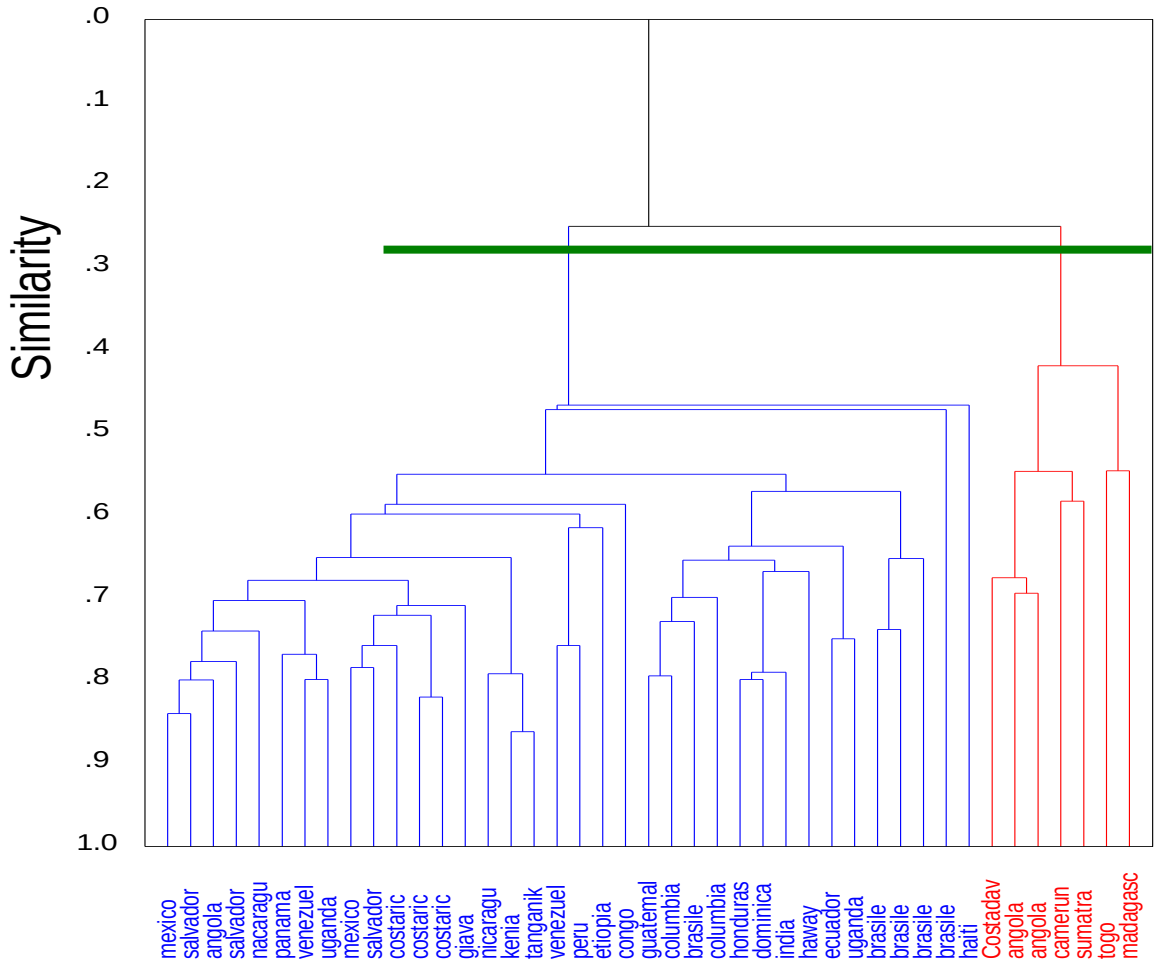
Data set COFFEE

- Due categorie: arabica e robusta

- Pretrattamento dei dati: autoscaling

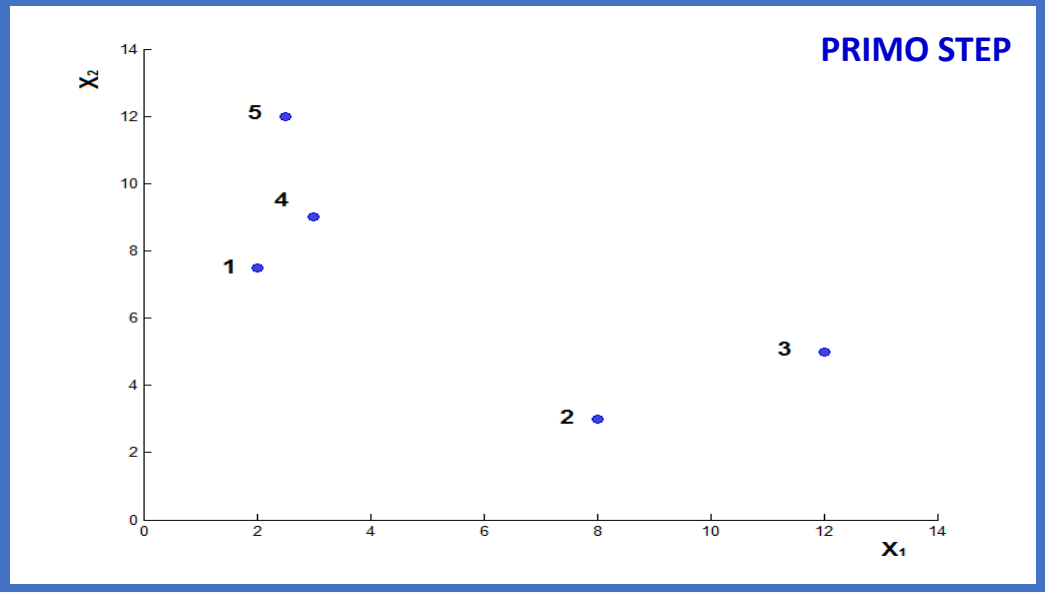
- Tecnica usata: metodo del legame medio non pesato

Tagliando ad un livello di similarità pari a 0.3 si ottengono due clusters che separano le due categorie

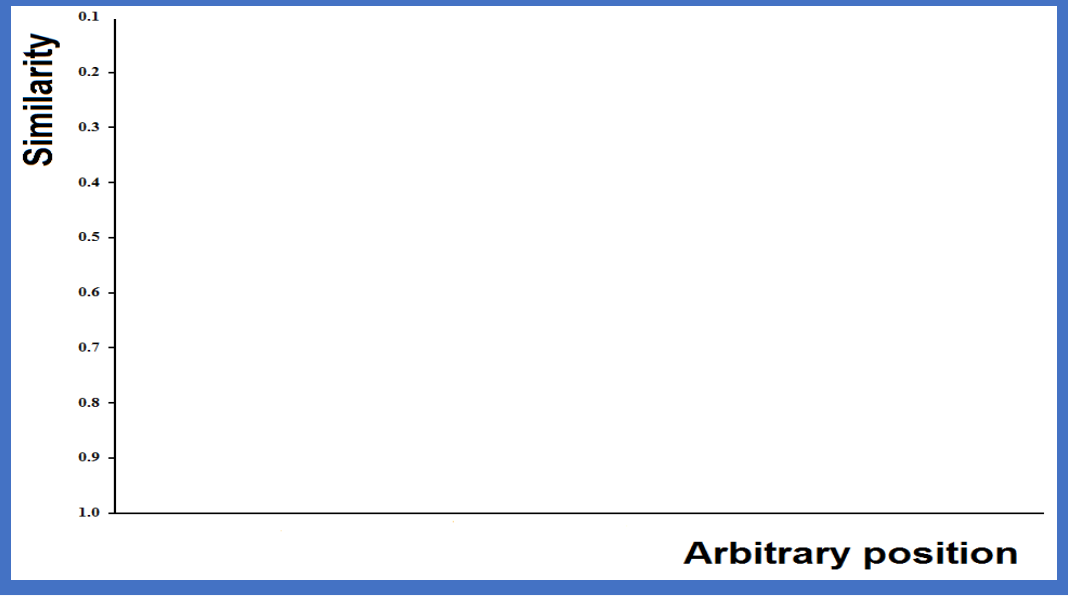


Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



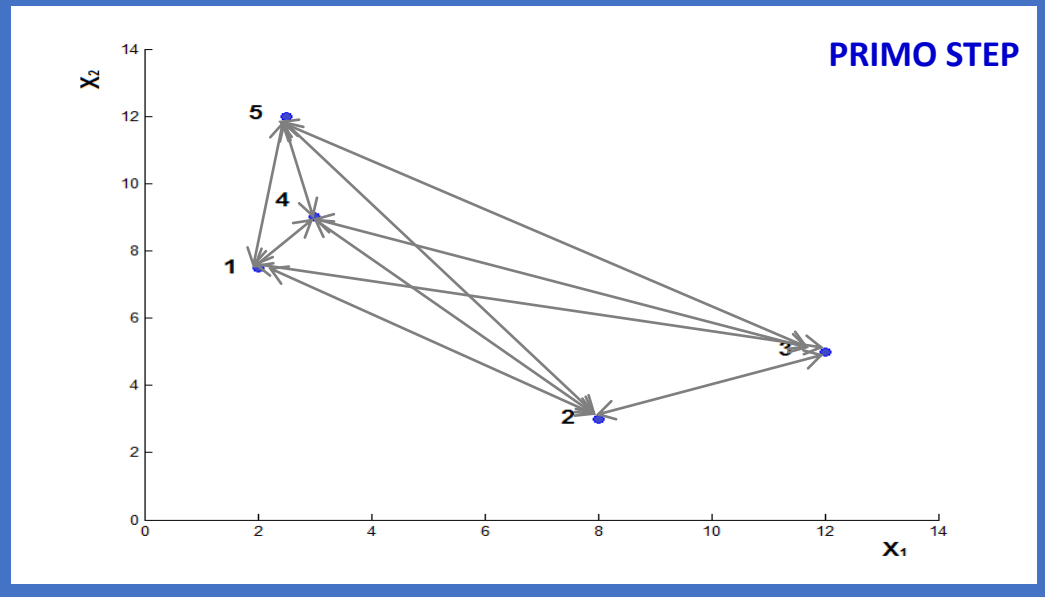
DATA SET

	X1	X2
1	2	7.5
2	8	3
3	12	5
4	3	9
5	2.5	12

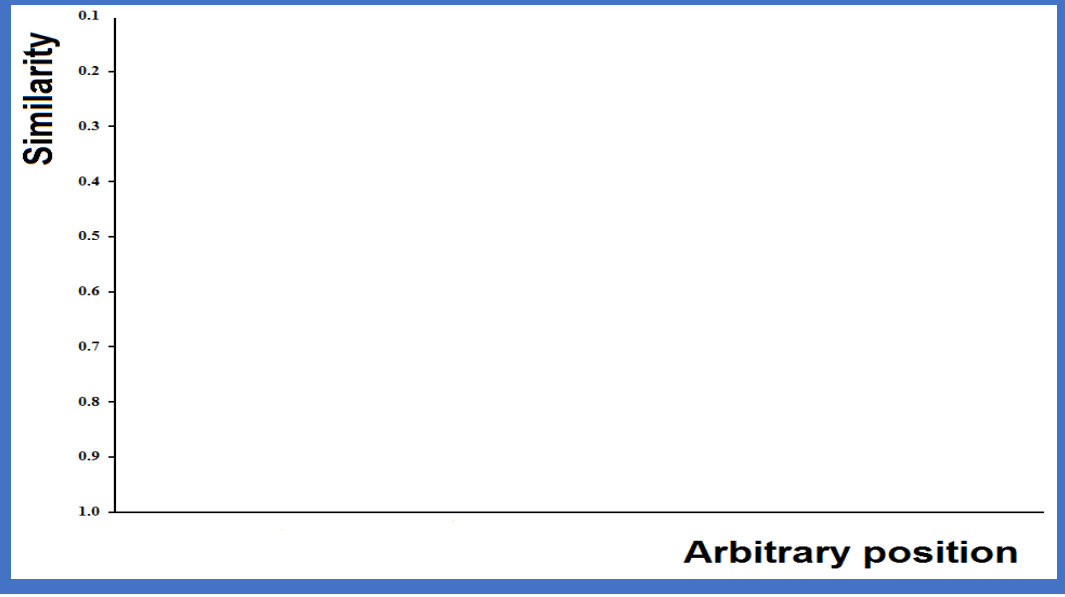
MATRICE DELLE SIMILARITÀ

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

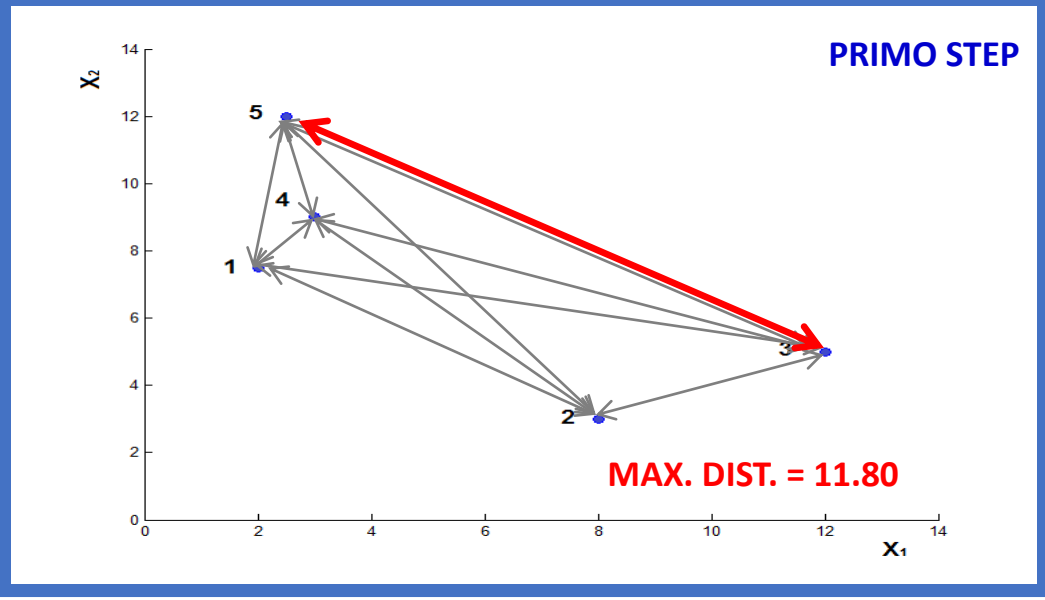
	1	2	3	4	5
1	0.00	7.50	10.31	1.80	4.53
2	7.50	0.00	4.47	7.81	10.55
3	10.31	4.47	0.00	9.85	11.80
4	1.80	7.81	9.85	0.00	3.04
5	4.53	10.55	11.80	3.04	0.00

MATRICE DELLE SIMILARITÀ

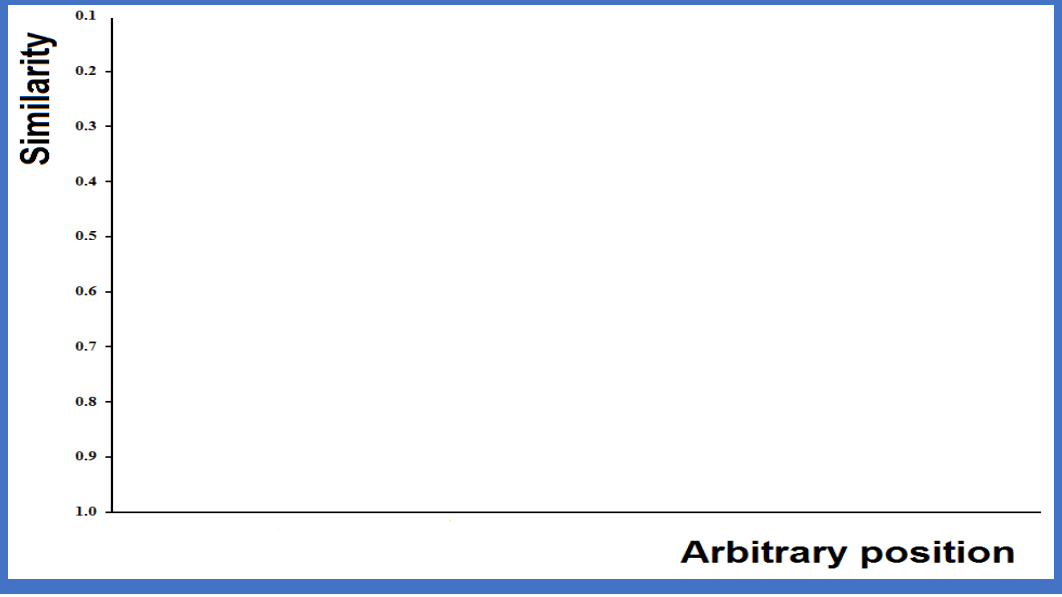
(è stata scelta la distanza euclidea)

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



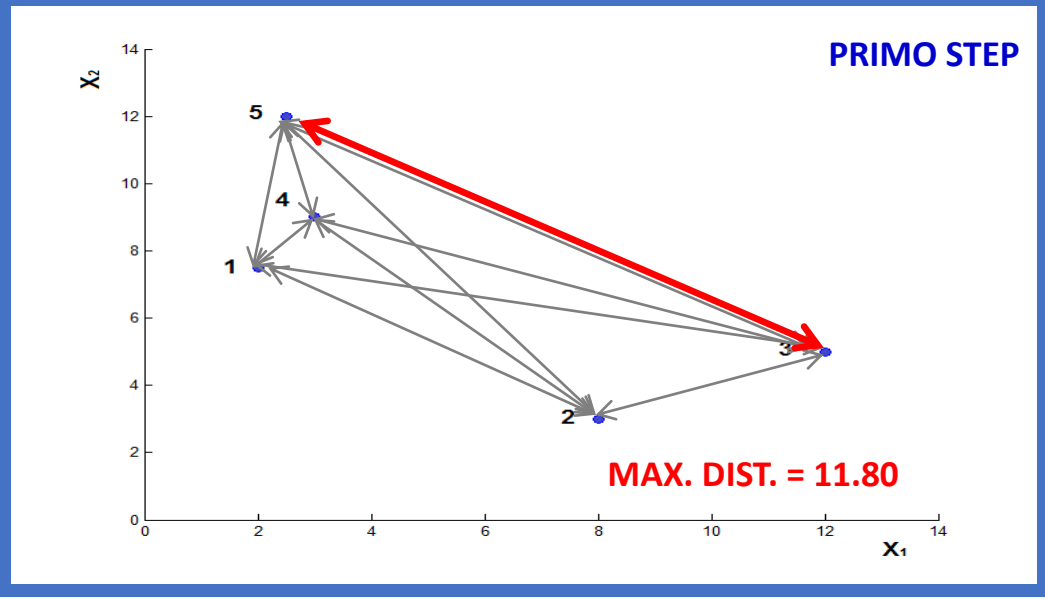
MATRICE DELLE DISTANZE

	1	2	3	4	5
1	0.00	7.50	10.31	1.80	4.53
2	7.50	0.00	4.47	7.81	10.55
3	10.31	4.47	0.00	9.85	11.80
4	1.80	7.81	9.85	0.00	3.04
5	4.53	10.55	11.80	3.04	0.00

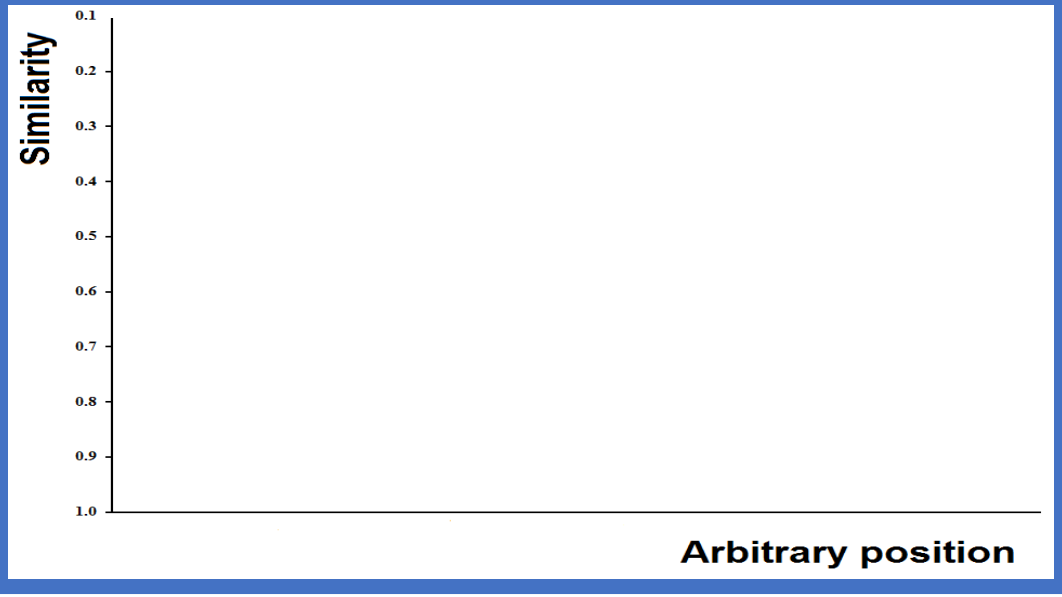
MATRICE DELLE SIMILARITÀ

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

	1	2	3	4	5
1	0.00	7.50	10.31	1.80	4.53
2	7.50	0.00	4.47	7.81	10.55
3	10.31	4.47	0.00	9.85	11.80
4	1.80	7.81	9.85	0.00	3.04
5	4.53	10.55	11.80	3.04	0.00

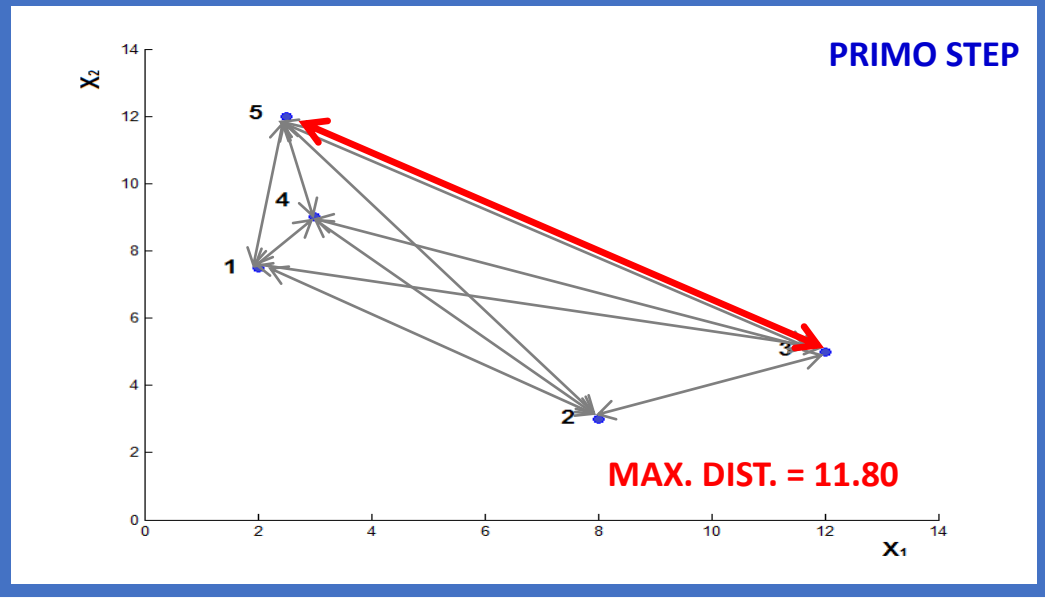
MAX. DIST. = 11.80

MATRICE DELLE SIMILARITÀ

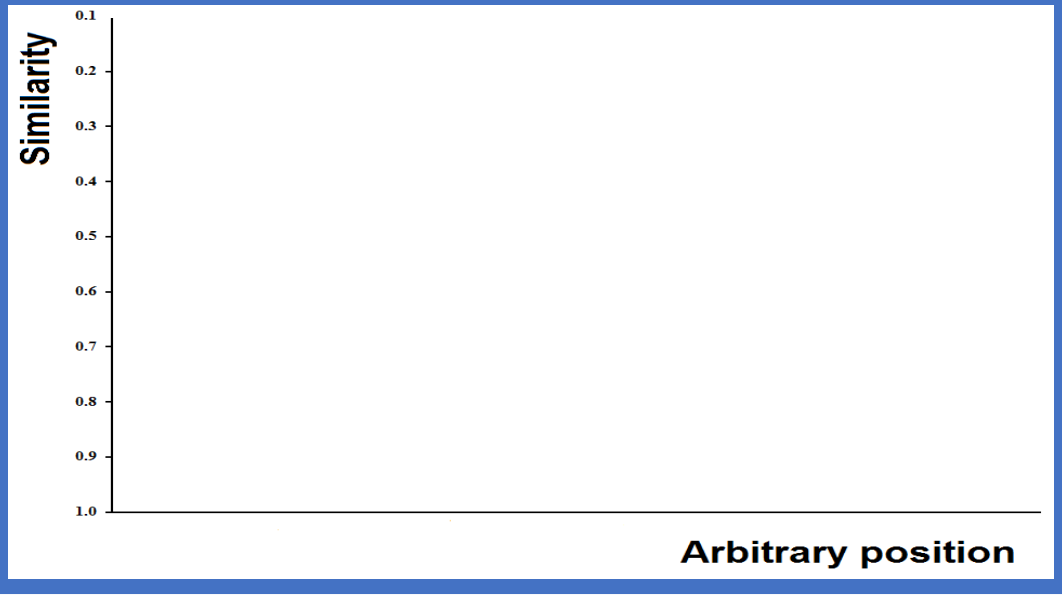
	1	2	3	4	5
1	1.00	0.36	0.13	0.85	0.62
2	0.36	1.00	0.62	0.34	0.11
3	0.13	0.62	1.00	0.17	0.00
4	0.85	0.34	0.17	1.00	0.74
5	0.62	0.11	0.00	0.74	1.00

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

	1	2	3	4	5
1	0.00	7.50	10.31	1.80	4.53
2	7.50	0.00	4.47	7.81	10.55
3	10.31	4.47	0.00	9.85	11.80
4	1.80	7.81	9.85	0.00	3.04
5	4.53	10.55	11.80	3.04	0.00

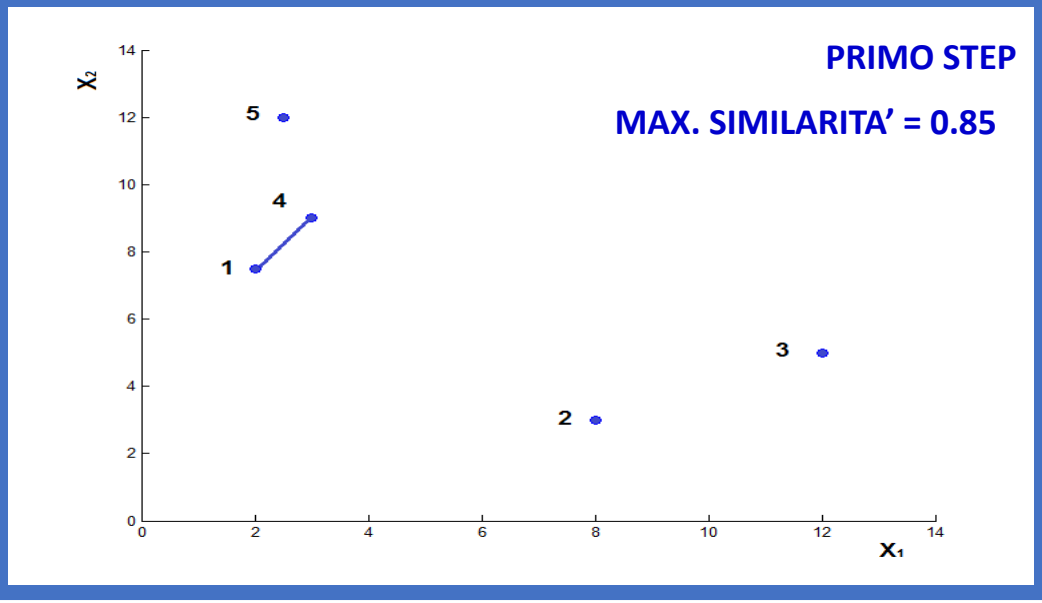
MAX. DIST. = 11.80

MATRICE DELLE SIMILARITÀ

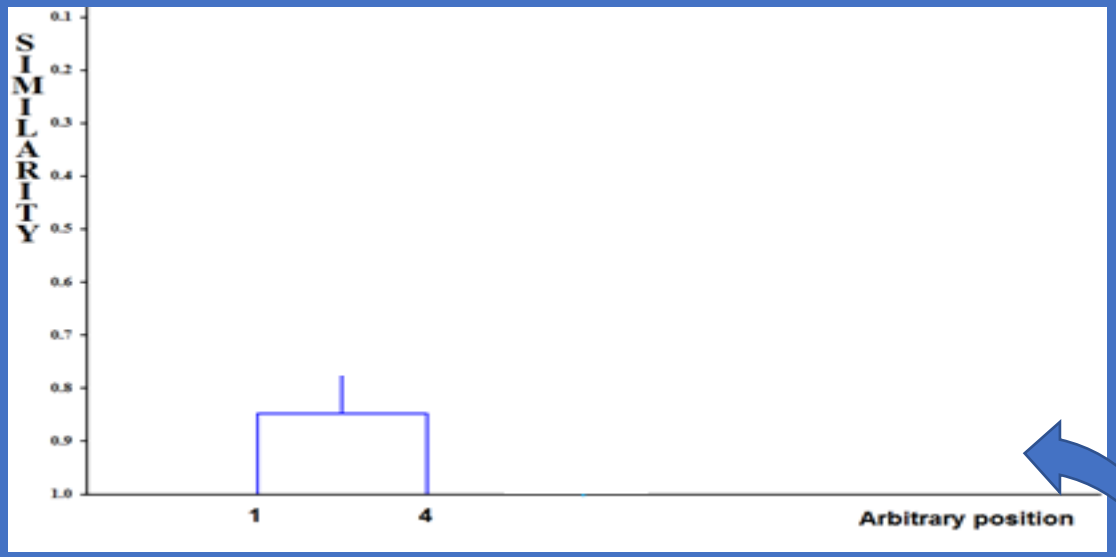
	1	2	3	4	5
1	1.00	0.36	0.13	0.85	0.62
2	0.36	1.00	0.62	0.34	0.11
3	0.13	0.62	1.00	0.17	0.00
4	0.85	0.34	0.17	1.00	0.74
5	0.62	0.11	0.00	0.74	1.00

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

	1	2	3	4	5
1	0.00	7.50	10.31	1.80	4.53
2	7.50	0.00	4.47	7.81	10.55
3	10.31	4.47	0.00	9.85	11.80
4	1.80	7.81	9.85	0.00	3.04
5	4.53	10.55	11.80	3.04	0.00

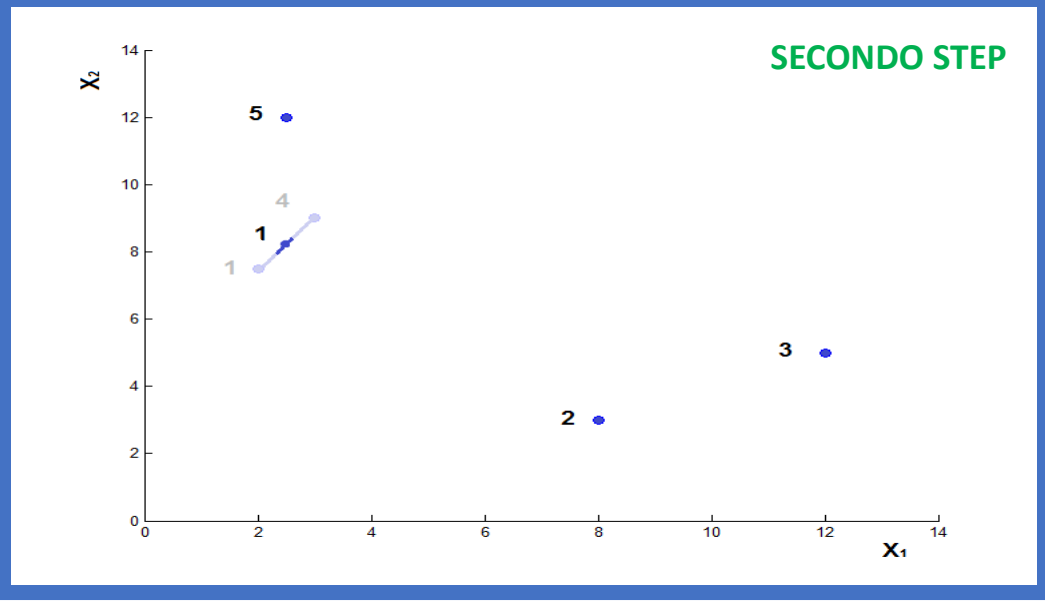
MAX. DIST. = 11.80

MATRICE DELLE SIMILARITÀ

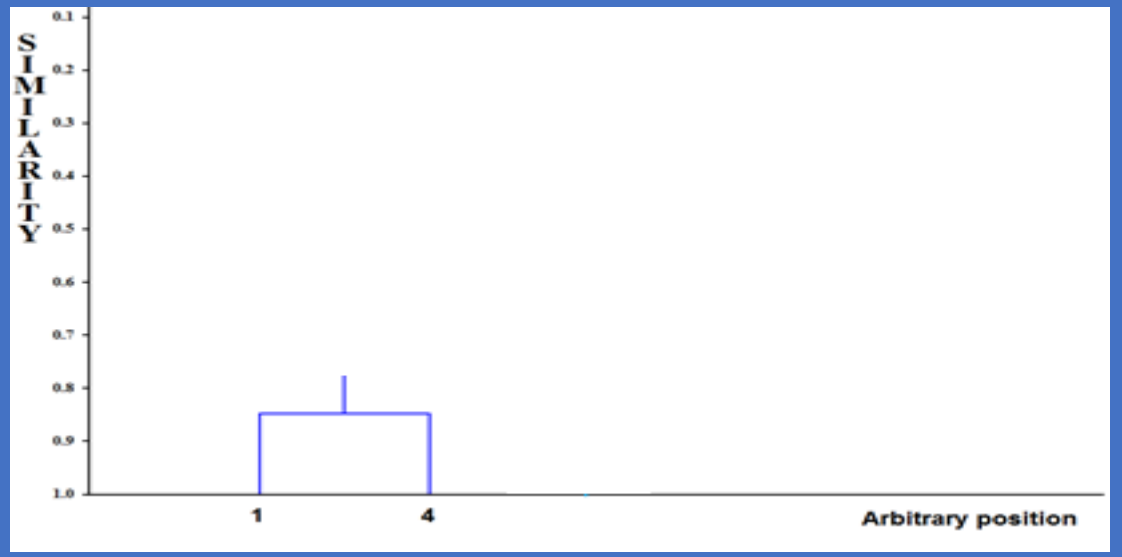
	1	2	3	4	5
1	1.00	0.36	0.13	0.85	0.62
2	0.36	1.00	0.62	0.34	0.11
3	0.13	0.62	1.00	0.17	0.00
4	0.85	0.34	0.17	1.00	0.74
5	0.62	0.11	0.00	0.74	1.00

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA

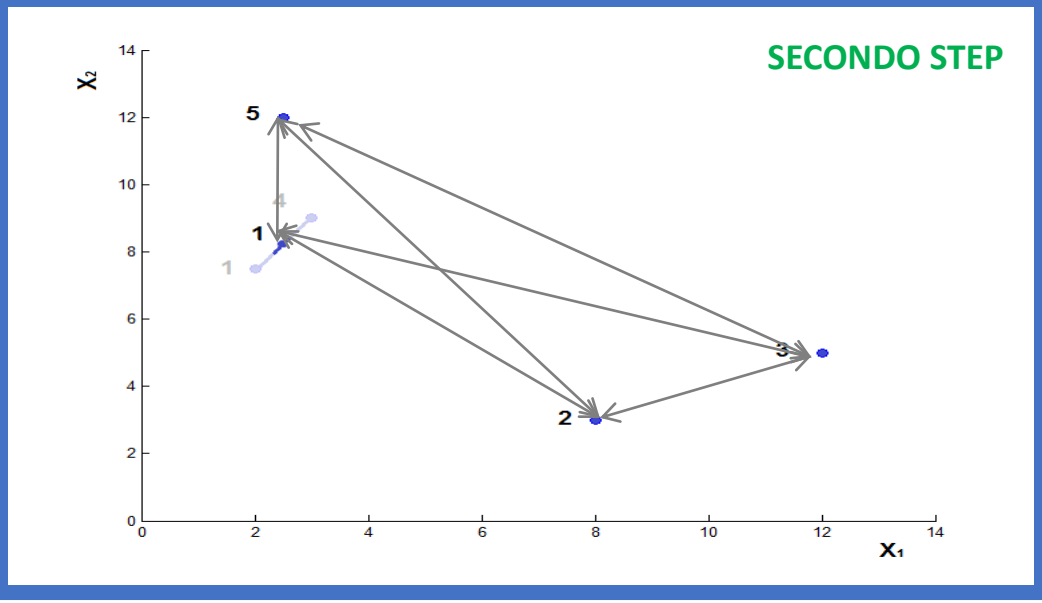


MATRICE DELLE DISTANZE

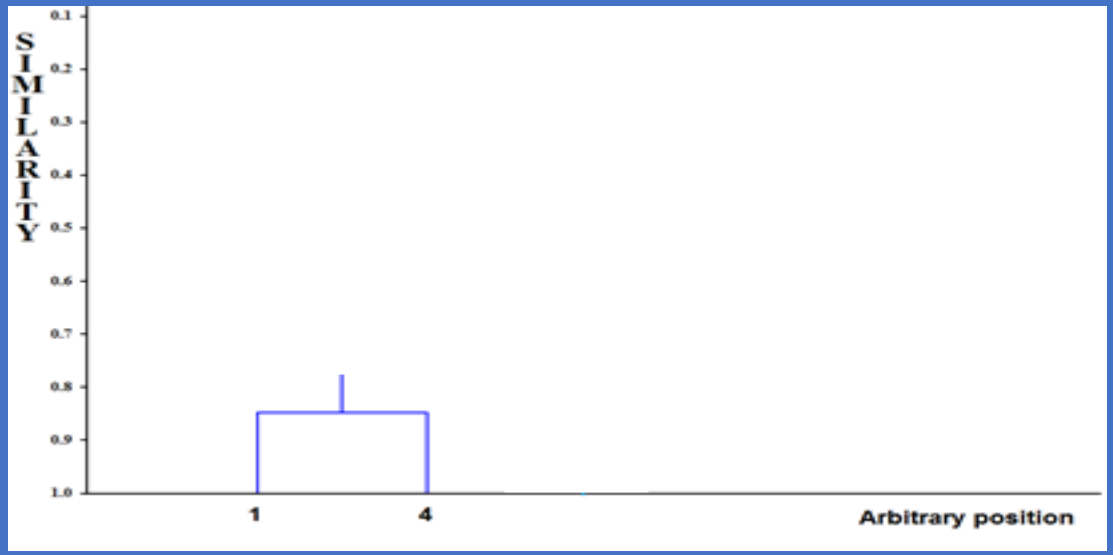
MATRICE DELLE SIMILARITÀ

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE

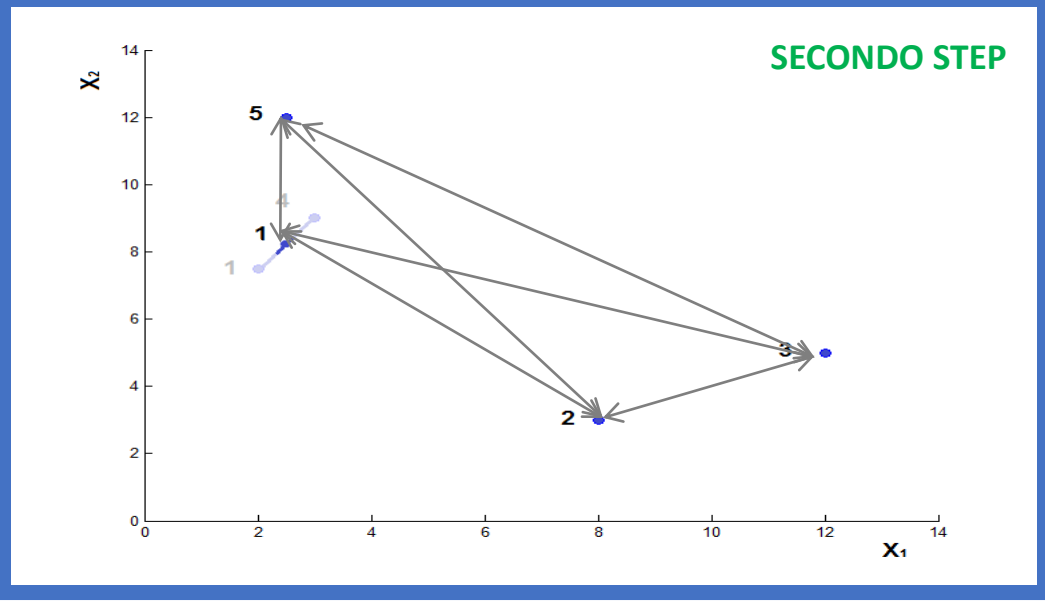


DENDROGRAMMA

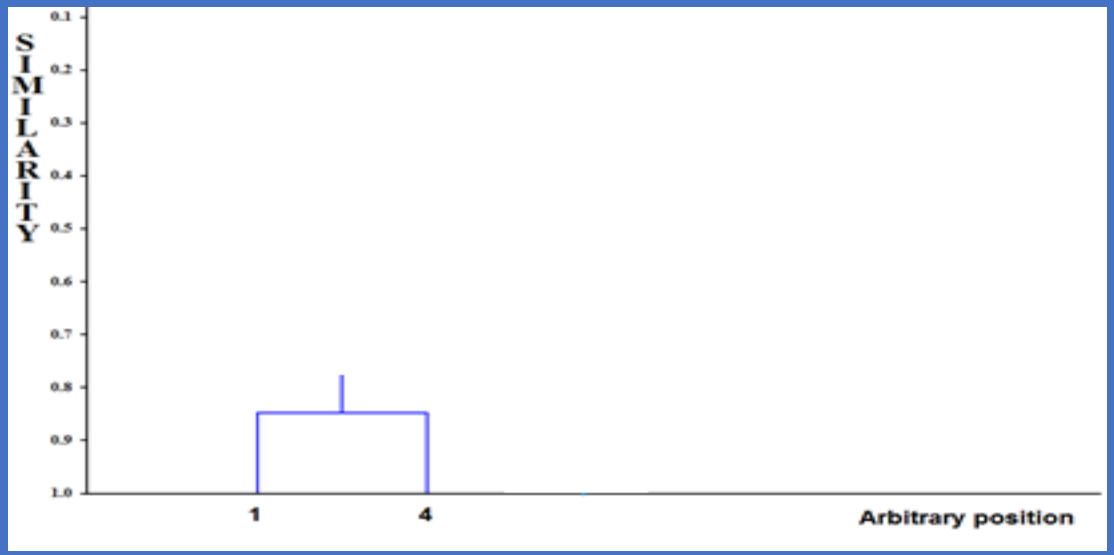


Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA

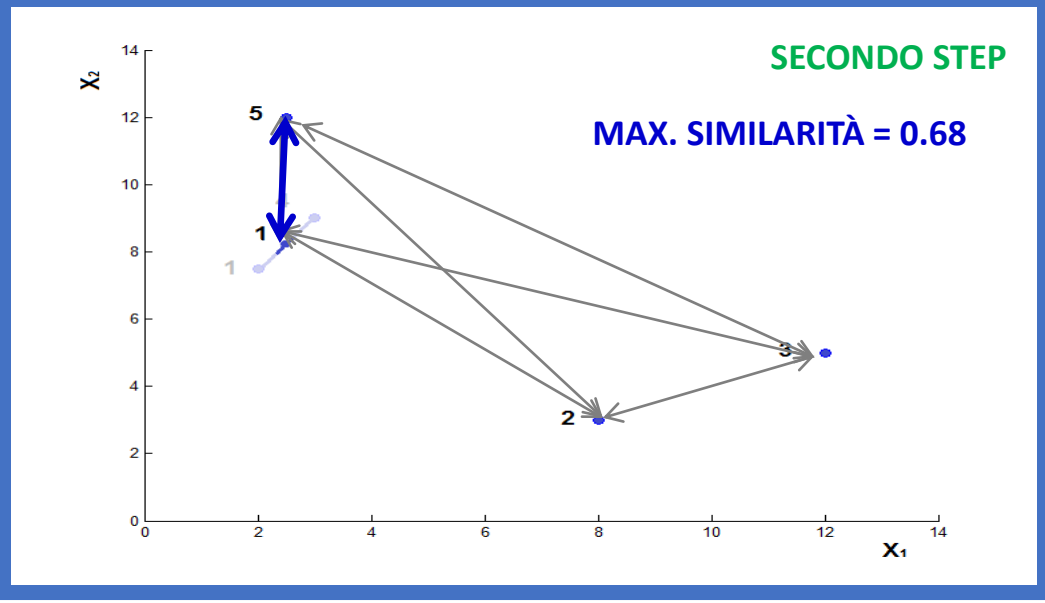


MATRICE DELLE DISTANZE

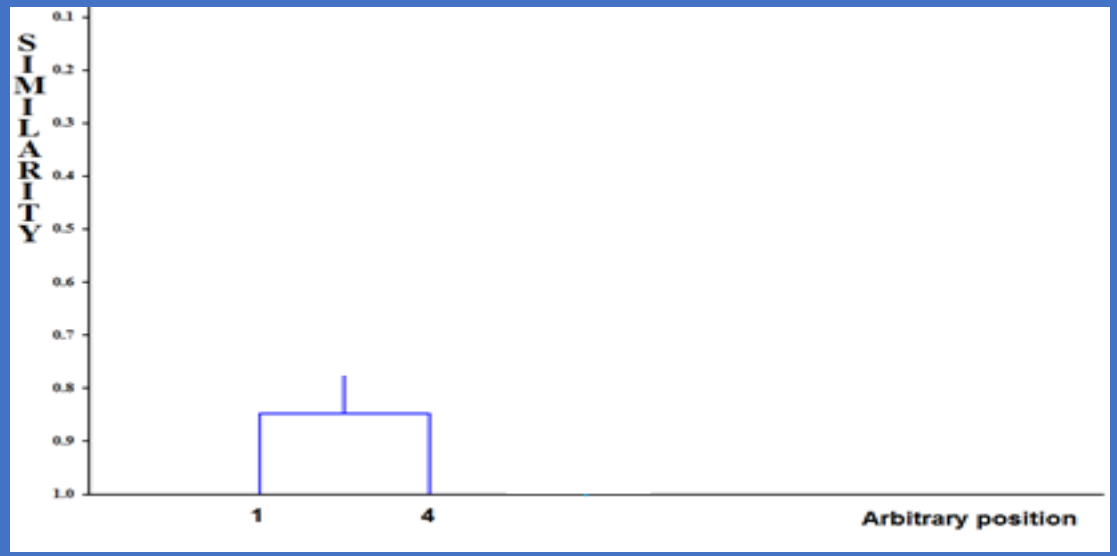
	1	2	3	5
1	0.00	7.60	10.04	3.75
2	7.60	0.00	4.47	10.55
3	10.04	4.47	0.00	11.80
5	3.75	10.55	11.80	0.00

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

	1	2	3	5
1	0.00	7.60	10.04	3.75
2	7.60	0.00	4.47	10.55
3	10.04	4.47	0.00	11.80
5	3.75	10.55	11.80	0.00

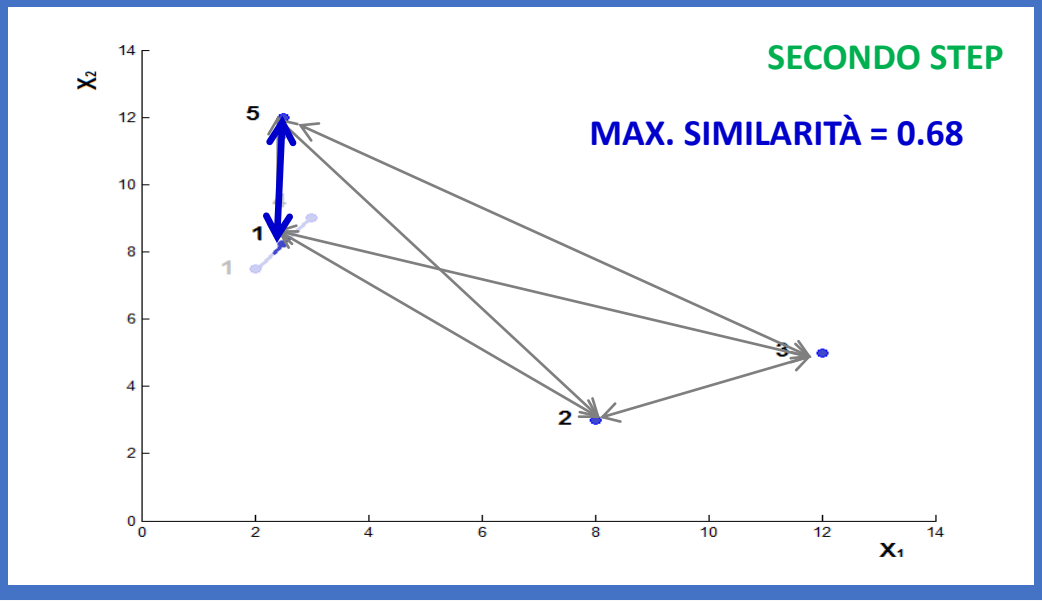
MAX. DIST. = 11.80

MATRICE DELLE SIMILARITÀ

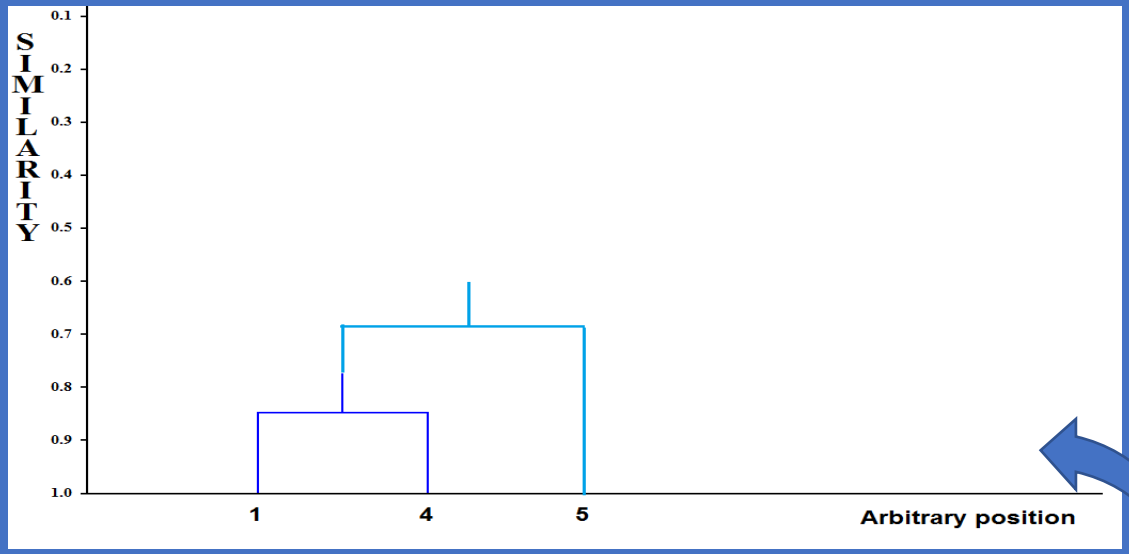
	1	2	3	5
1	1.00	0.36	0.15	0.68
2	0.36	1.00	0.62	0.11
3	0.15	0.62	1.00	0.00
5	0.68	0.11	0.00	1.00

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

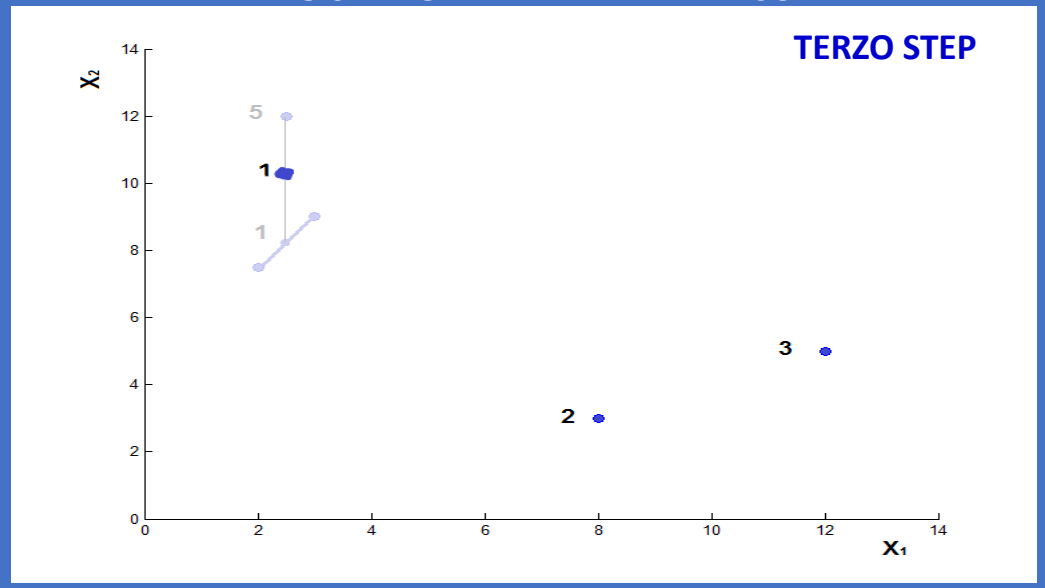
	1	2	3	5
1	0.00	7.60	10.04	3.75
2	7.60	0.00	4.47	10.55
3	10.04	4.47	0.00	11.80
5	3.75	10.55	11.80	0.00

MATRICE DELLE SIMILARITÀ

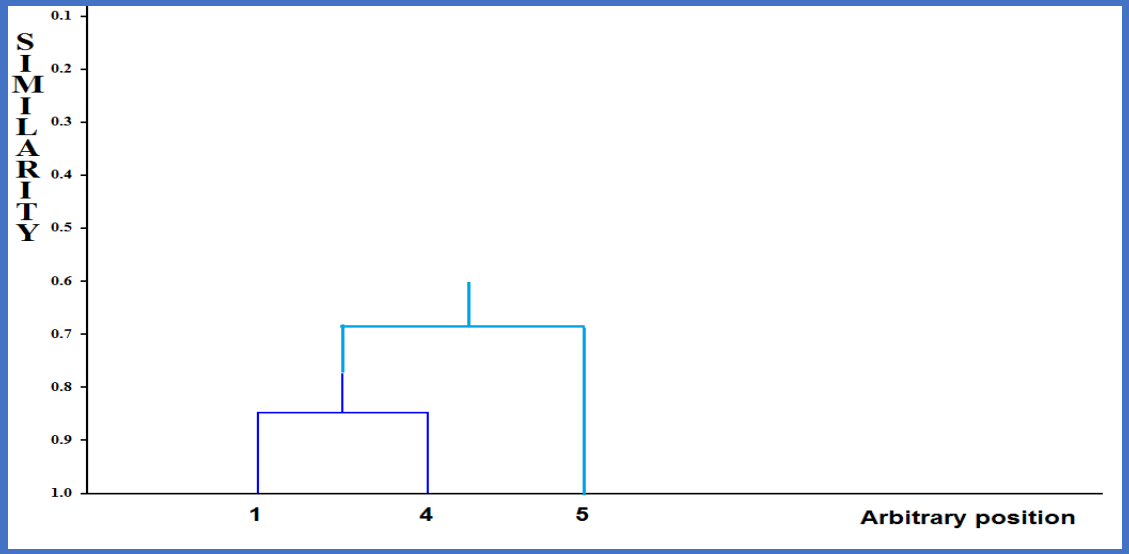
	1	2	3	5
1	1.00	0.36	0.15	**0.68**
2	0.36	1.00	0.62	0.11
3	0.15	0.62	1.00	0.00
1	**0.68**	0.11	0.00	1.00

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA

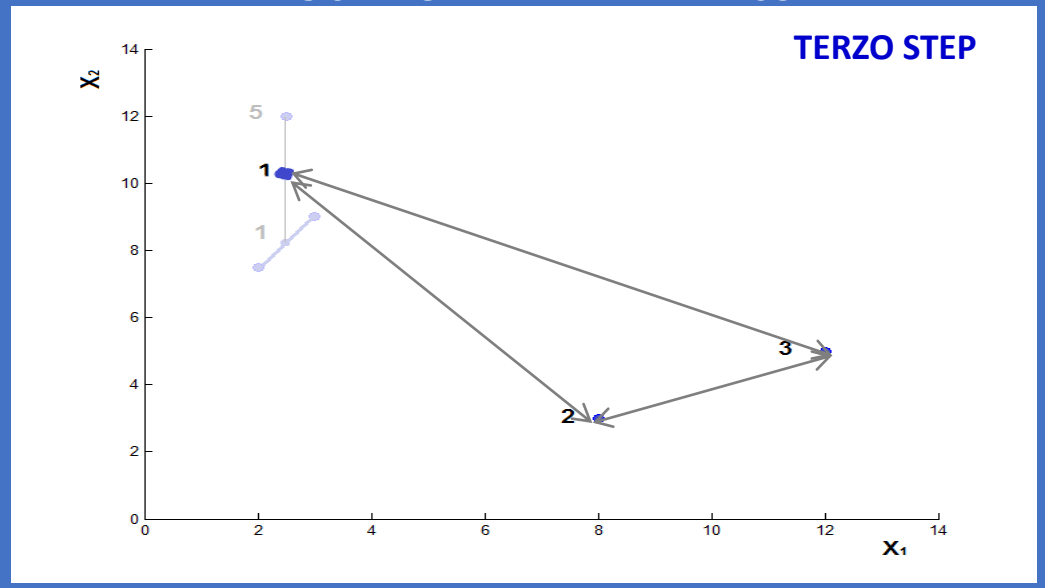


MATRICE DELLE DISTANZE

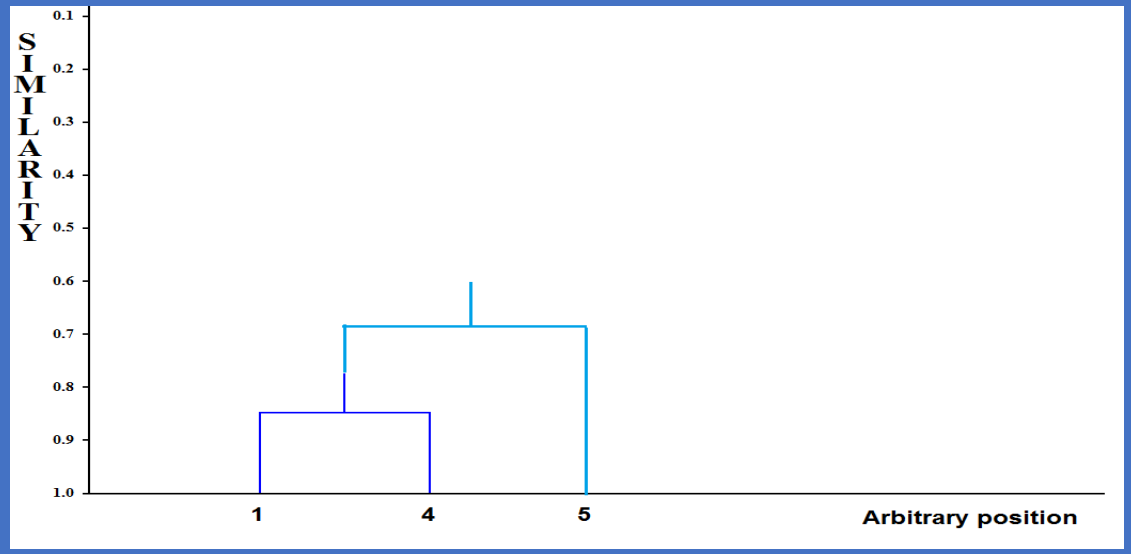
MATRICE DELLE SIMILARITÀ

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



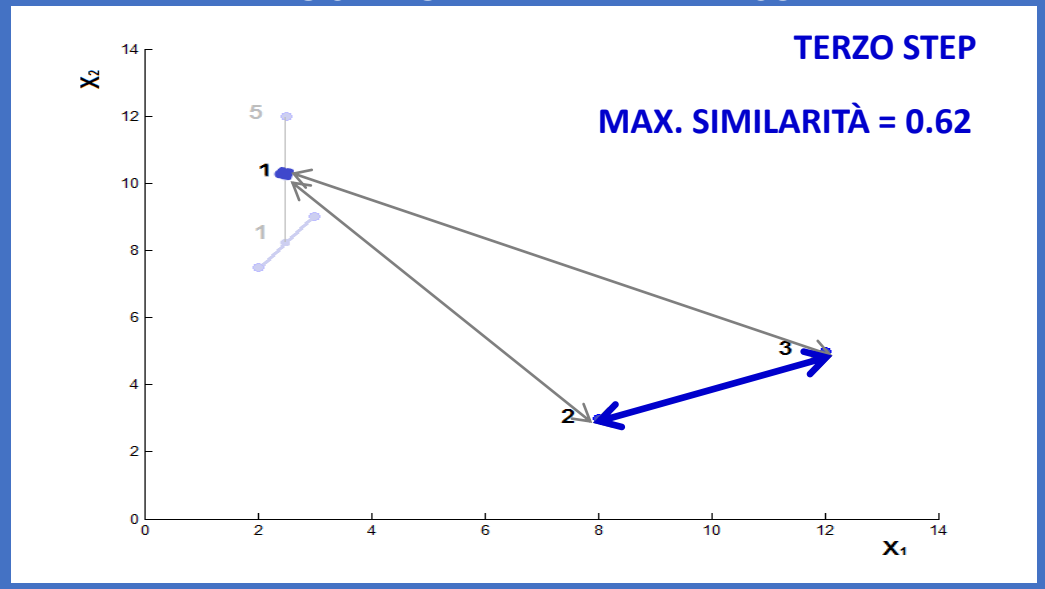
MATRICE DELLE DISTANZE

	1	2	3
1	0.00	9.00	10.79
2	9.00	0.00	4.47
3	10.79	4.47	0.00

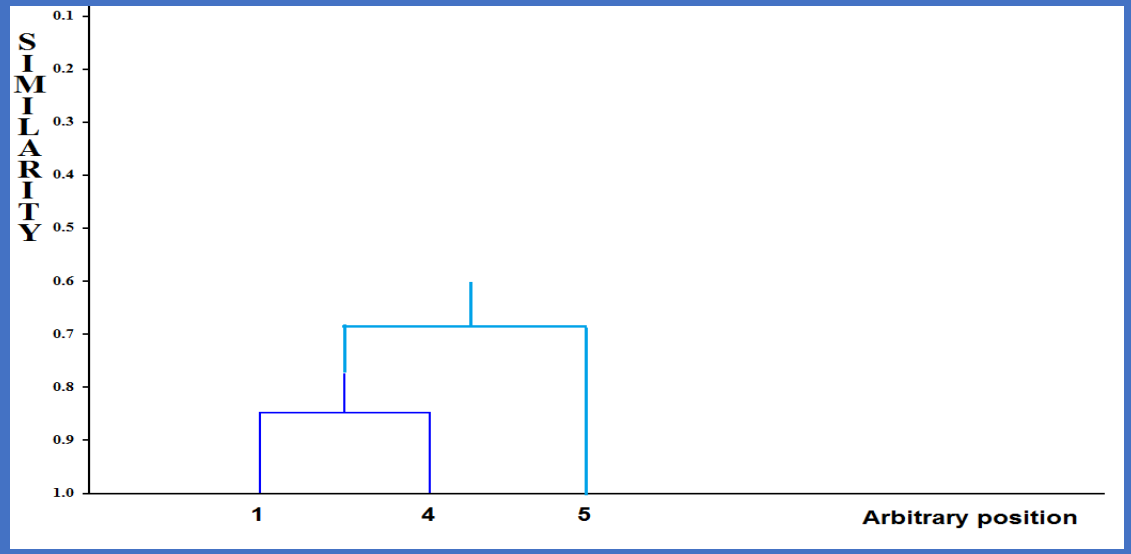
MATRICE DELLE SIMILARITÀ

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

	1	2	3
1	0.00	9.00	10.79
2	9.00	0.00	4.47
3	10.79	4.47	0.00

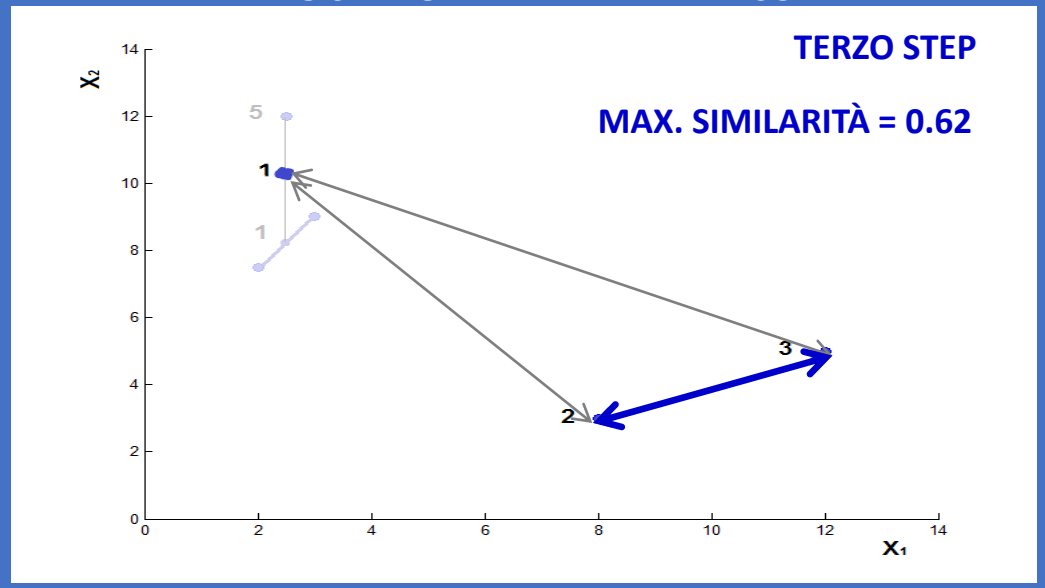
MAX. DIST. = 11.80

MATRICE DELLE SIMILARITÀ

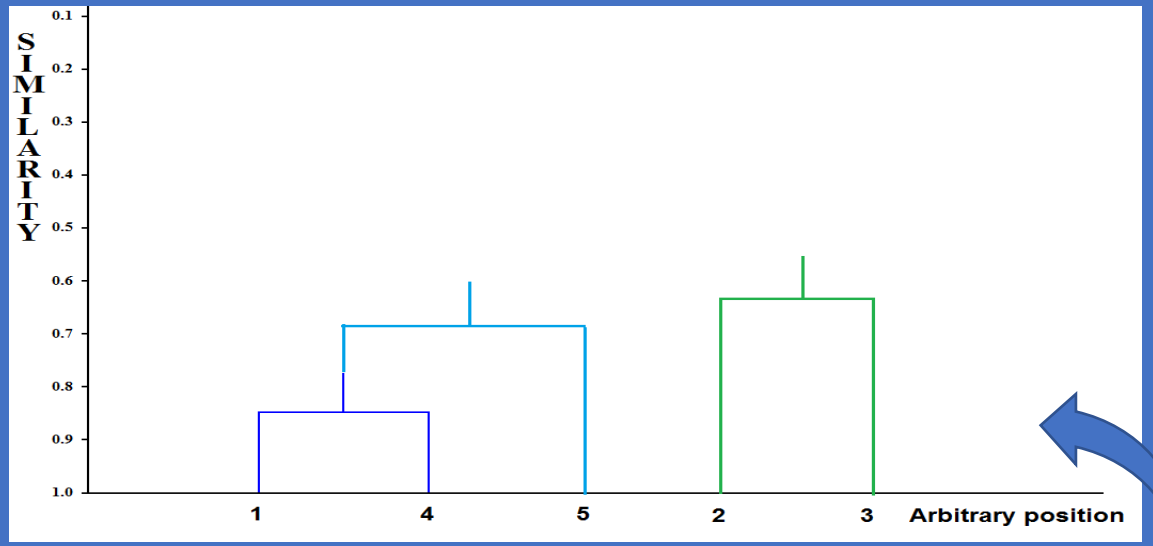
	1	2	3
1	1.00	0.24	0.09
2	0.24	1.00	0.62
3	0.09	0.62	1.00

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

	1	2	3
1	0.00	9.00	10.79
2	9.00	0.00	4.47
3	10.79	4.47	0.00

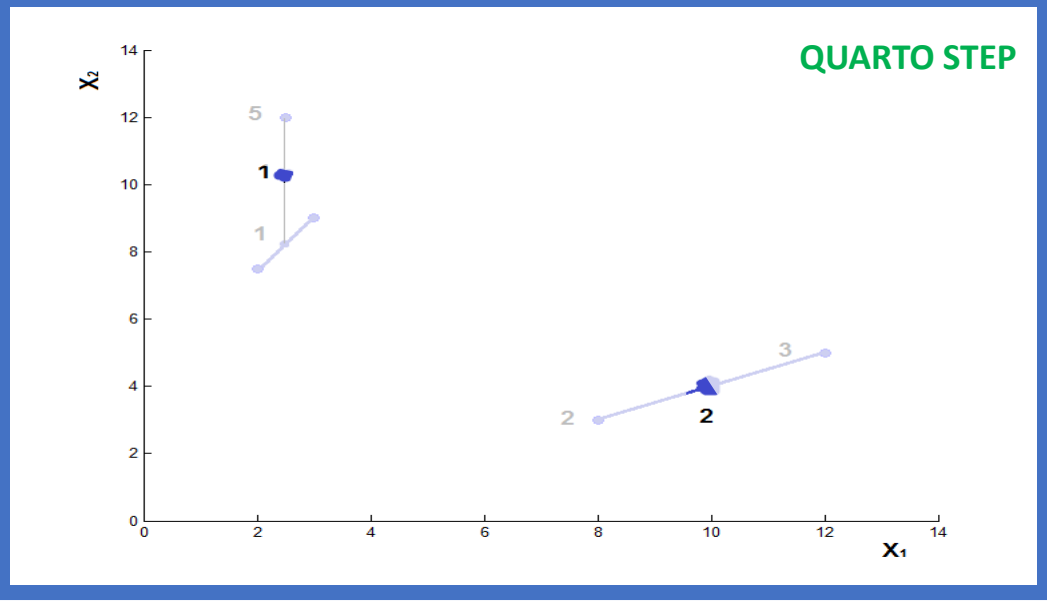
MAX. DIST. = 11.80

MATRICE DELLE SIMILARITÀ

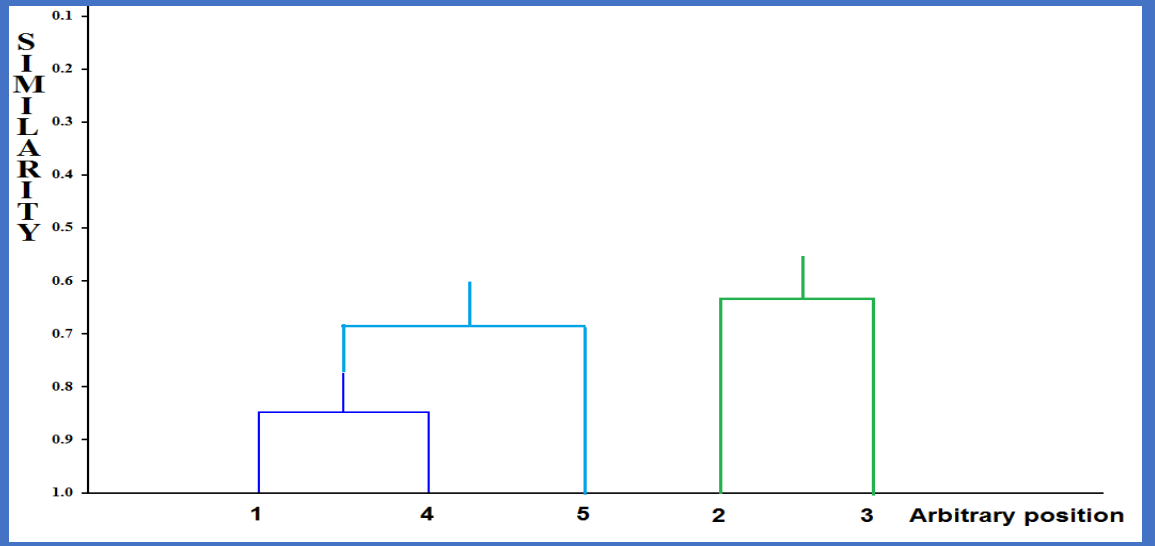
	1	2	3
1	1.00	0.24	0.09
2	0.24	1.00	0.62
3	0.09	0.62	1.00

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA

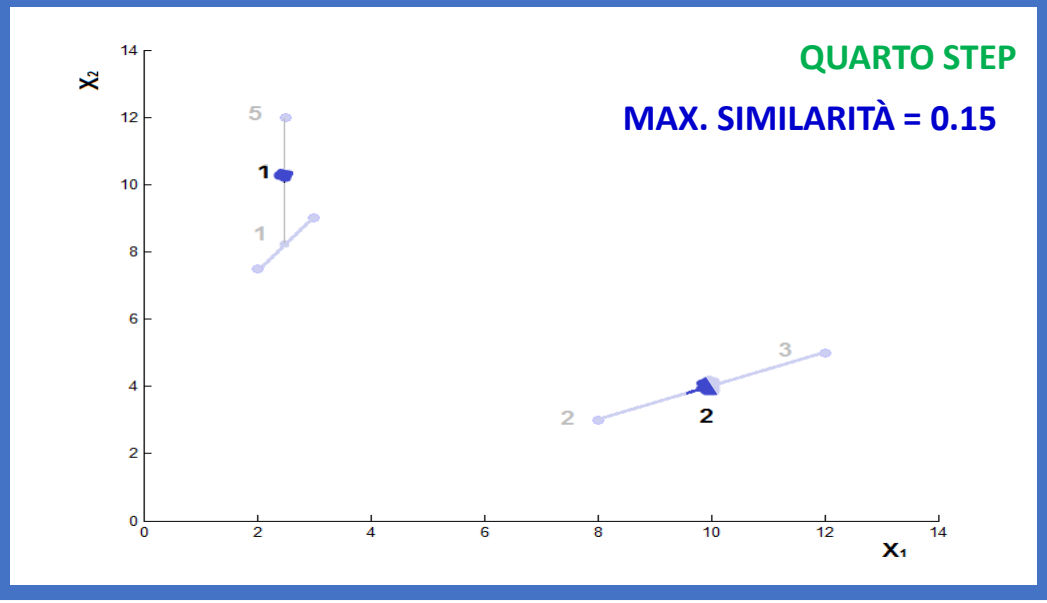


MATRICE DELLE DISTANZE

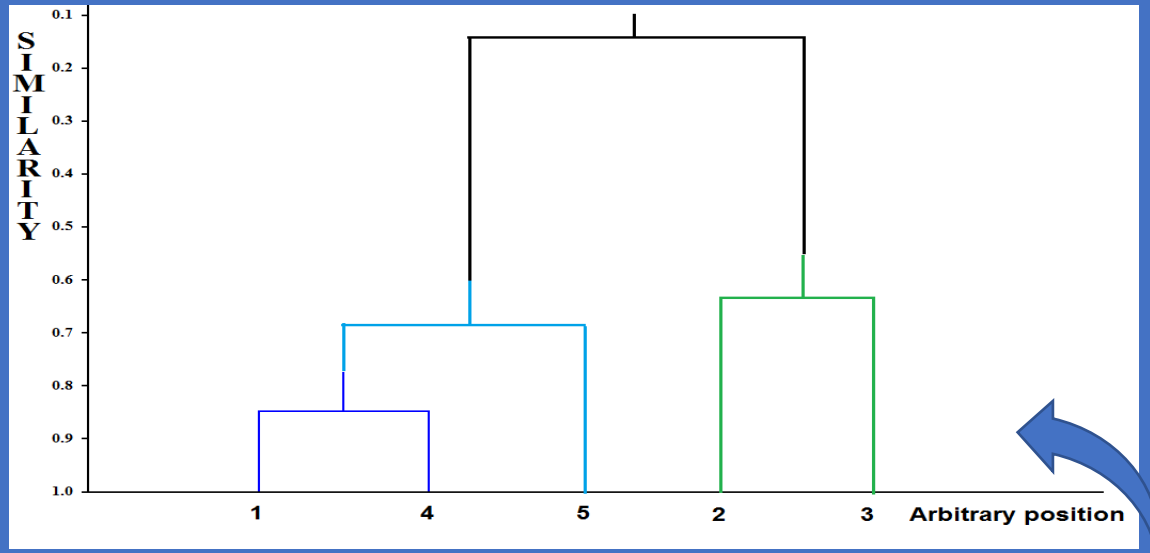
MATRICE DELLE SIMILARITÀ

Esempio: Metodo agglomerativo – legame medio pesato

DATI NELLO SPAZIO DELLE VARIABILI MISURATE



DENDROGRAMMA



MATRICE DELLE DISTANZE

	1	2
1	0.00	9.68
2	9.68	0.00

MAX. DIST. = 11.80

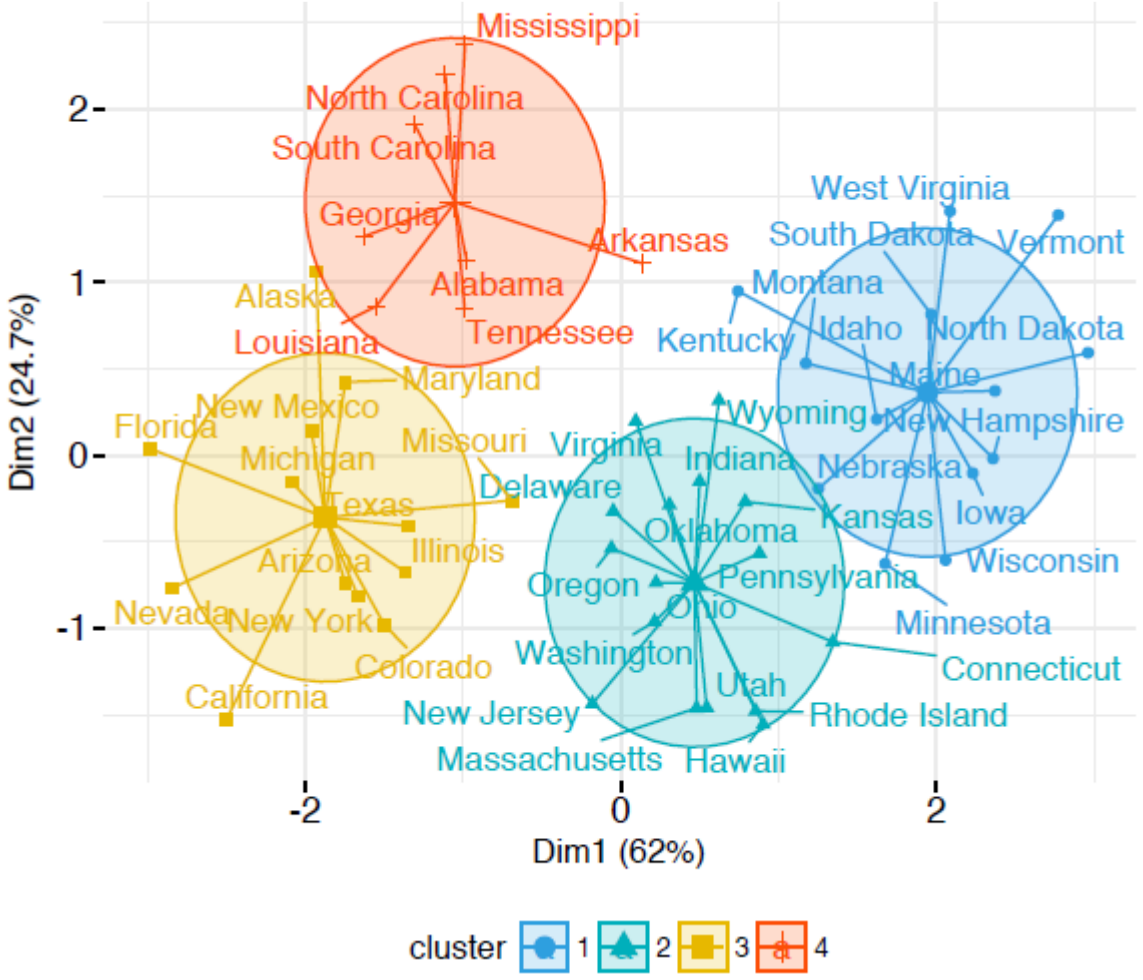
MATRICE DELLE SIMILARITÀ

	1	2
1	1.00	**0.15**
2	**0.15**	1.00

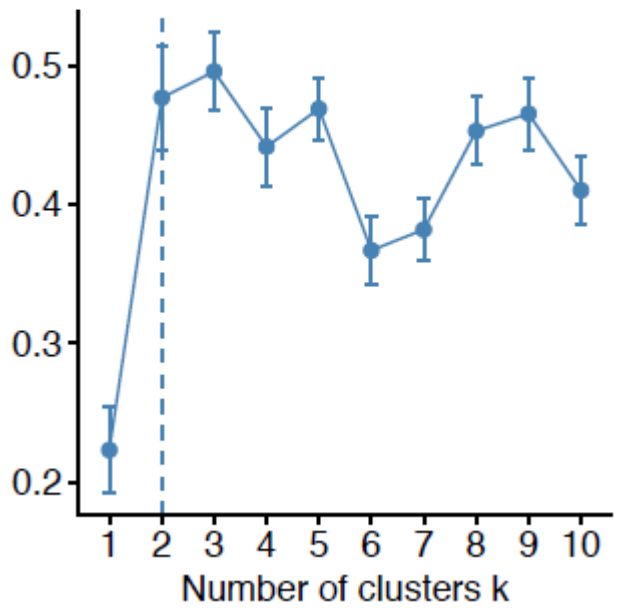
La Cluster Analysis

Tipologie di Cluster Analysis – Partitioning Clustering (ripartizione in *N* gruppi)

Partitioning Clustering



Numero ottimale di N cluster

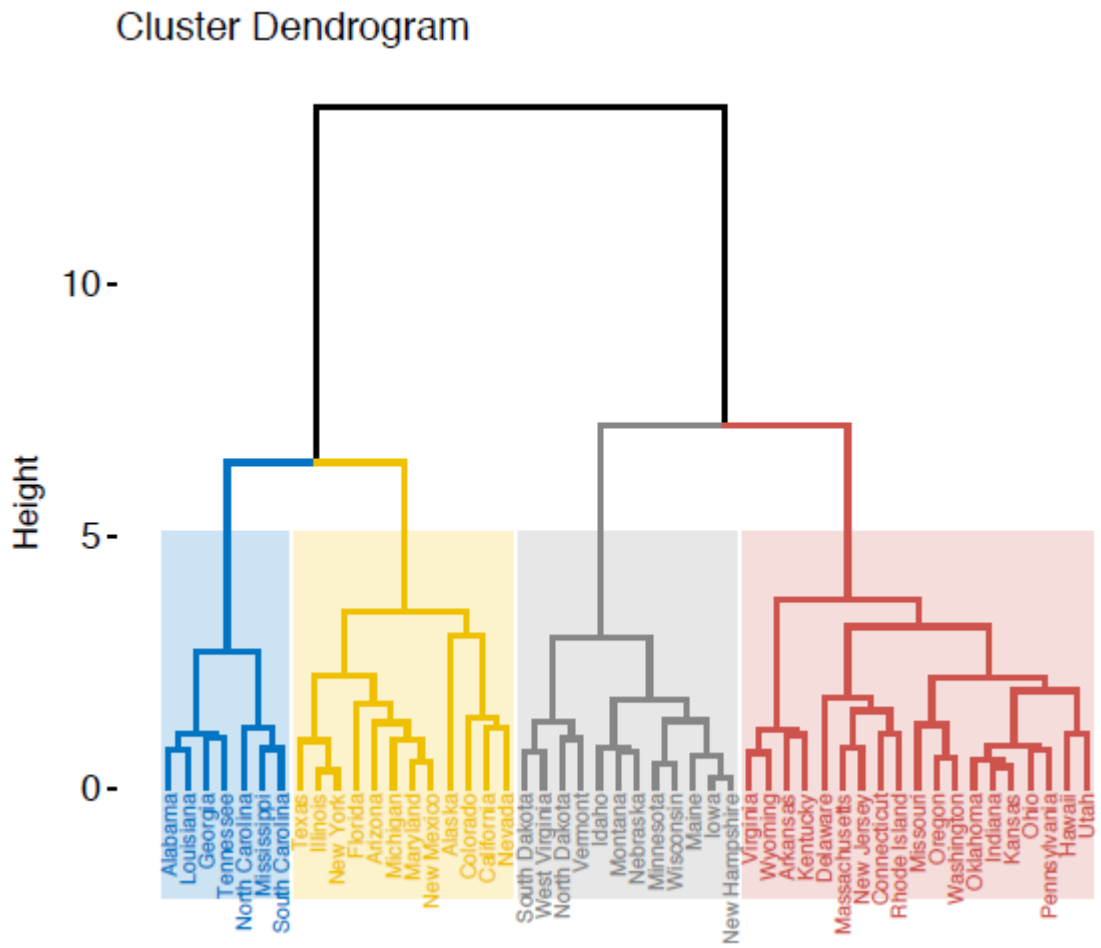


Il numero di N gruppi va stabilito a priori

k-means, CLARA (Clustering Large Applications)

La Cluster Analysis

Tipologie di Cluster Analysis – Hierarchical Clustering (seguono una gerarchia sulla base delle distanze)



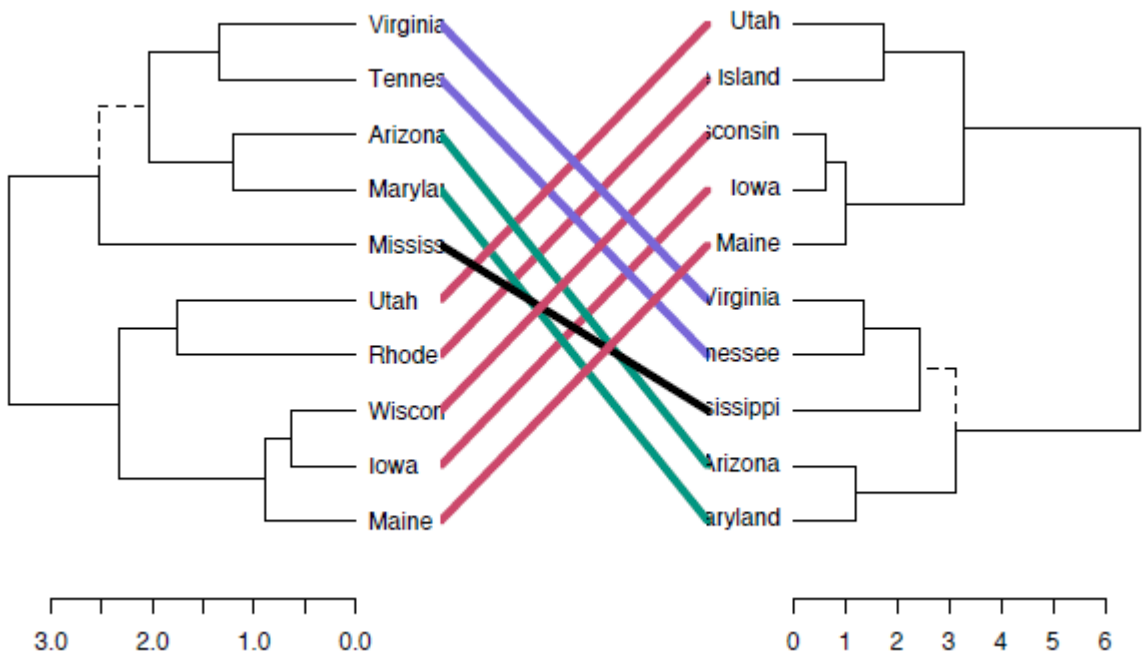
Si ottiene un DENDROGRAMMA

Agglomerativi, divisivi, …

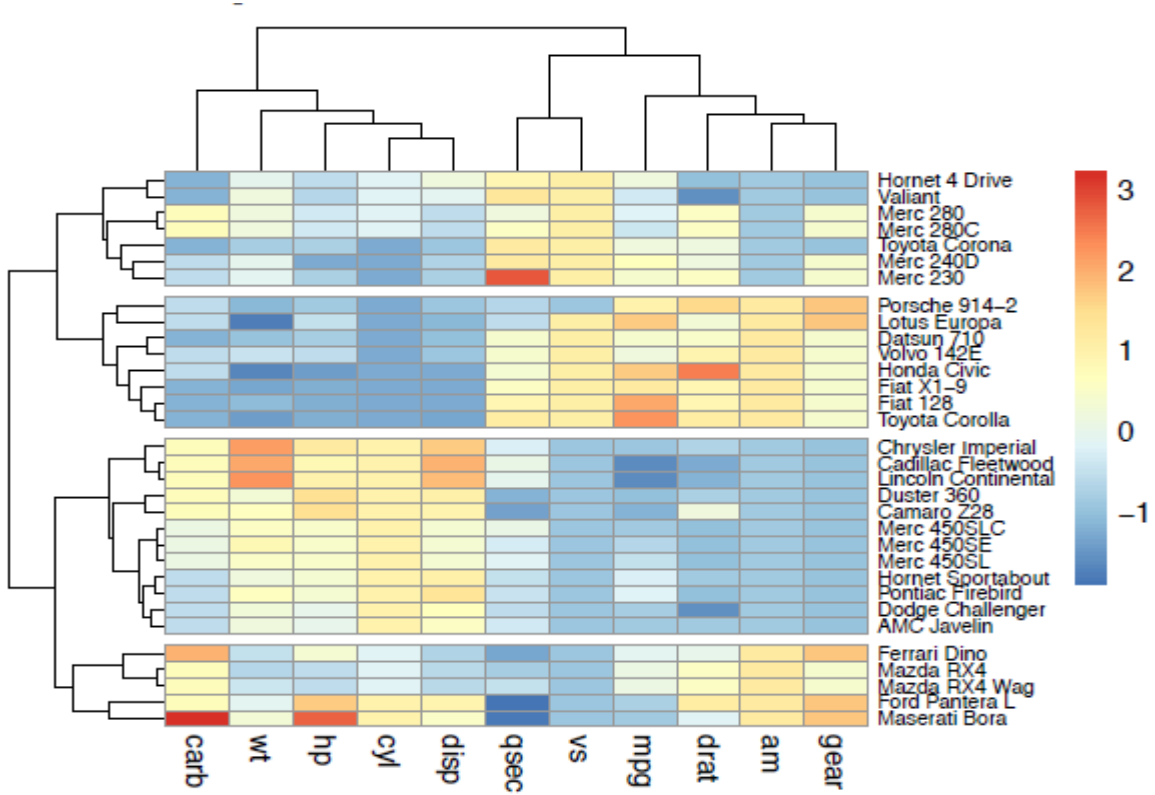
La Cluster Analysis

Tipologie di Cluster Analysis – Hierarchical Clustering (seguono una gerarchia sulla base delle distanze)

2 dendrogrammi



Heatmap

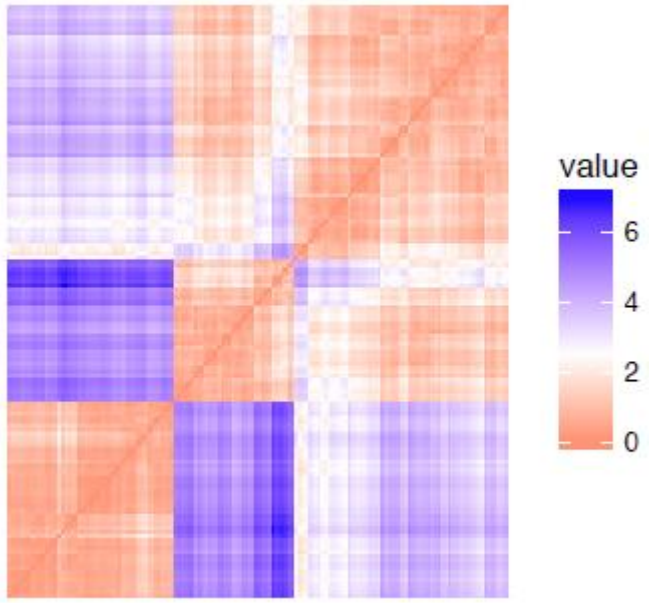


SI POSSONO CONFRONTARE I RISULTATI DI DENDROGRAMMI OTTENUTI CON METODI DIVERSI

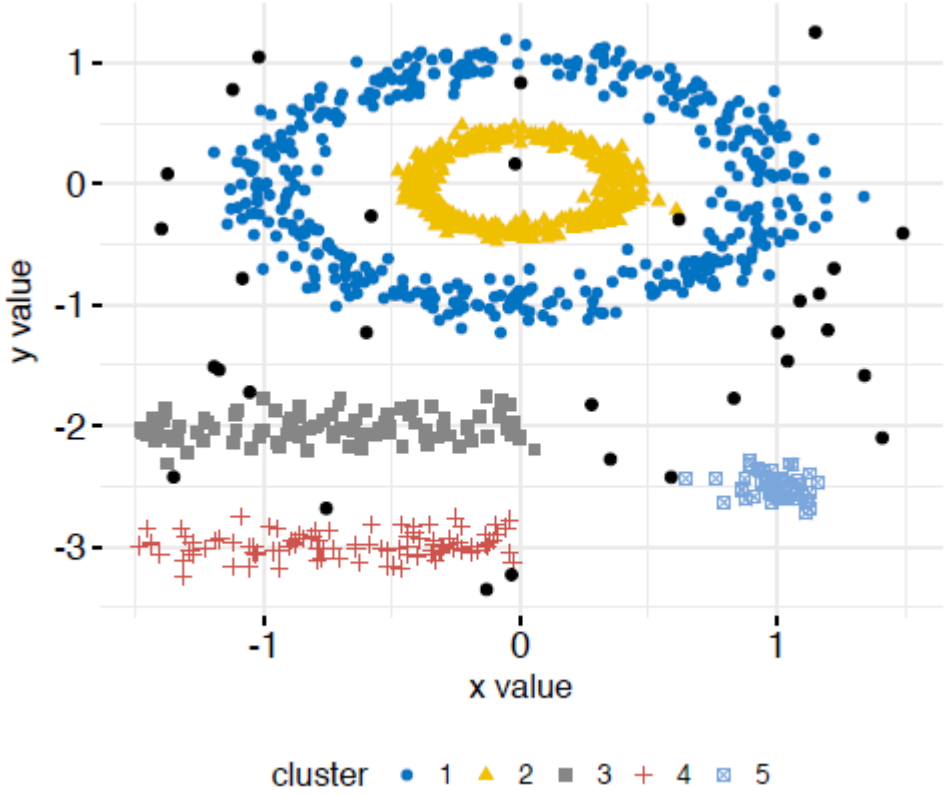
La Cluster Analysis

Altre tipologie di Cluster Analysis

Cluster di tendenza

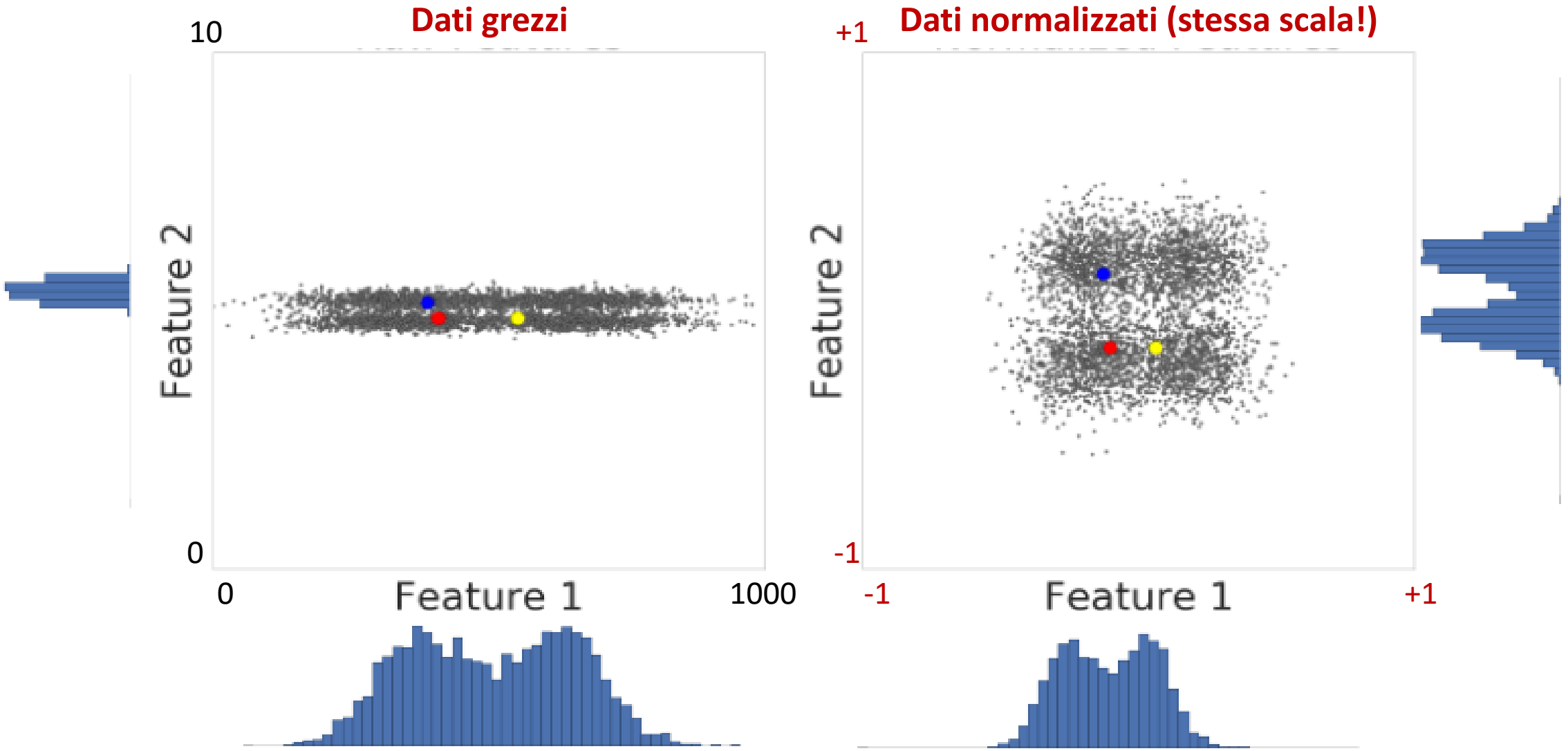


Cluster di densità



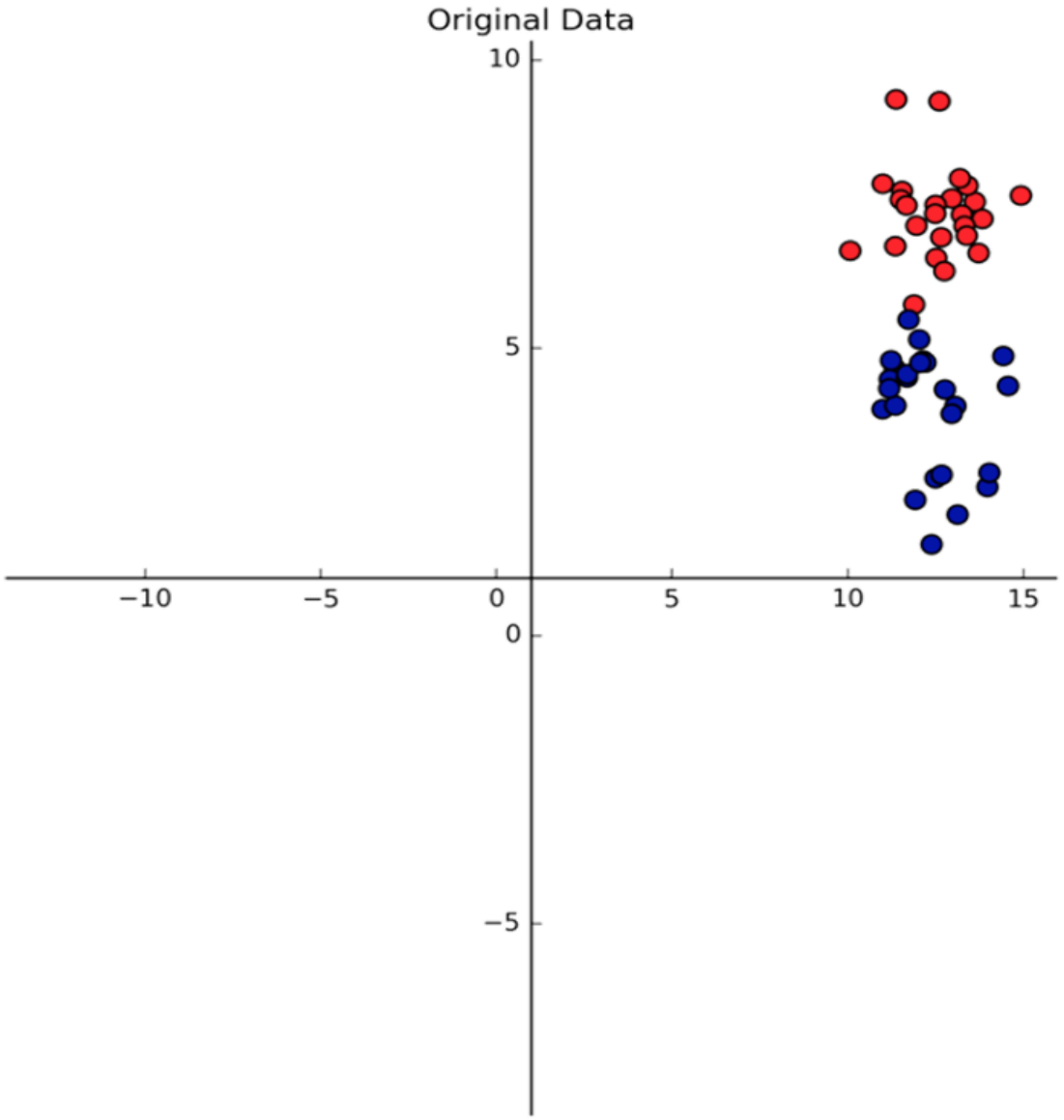
La Cluster Analysis

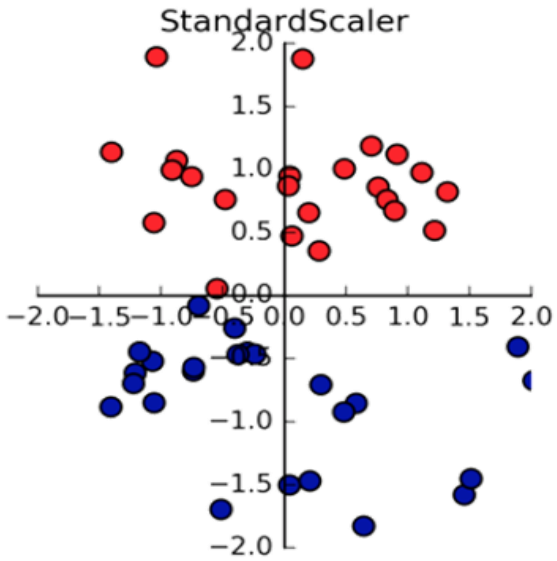
Prima di effettuare clustering: Normalizzazione dei Dati!

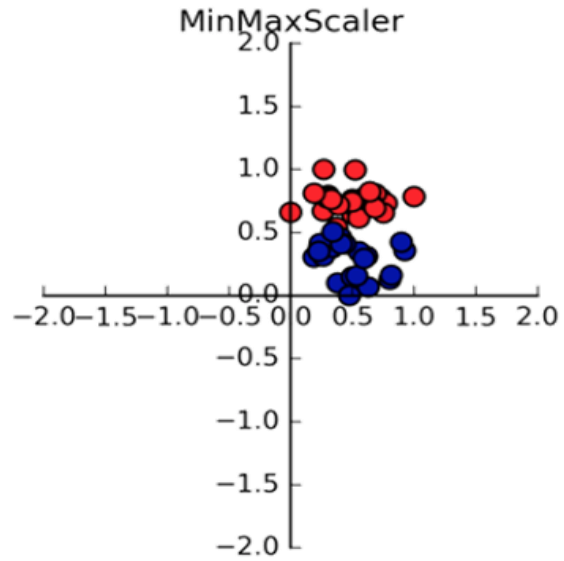


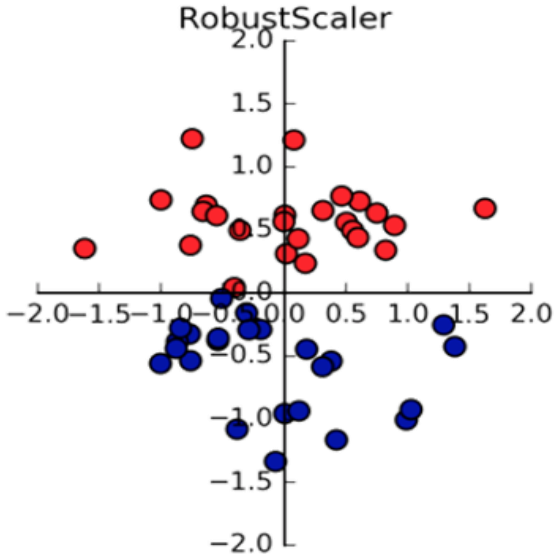
La Cluster Analysis

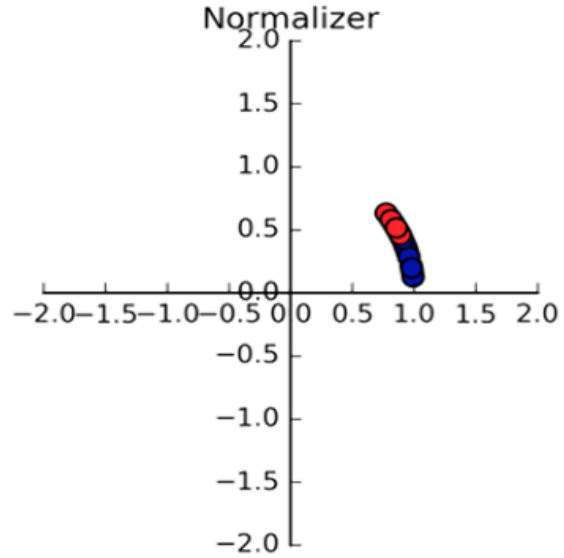
Prima di effettuare clustering: Normalizzazione dei Dati!







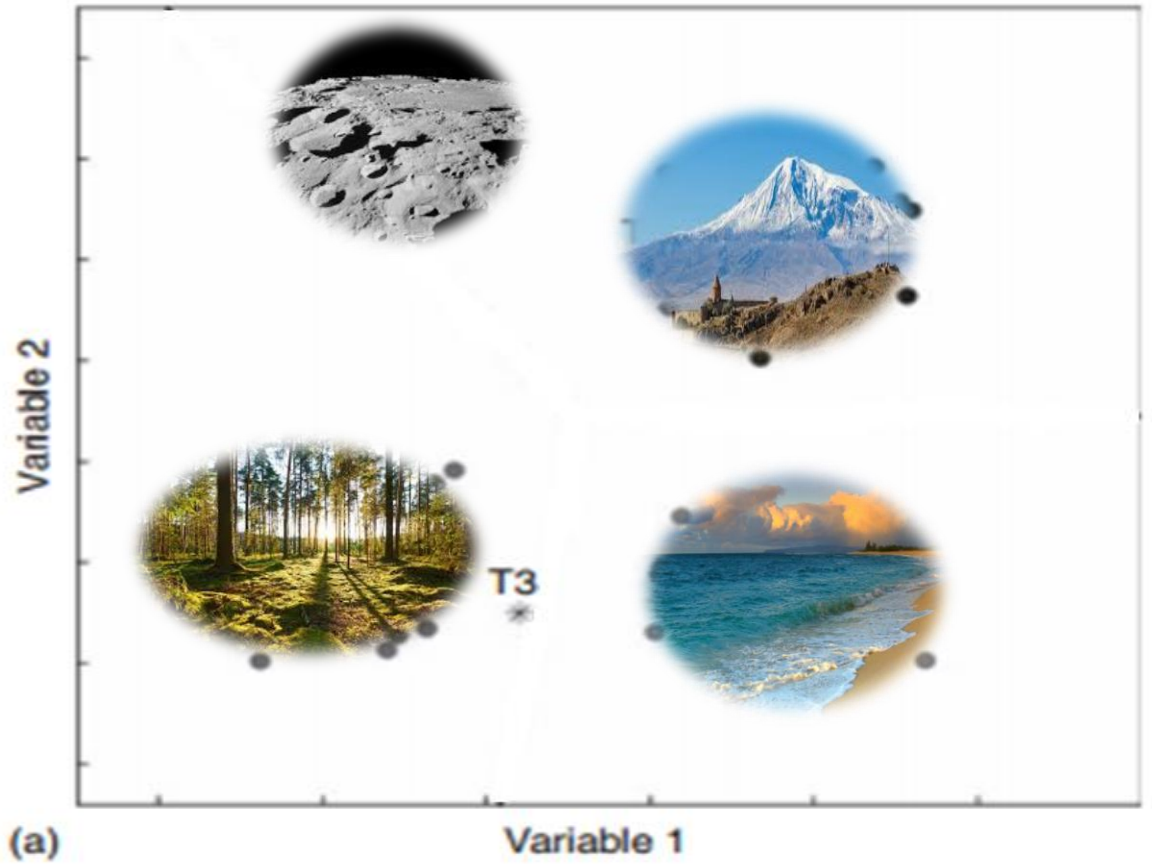




L’analisi multivariata di classificazione e modellamento

L’analisi multivariata di classificazione e modellamento

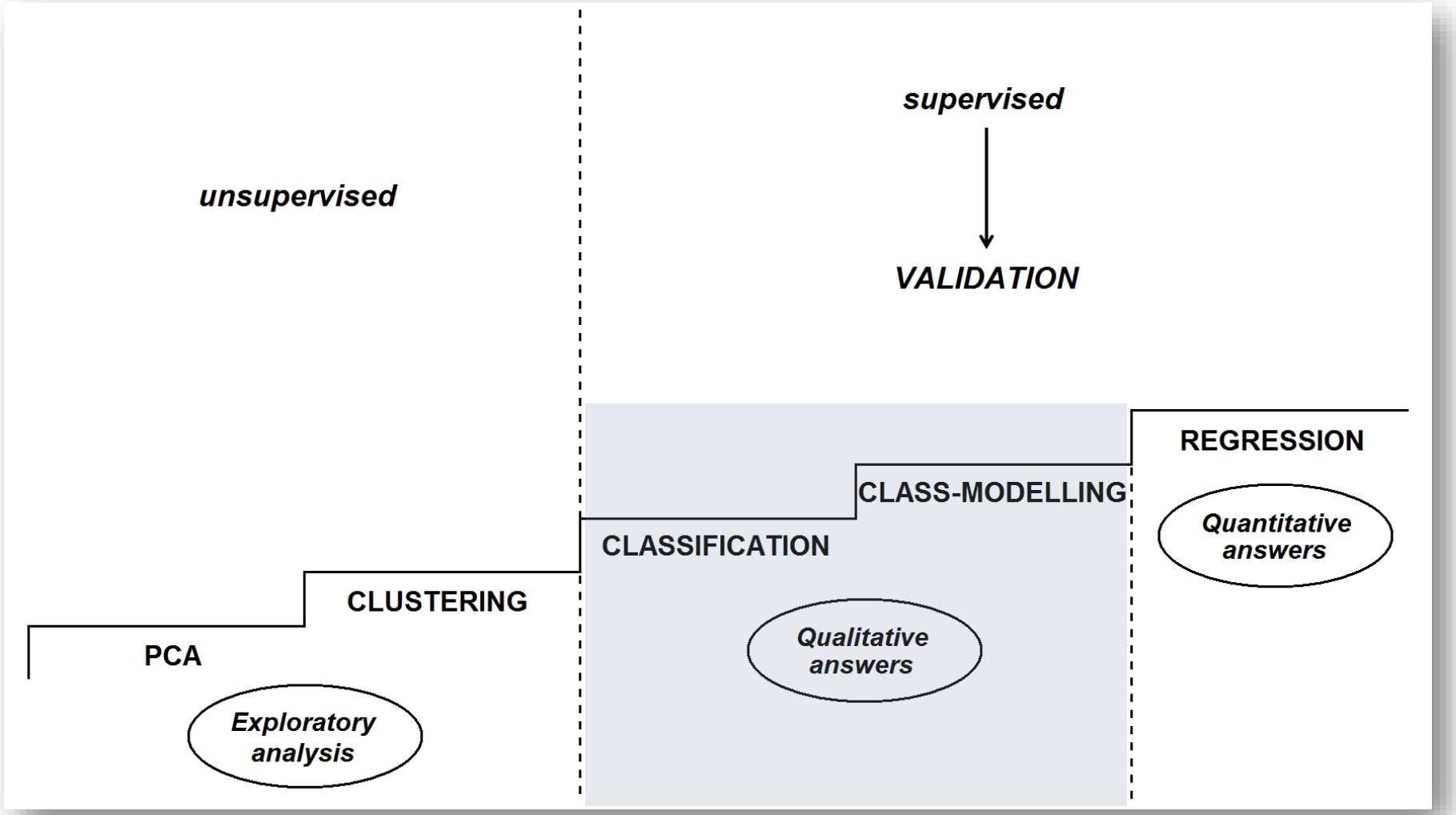
UNSUPERVISED (ESPLORATIVI)



- **PCA** (Principal Component Analysis)
- **CA** (Cluster Analysis)

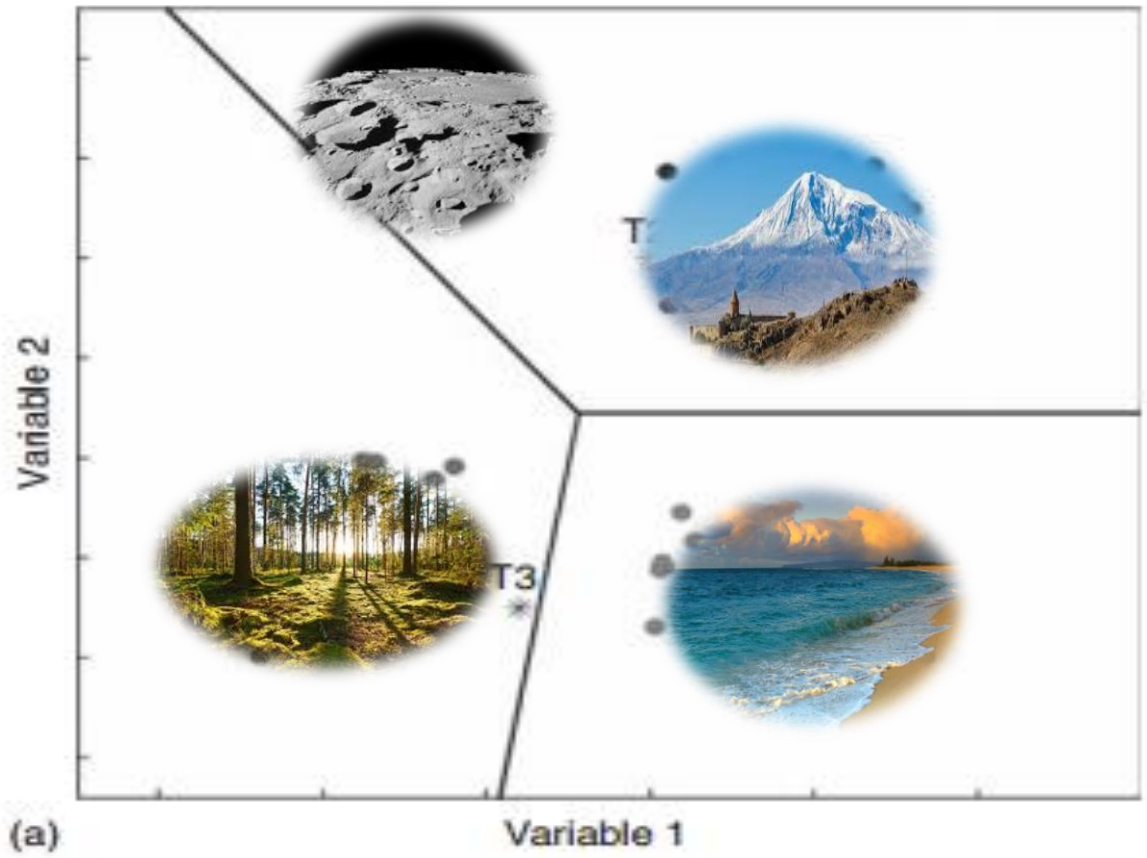
L’analisi multivariata di classificazione e modellamento

Metodi



L’analisi multivariata di classificazione e modellamento

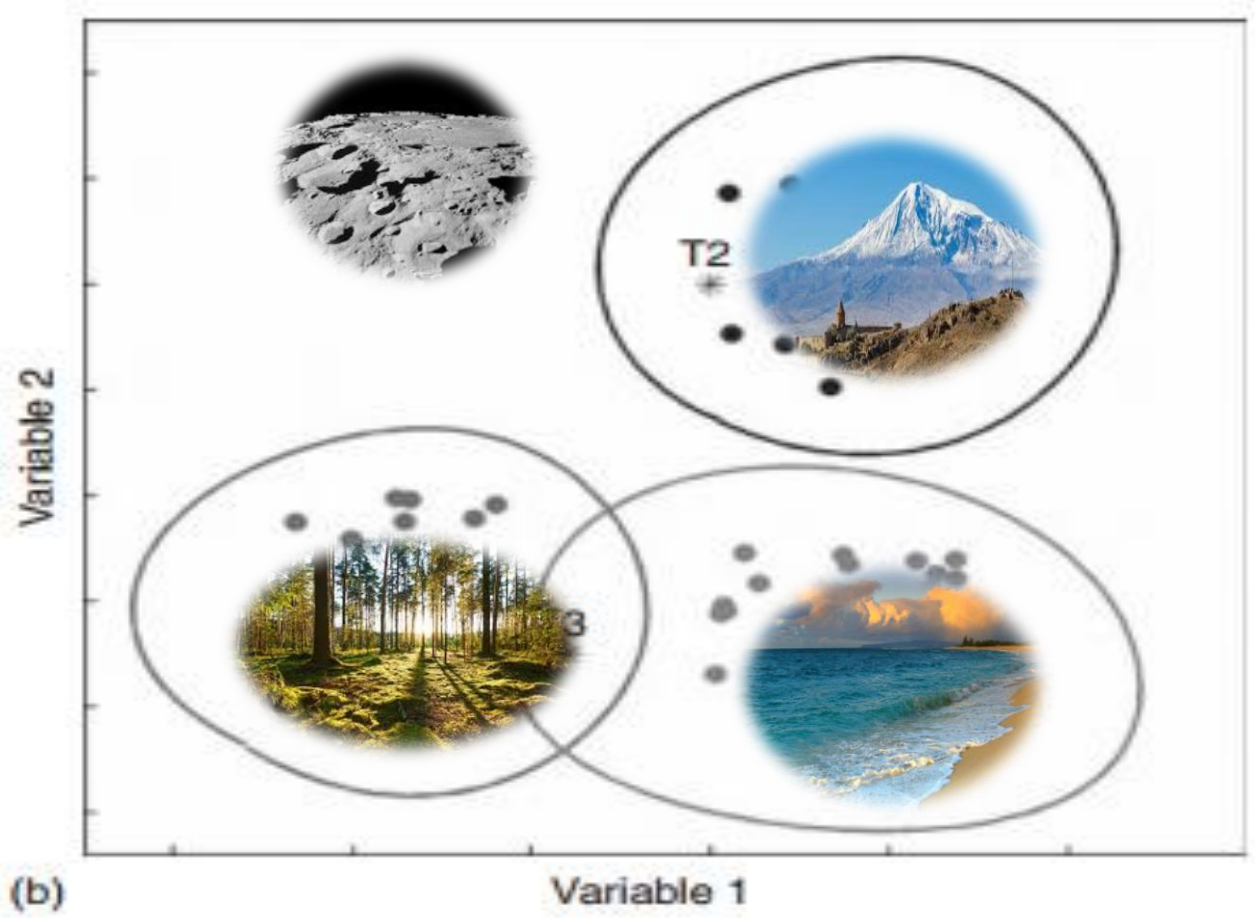
SUPERVISED (CLASSIFICAZIONE)



- **LDA** (Linear Discriminant Analysis)
- **SVM** (Support Vector Machine)

L’analisi multivariata di classificazione e modellamento

SUPERVISED (MODELLAMENTO)



- **SIMCA** (Soft-Indipendent Models of Class Analogy)
- **UNEQ** (Unequal Dispersed Class)

L’analisi multivariata di classificazione e modellamento

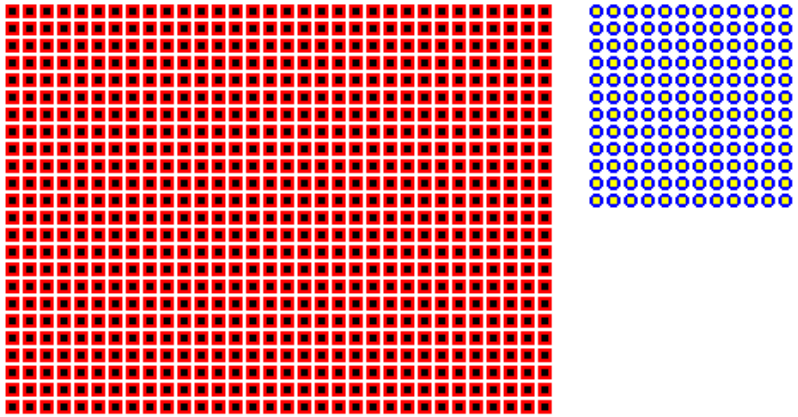
CAMPIONAMENTO

Se non costruito correttamente, è l’approccio più rischioso! I dati vanno selezionati nel modo più rappresentativo possibile.

L’analisi multivariata di classificazione e modellamento

CAMPIONAMENTO

Popolazione



Probabilità a-priori p(g)

Campione statistico - campionamento

Proporzionale

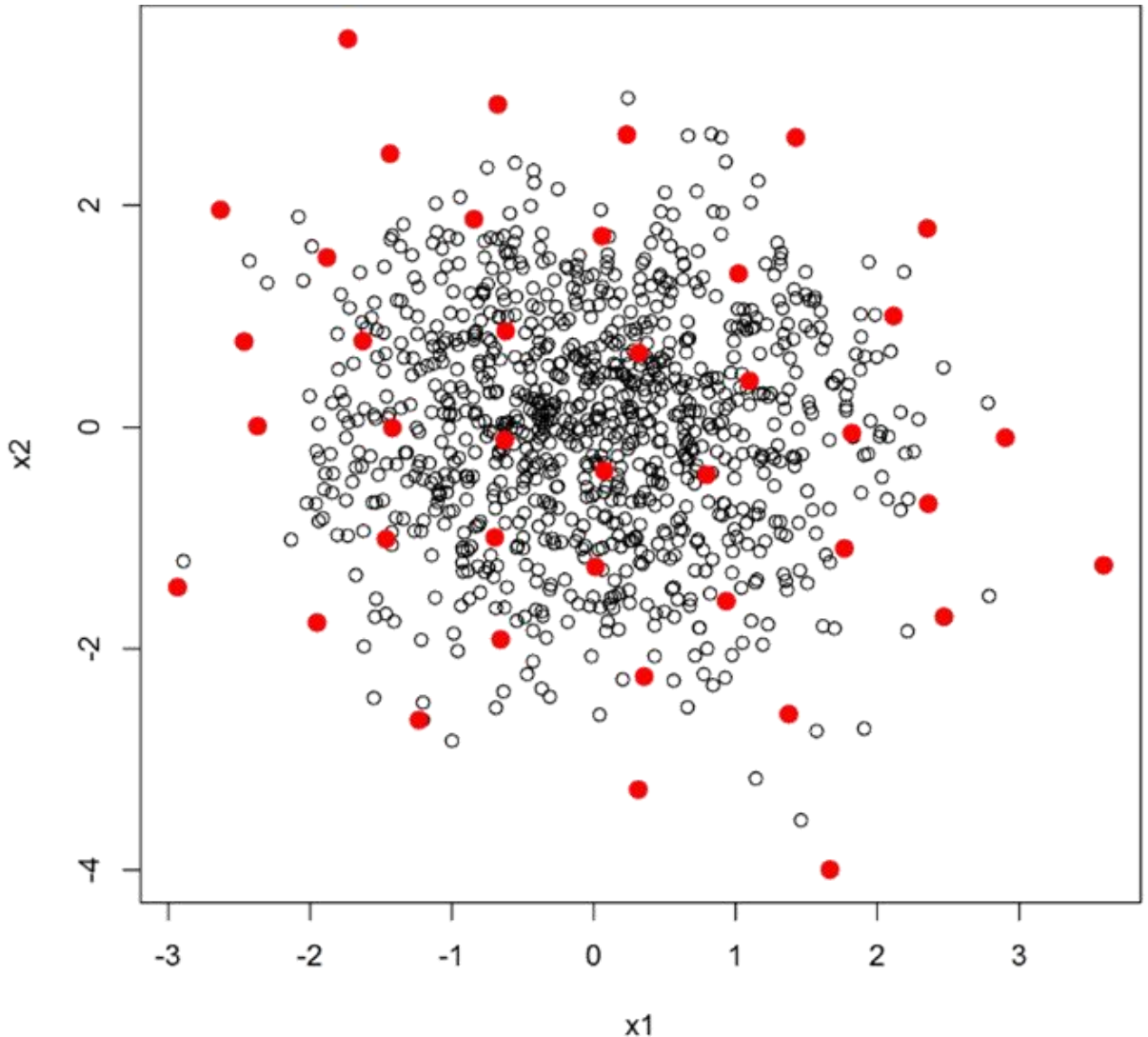
Di dimensioni predefinite

Limitato (per motivi pratici)

L’analisi multivariata di classificazione e modellamento

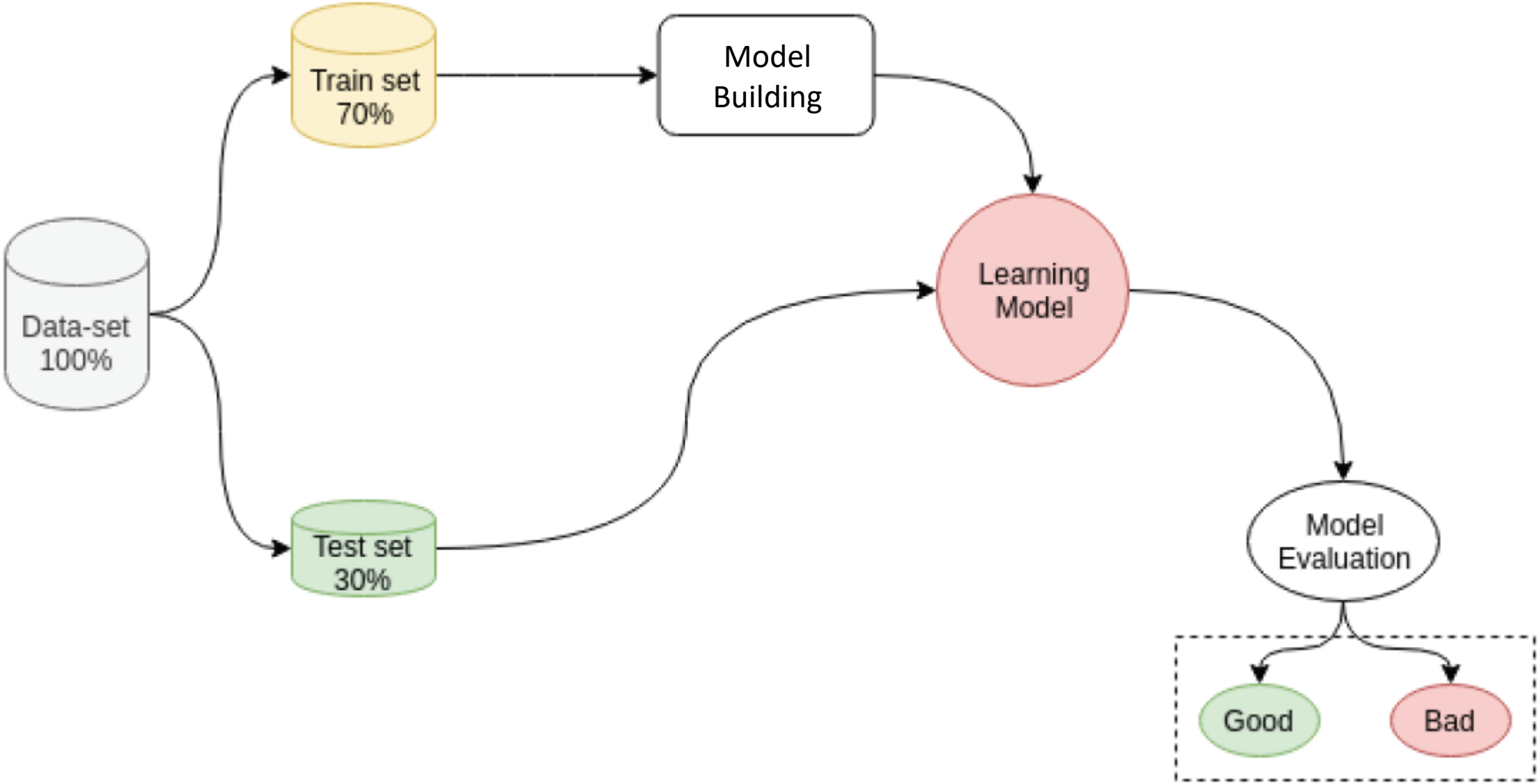
CAMPIONAMENTO

Kennard & Stone algorithm



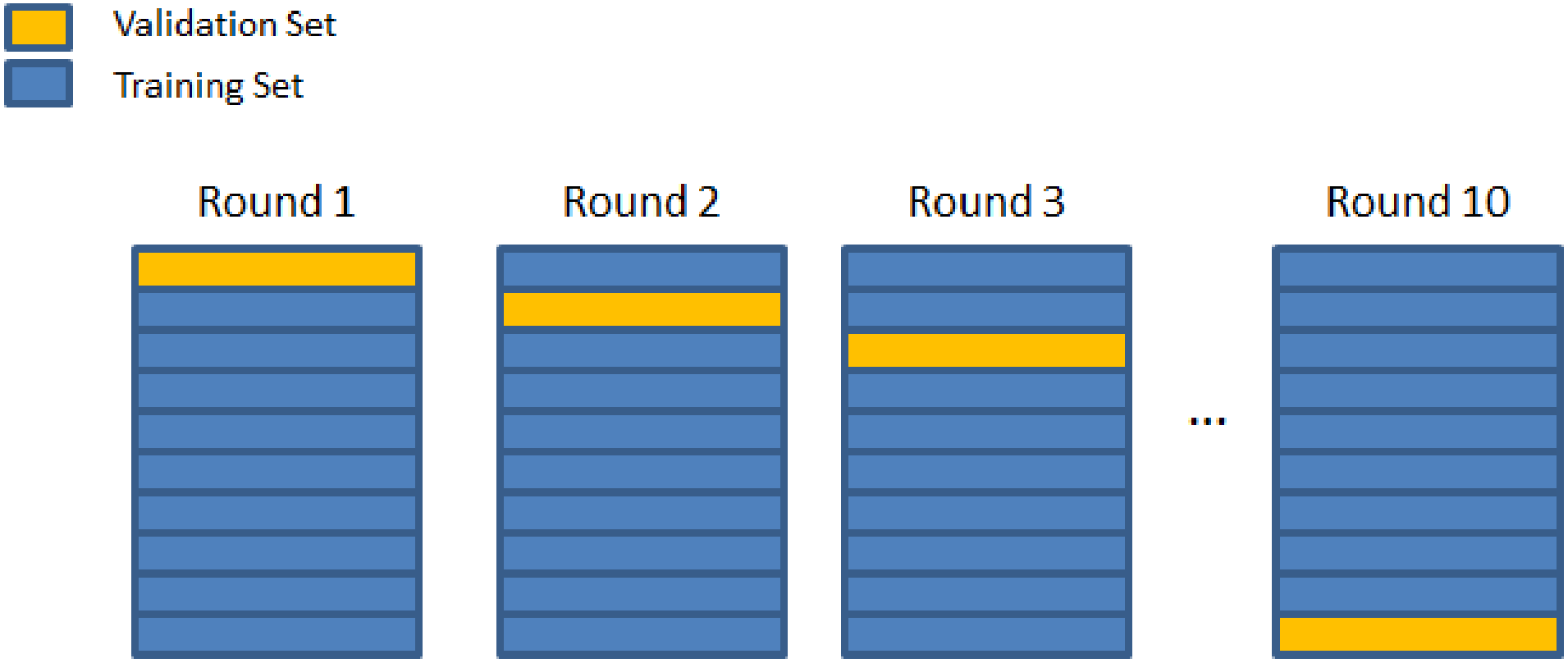
L’analisi multivariata di classificazione e modellamento

Validazione e Cross-validazione



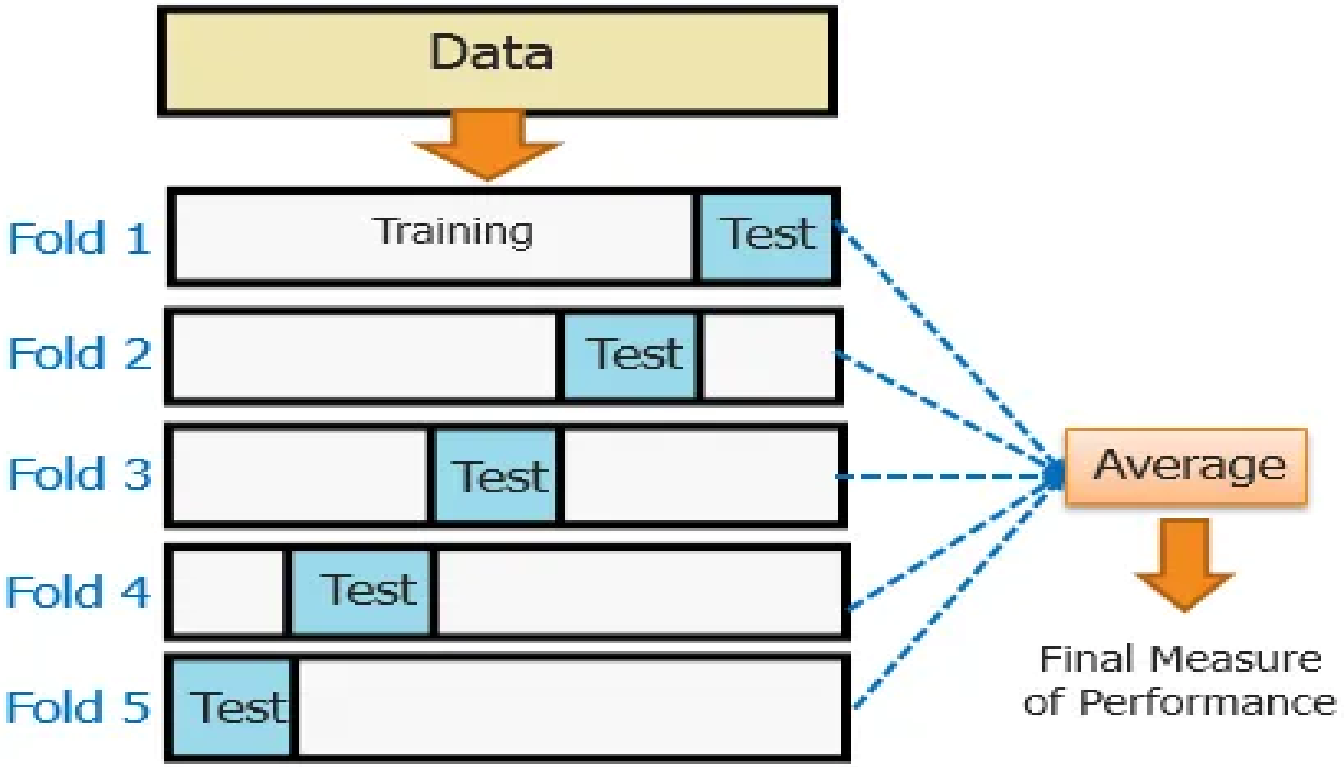
L’analisi multivariata di classificazione e modellamento

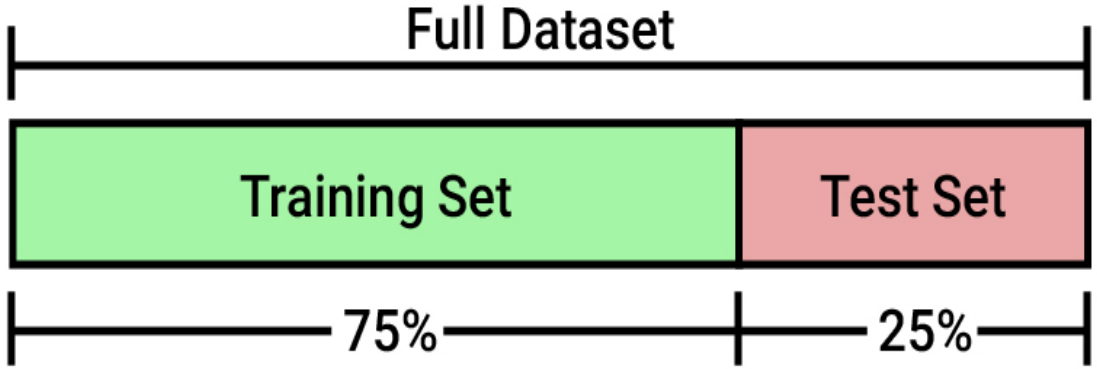
Cross-validazione – Leave-One-Out

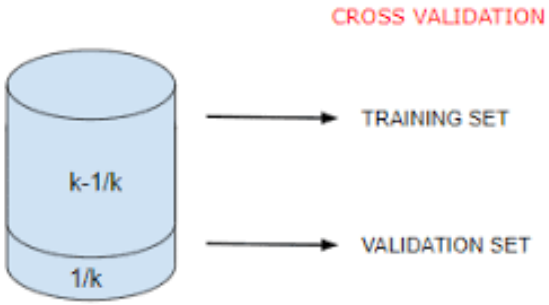


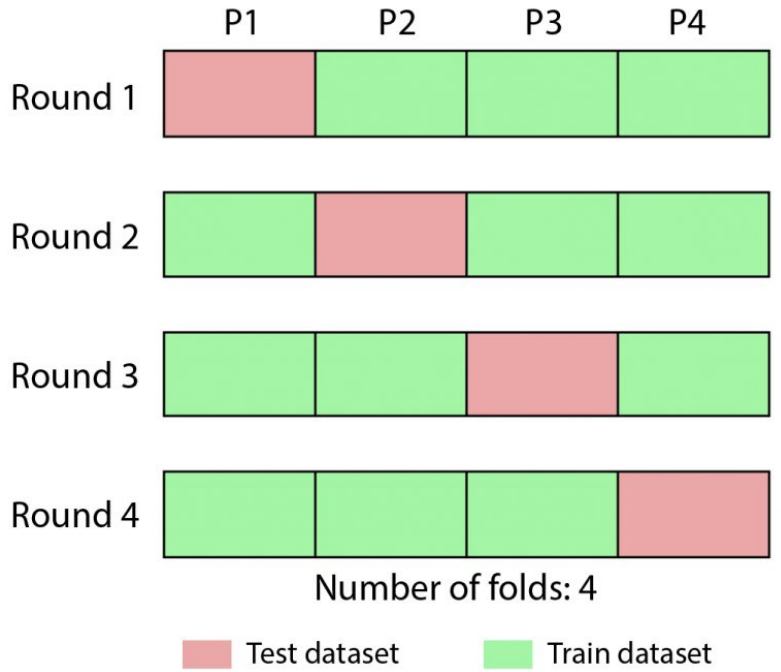
L’analisi multivariata di classificazione e modellamento

Cross-validazione – K-fold









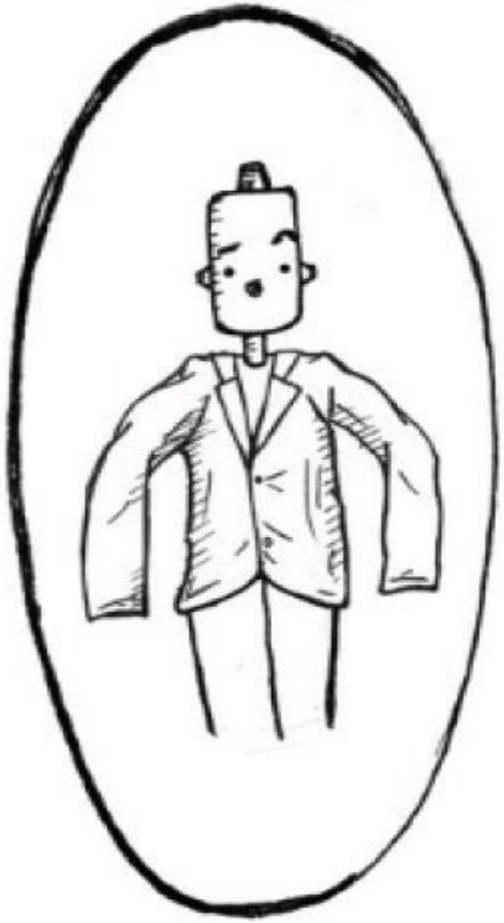
L’analisi multivariata di classificazione e modellamento

Cross-validazione – K-fold

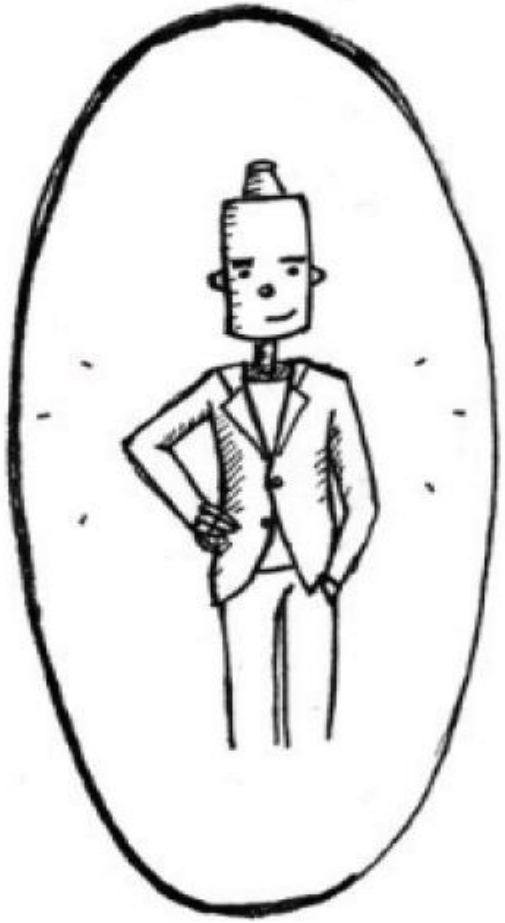
	Venetian Blinds	Contiguous Blocks	Random Subsets
Test sample selection scheme			

L’analisi multivariata di classificazione e modellamento

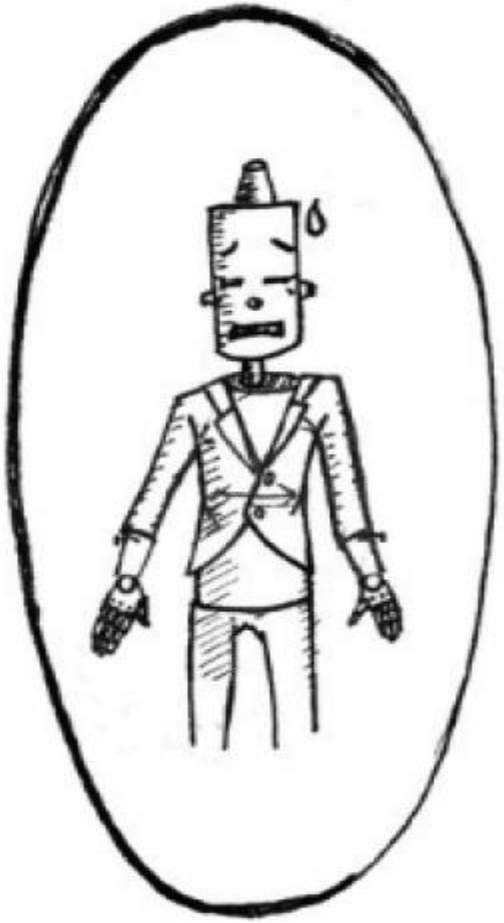
Fitting



Underfitting



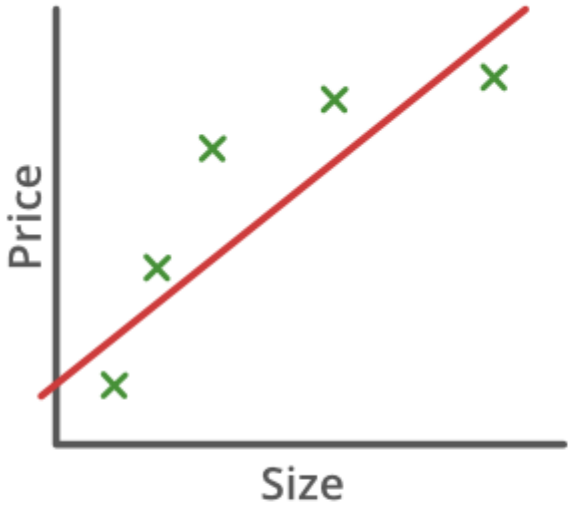
Fitting



Overfitting

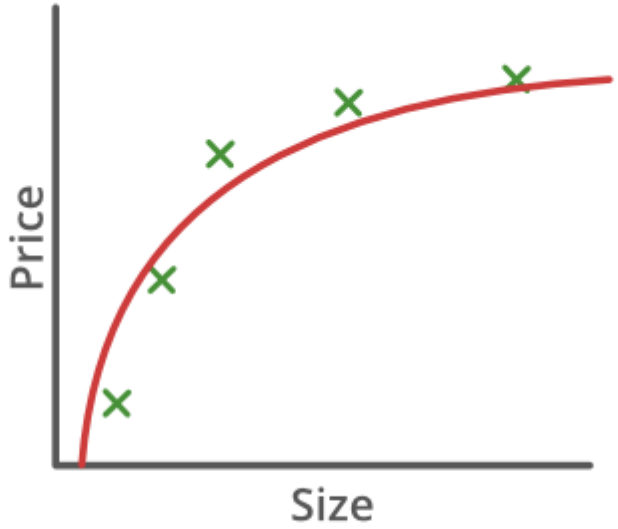
L’analisi multivariata di classificazione e modellamento

Selezione delle variabili (+ validazione)



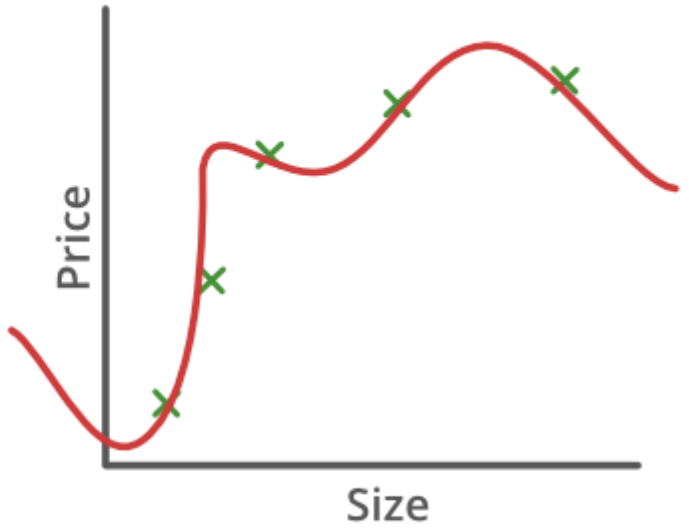
$\theta_0 + \theta_1 x$

Underfitting



$\theta_0 + \theta_1 x + \theta_2 x^2$

Fitting

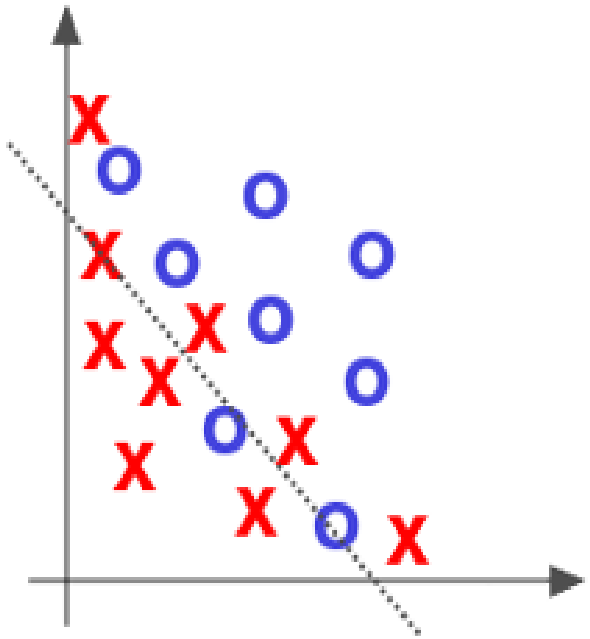


$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_2 x^2 + \theta_2 x^2$

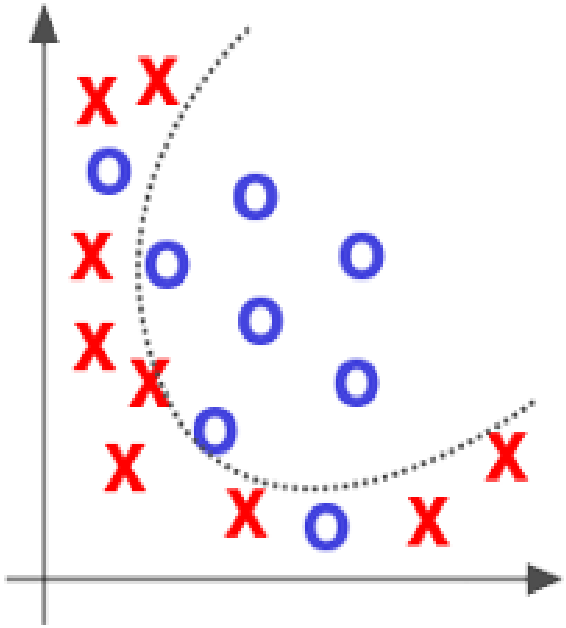
Overfitting

L’analisi multivariata di classificazione e modellamento

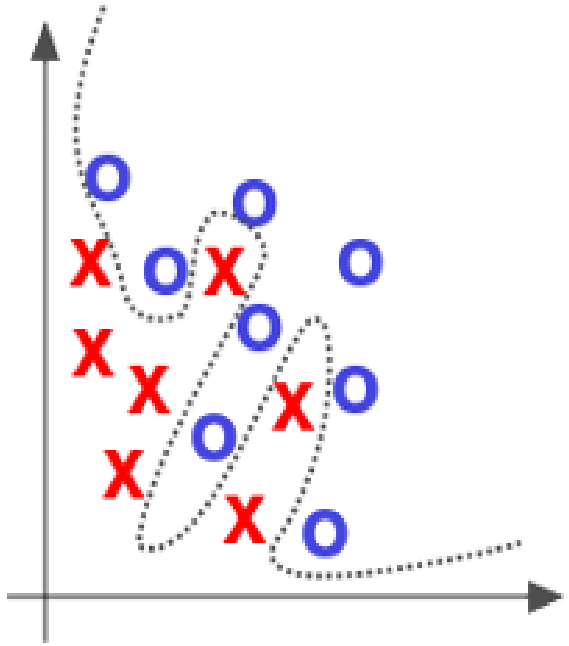
Selezione delle variabili (+ validazione)



Underfitting

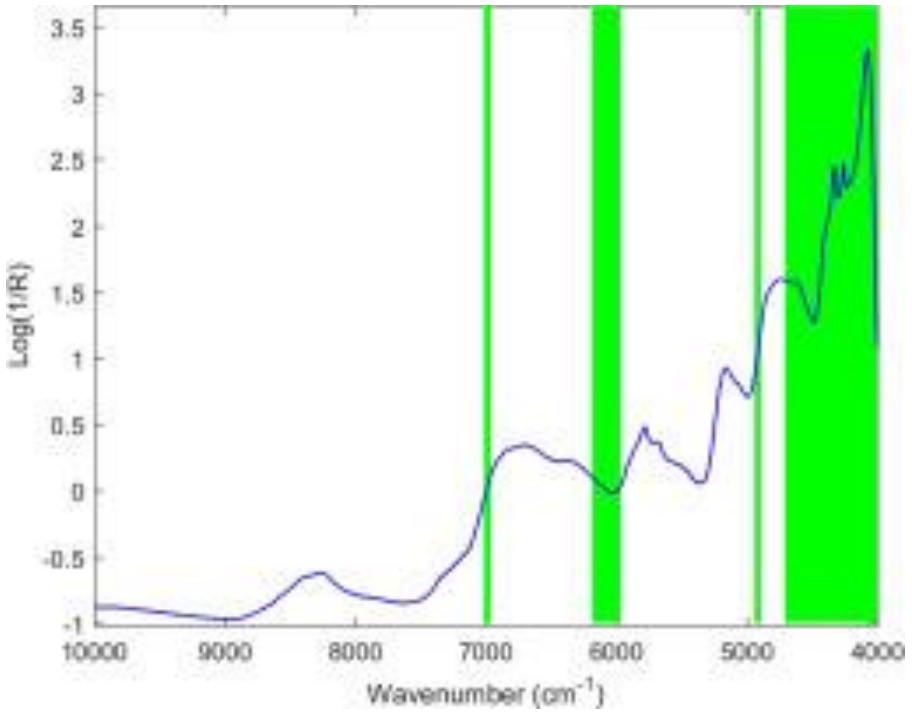


Fitting

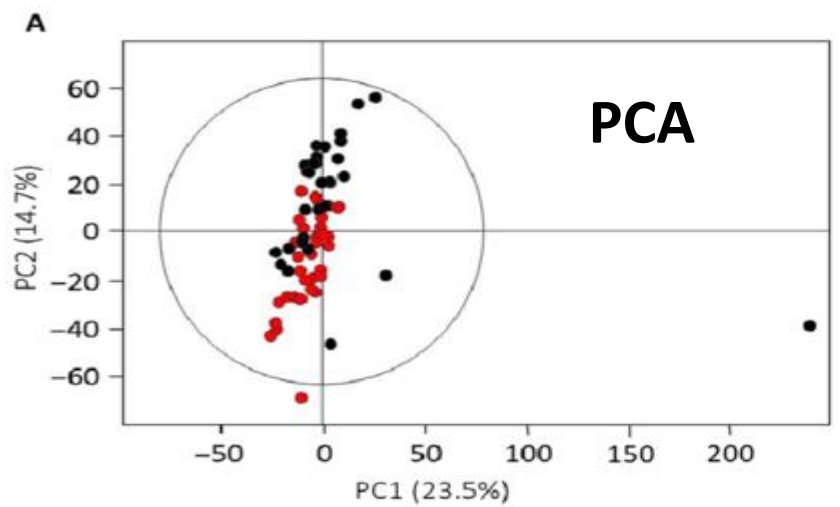


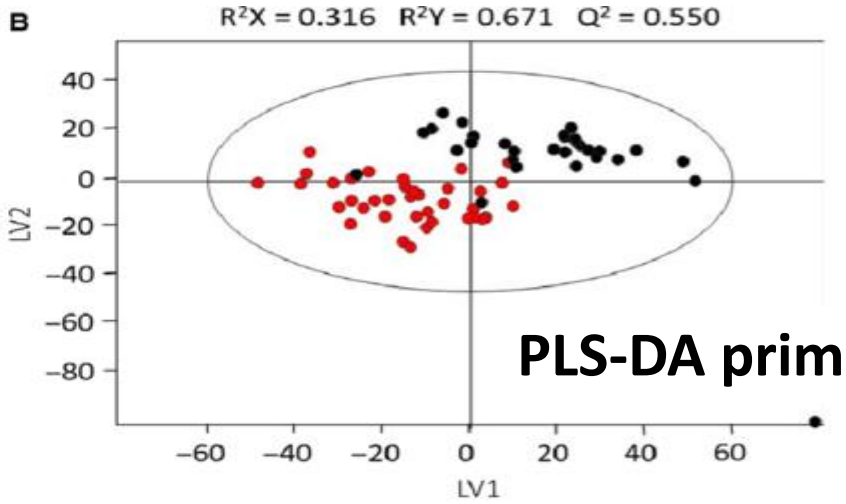
Overfitting

Variables/Features Selection (FS)

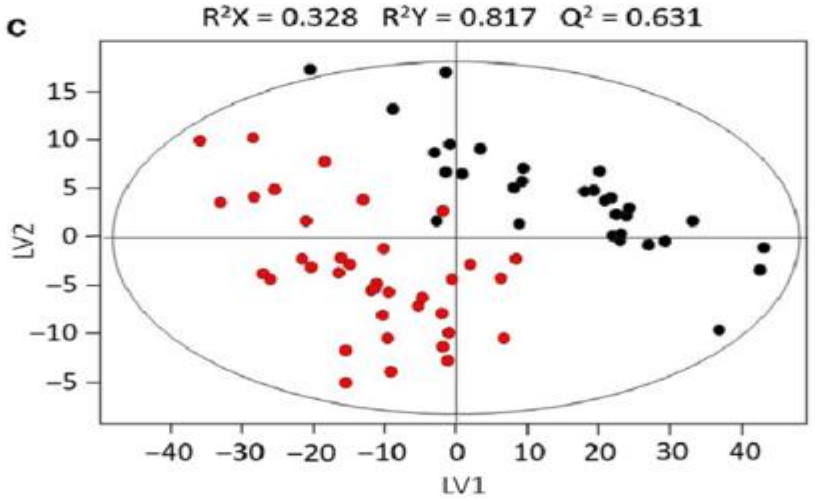


Variables/Features Selection (FS)



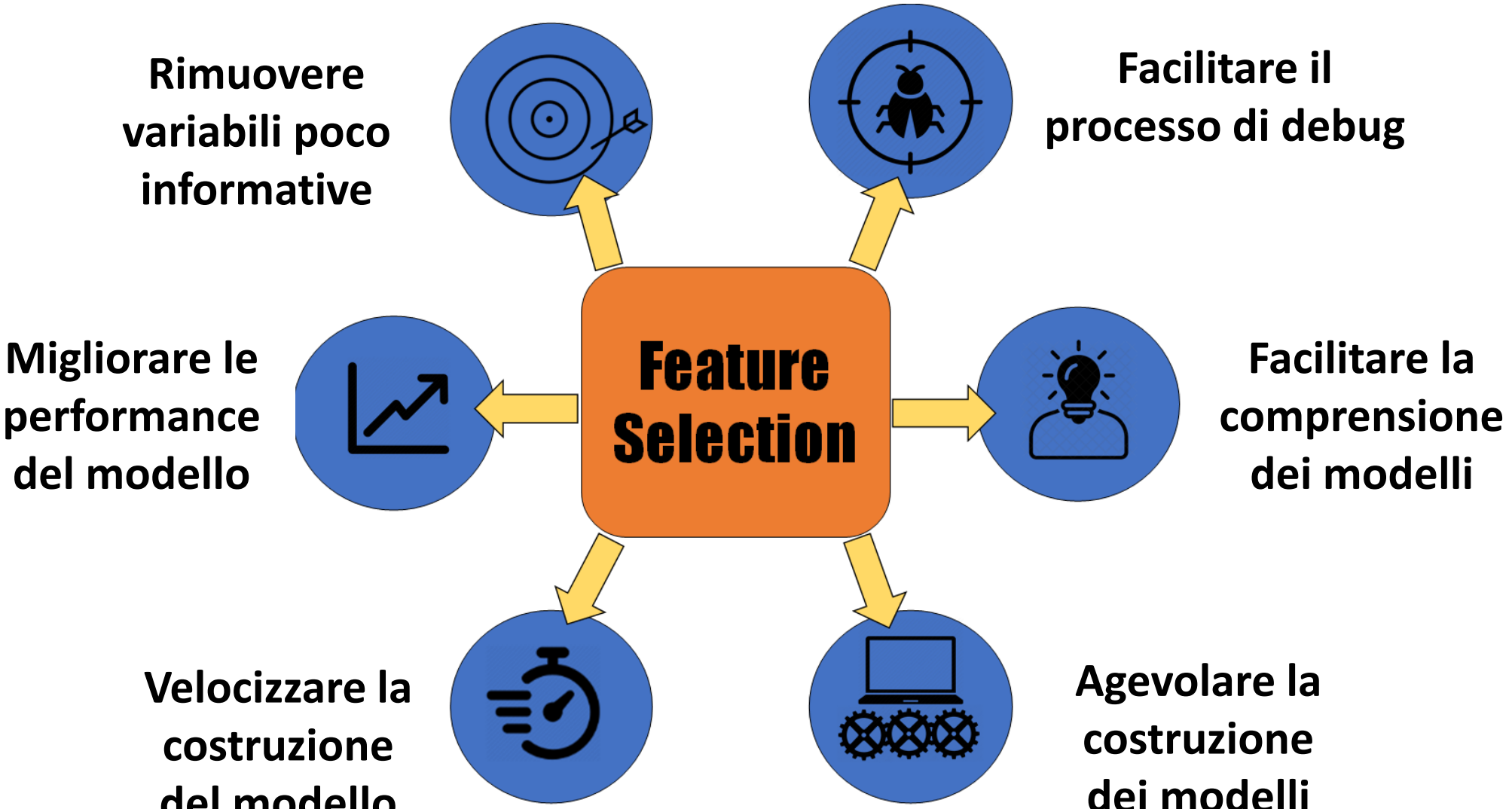


PLS-DA prima di FS

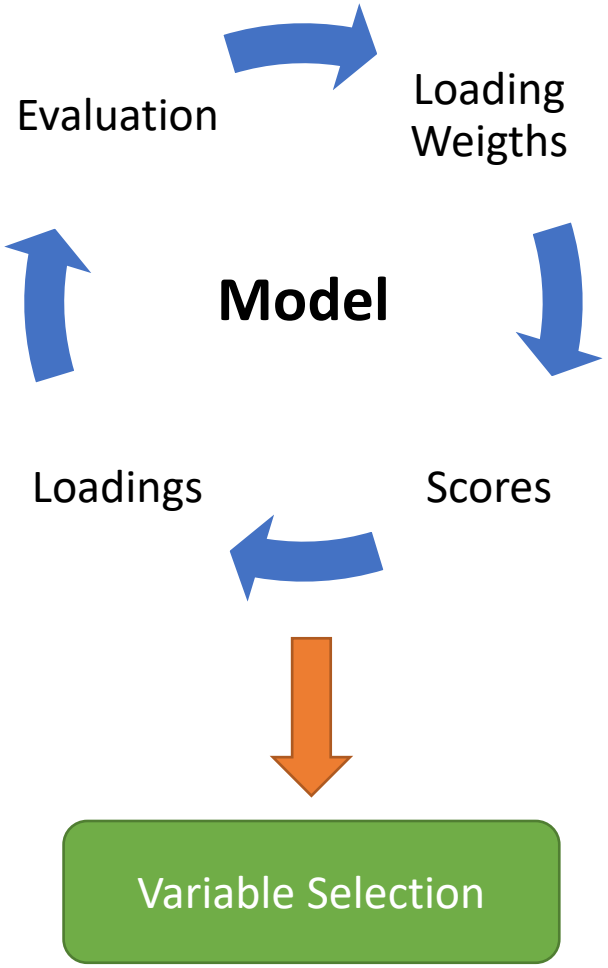


PLS-DA dopo FS

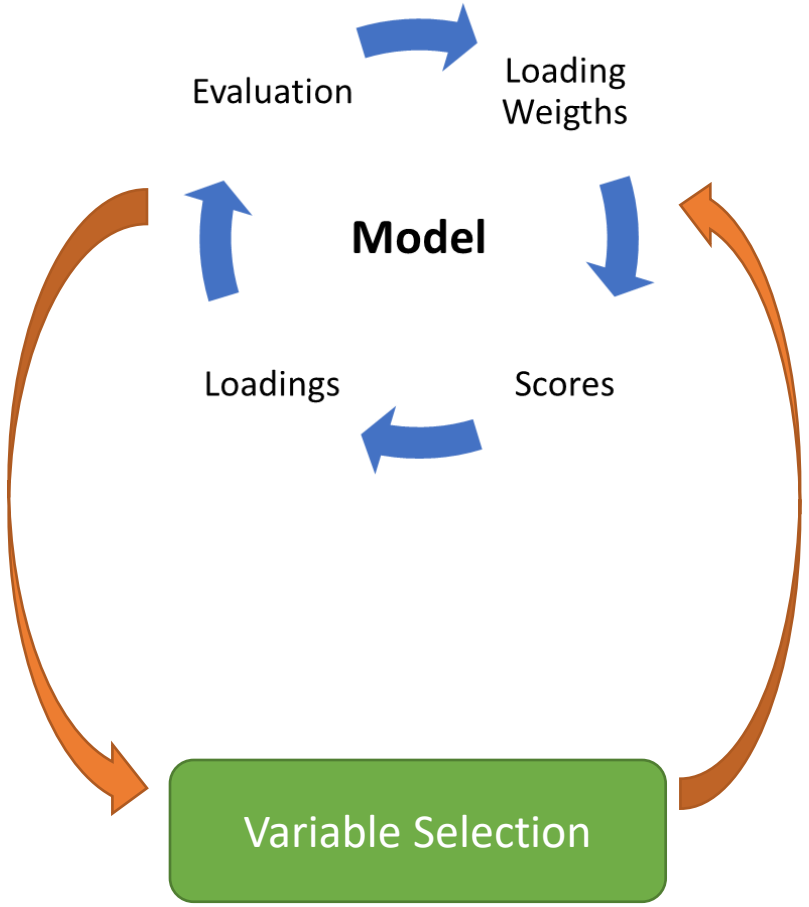
Variables/Features Selection (FS)



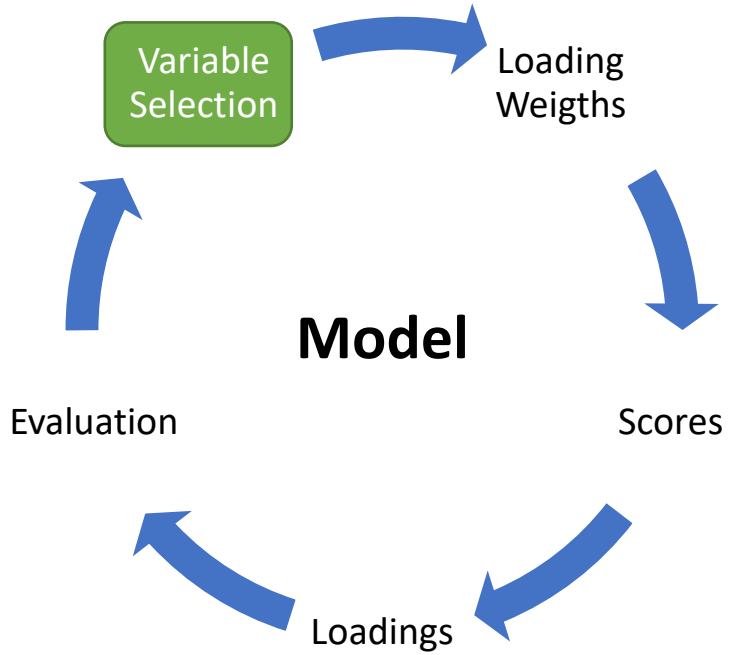
Variables/Features Selection (FS)



1. Filter methods

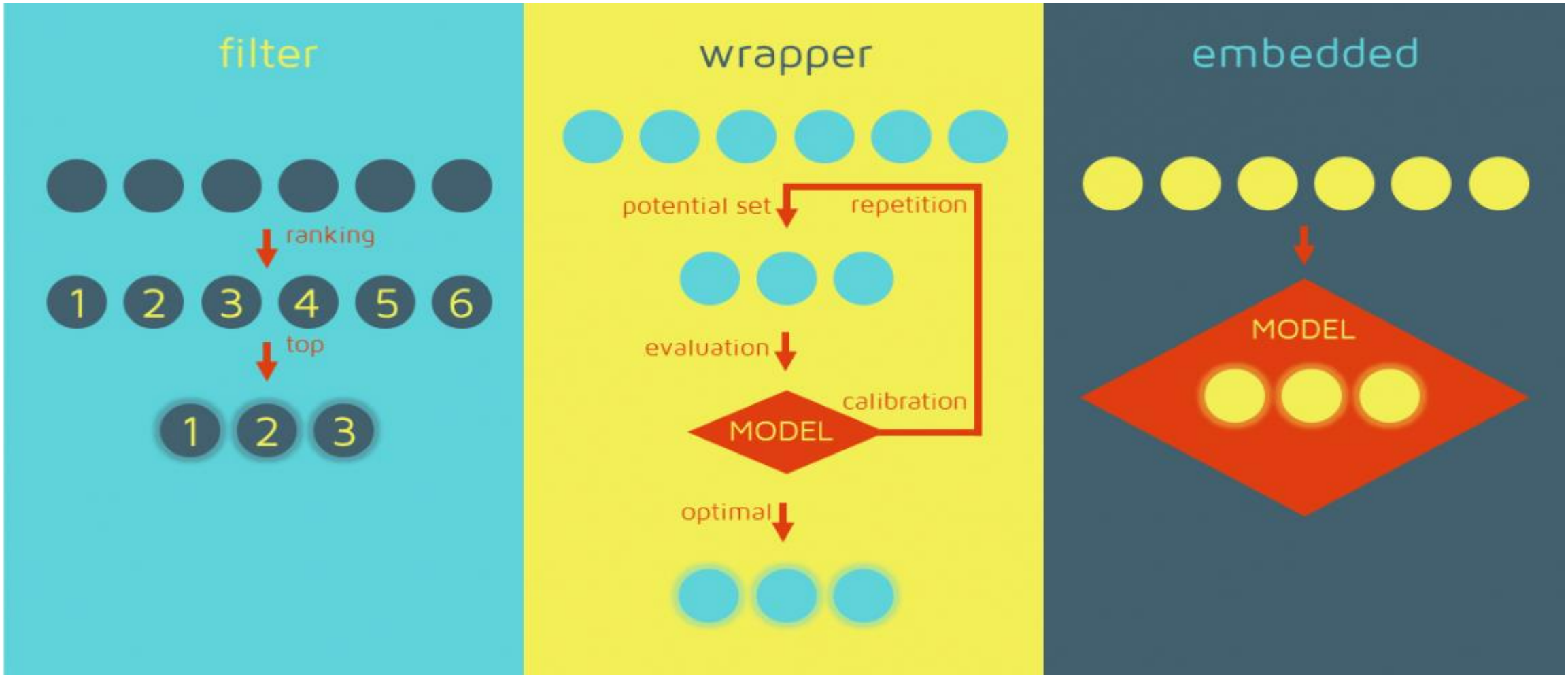


2. Wrapper methods



3. Embedded methods

Variables/Features Selection (FS)

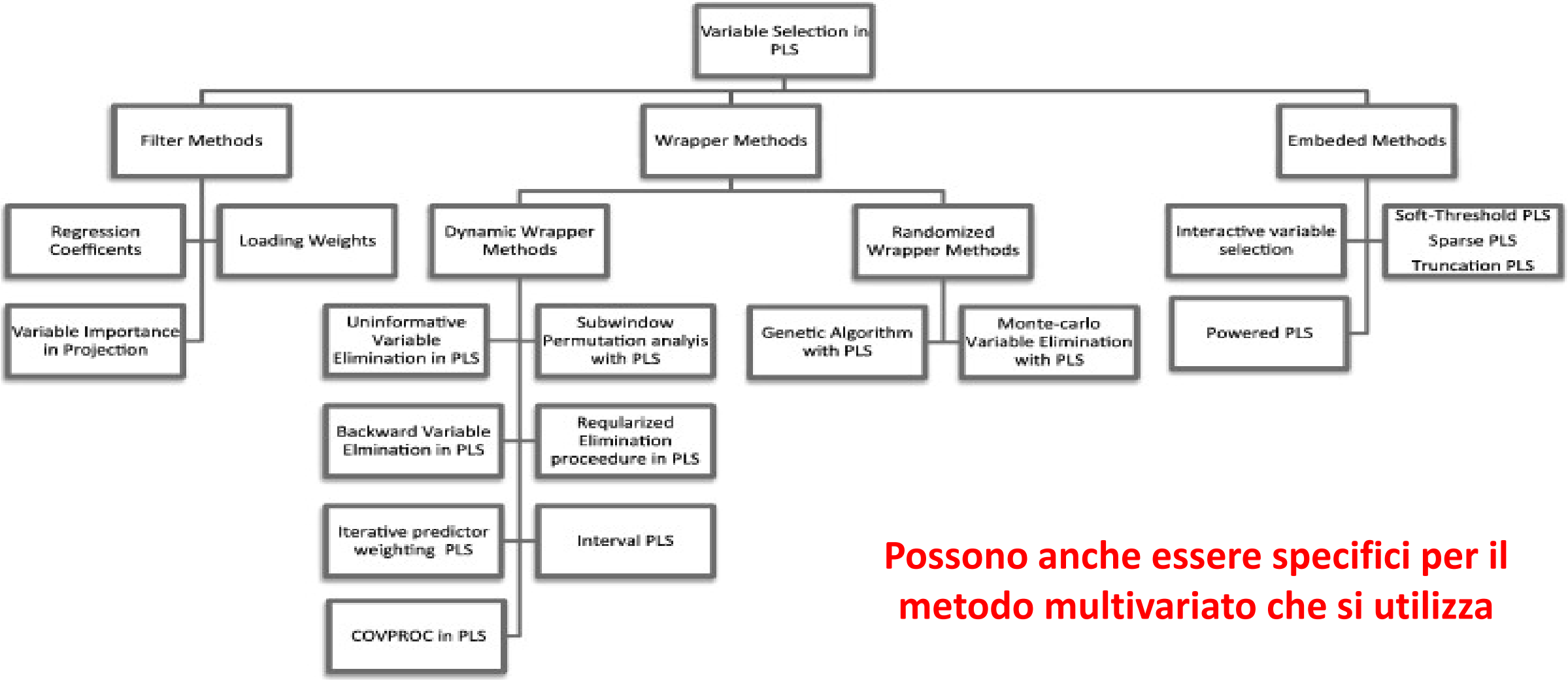


Variables/Features Selection (FS)

	Method	Classifier
Filter	VIP	PLS-DA
	SR	PLS-DA
	Discriminant power	SIMCA
	Regression coefficients, weights	PLS-DA
	Loadings, canonical vectors	PCA-LDA, LDA
Wrapper	UVE	PLS-DA
	GA	PLS-DA, LDA, SVM, ANN, etc.
	Interval based (iPLS, iECVA)	PLS-DA, ECVA
Embedded	Sparsity in regression coefficients, weights	PLS-DA, SVM
	Sparsity in loadings	PCA-LDA
	Sparsity in Mahalanobis distance	LDA, SHM

PCA

Variables/Features Selection (FS)



Possono anche essere specifici per il metodo multivariato che si utilizza

FILTER

Filter methods

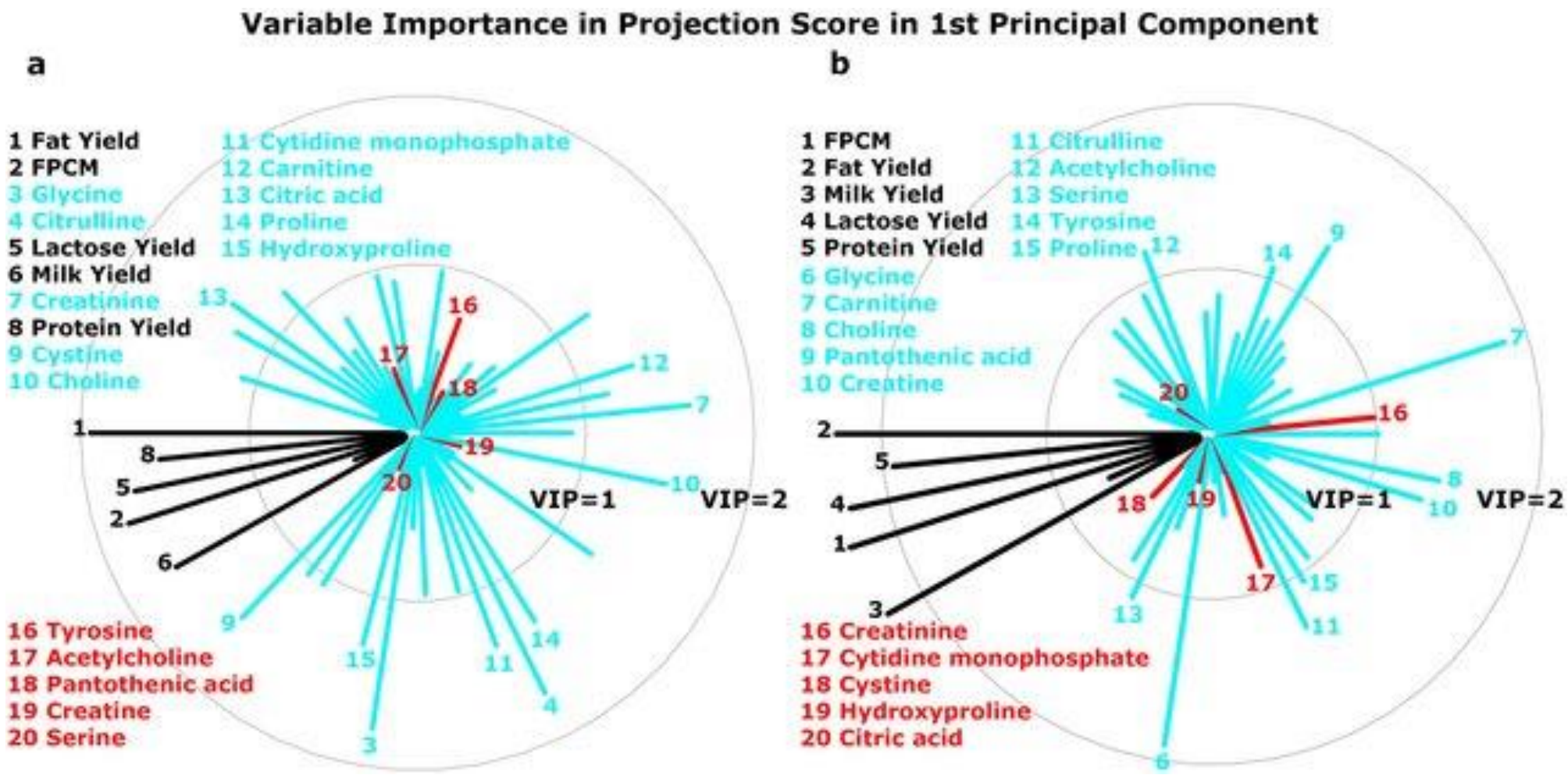
1. Variable importance in projection (VIP);
2. Selectivity Ratio (SR);
3. PCA

Filter: la selezione delle variabili viene effettuata secondo un criterio di valutazione che è in grado di QUANTIFICARE quanto un determinato insieme di variabili (subset) è in grado di discriminare 2 o più classi (o di effettuare una regressione).

Parametri classificatori (classifier): T^2, Q, p-value, RMSECV, RMSEP, PRESS, ecc.

FILTER

Variable importance in projection (VIP)



FILTER

Variable importance in projection (VIP)

$$v_j = \sqrt{p \sum_{a=1}^{A} \left[SS_a (w_{aj} / \|w_a\|)^2 \right] / \sum_{a=1}^{A} (SS_a)}$$

VIP della j-ima variabile

Loading

SS_a è la somma dei quadrati dei pesi normalizzati per la a-ima componente

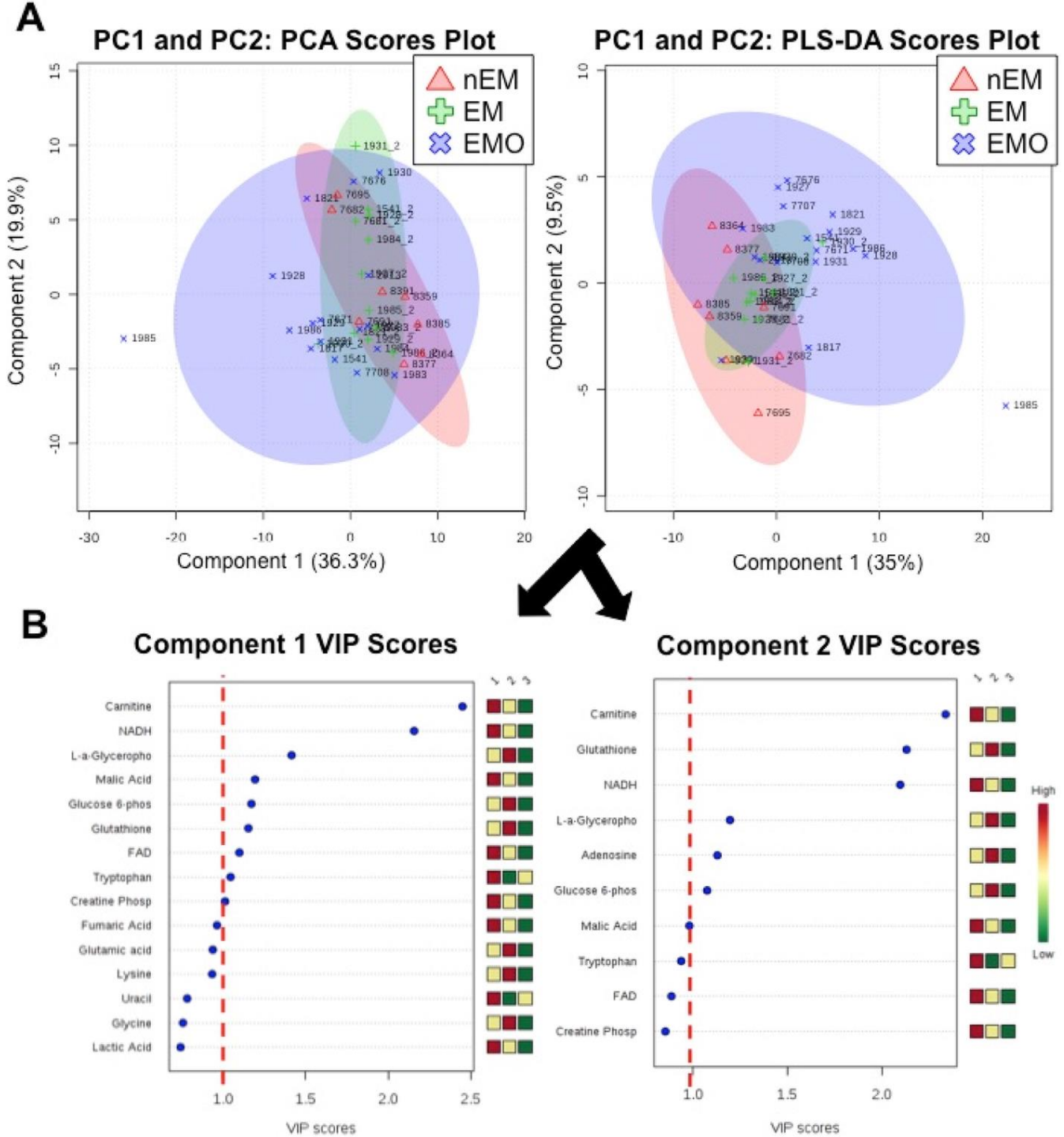
Peso della j-ima variabile per il modello che usa la a-ima componente

Matrice dei dati originali (trasposta)

$$w_a = X'_{a-1} y_{a-1}$$

Classe/responso

FILTER


A
PC1 and PC2: PCA Scores Plot
nEM
EM
EMO
Component 2 (19.9%)
Component 1 (36.3%)
PC1 and PC2: PLS-DA Scores Plot
Component 2 (9.5%)
Component 1 (35%)
B
Component 1 VIP Scores
Carnitine
NADH
L-a-Glyceropho
Malic Acid
Glucose 6-phos
Glutathione
FAD
Tryptophan
Creatine Phosp
Fumaric Acid
Glutamic acid
Lysine
Uracil
Glycine
Lactic Acid
VIP scores
Component 2 VIP Scores
Carnitine
Glutathione
NADH
L-a-Glyceropho
Adenosine
Glucose 6-phos
Malic Acid
Tryptophan
FAD
Creatine Phosp
VIP scores
High
Low

FILTER

Selectivity Ratio (SR)

Score

Loading

$$SS_{i,explained} = \|\mathbf{t}_{TPi}\mathbf{p}_{TPi}'\|^2$$

$$SS_{i,residual} = \|\mathbf{e}_{TPi}\|^2$$

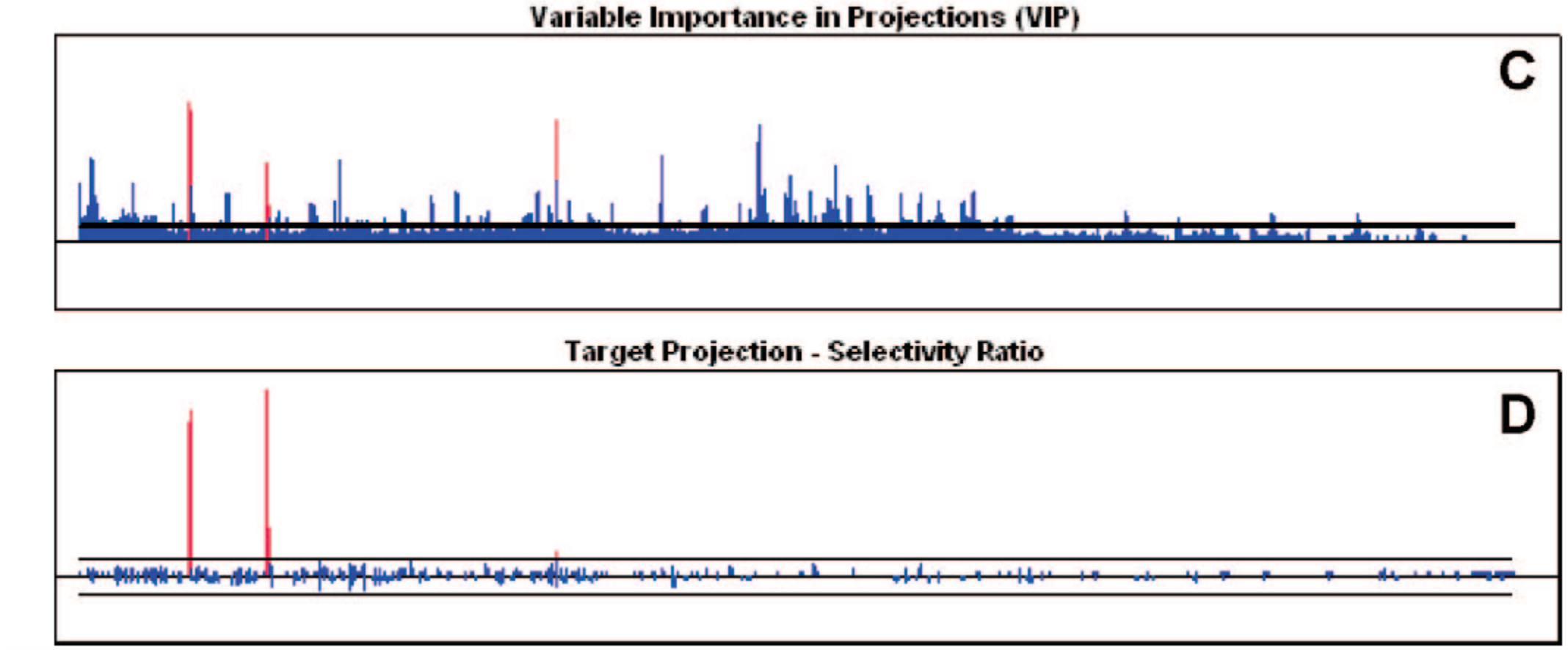
Errore (residuo)

$$SR_i = SS_{explained,i}/SS_{residual,i}$$

FILTER

Selectivity Ratio (SR)

DIFFERENTI RESULTATI IN SEGUITO ALLA SELEZIONE DELLE VARIABILI



FILTER

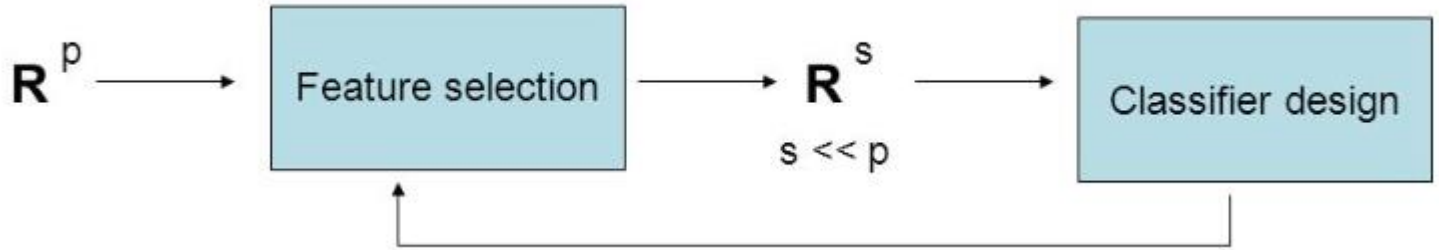
Problemi con i metodi Filter

1. **Ridondanza** delle variabili selezionate: le variabili sono considerate come INDIPENDENTI e non si tiene conto della eventuale presenza di correlazione (problema in spettroscopia);
2. Alcuni metodi sono in grado di valutare una eventuale **interazione** tra variabili (ci sono metodi di filtering più performanti rispetto ad altri);
3. Non è possibile definire a priori **quale classificatore** (*classifier*) utilizzare per valutare le variabili, al fine di definire se quelle selezionate siano significative o meno. Di conseguenza, è necessario valutare molteplici classificatori (RMSECV, RMSEP, PRESS, ma anche **p-value, Hotelling's T^2 e residui Q**) e vedere se il modello che si ottiene è migliorato.

WRAPPER

Wrapper methods

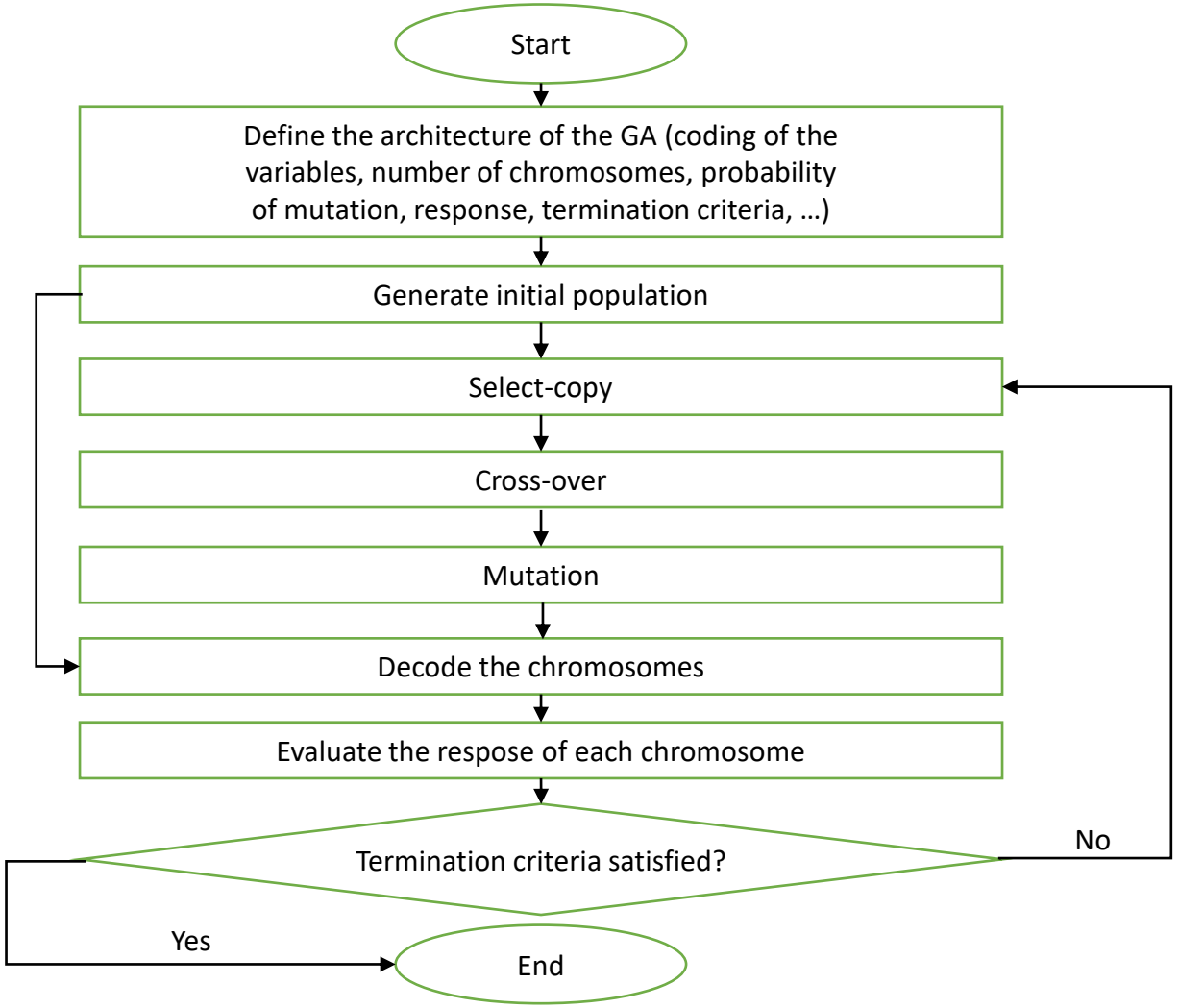
1. Algoritmi genetici (GA)
2. Uninformative variable elimination (UVE);
3. Backward/forward variable elimination (BVE);
4. Interval PLS (iPLS)



Approccio iterativo: si valutano molti subset di variabili e a questi si da un punteggio (score). Si usa il subset che, in termini di classificazione o regressione (RMSEP, RMSECV, PRESS) presenta il risultato migliore.

WRAPPER

Algoritmi Genetici (GA)



J. Chemometrics 2012 Apr;26:345-351

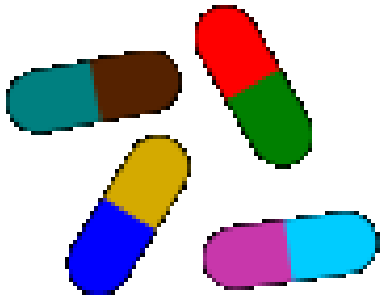
WRAPPER

Algoritmi Genetici (GA)

4) Si effettuano mutazioni casuali

3) Si costruiscono nuove combinazioni (progenie)

1) Si combinano le variabili in modo casuale



2) Si trovano le combinazioni migliori

WRAPPER

Uninformative variable elimination (UVE)

The UVE-PLS algorithm can be summarized as follows:

1. Determination of the optimal model complexity (A) on **X**, with the lowest RMSEP as the criterion[15]

$$\text{RMSEP} = \left(\sum_{i=1}^{n} (\hat{y}_i - y_i)^2/n\right)^{1/2} \qquad (4)$$

2. Generation of the artificial variable matrix **R**[16] and its multiplication by a small constant (10^{-10}). This yields the matrix **R** (n, p) with the number of variables p equal to the number of variables in **X**. The a priori probability to make an error in selection, i.e., to eliminate an informative or to retain an uninformative variable is then the same in both **X** and **R**. Inclusion of **R** with **X** (n, p). The resulting matrix is called **XR** (n, $2p$), the p first columns being those of **X** and the p last ones being those of **R**.

3. Calculation of PLS models for **XR** according to a leave-one-out procedure. The number of factors retained (A) is the same as for **X**. This yields n PLS models each with $2p$ regression coefficients b. They are collected in a matrix **B** (n, $2p$).

4. Determination for each variable j (i.e., both the experimental and random variables) of b_j ($b_j = \sum_{i=1}^{n} b_{ij}/n$), i.e., the mean of the column vector j from **B** and the standard deviation of that column vector

$$s(b_j) = \left(\sum_{i=1}^{n} (b_{ij} - b_j)^2/(n-1)\right)^{1/2} \qquad (5)$$

5. Determination for each variable j of the criterion $c_j = b_j/s(b_j)$.

6. Determination of $\max(\text{abs}(c_{\text{artif}}))$, i.e., the highest absolute value of c among all c for artificial variables.

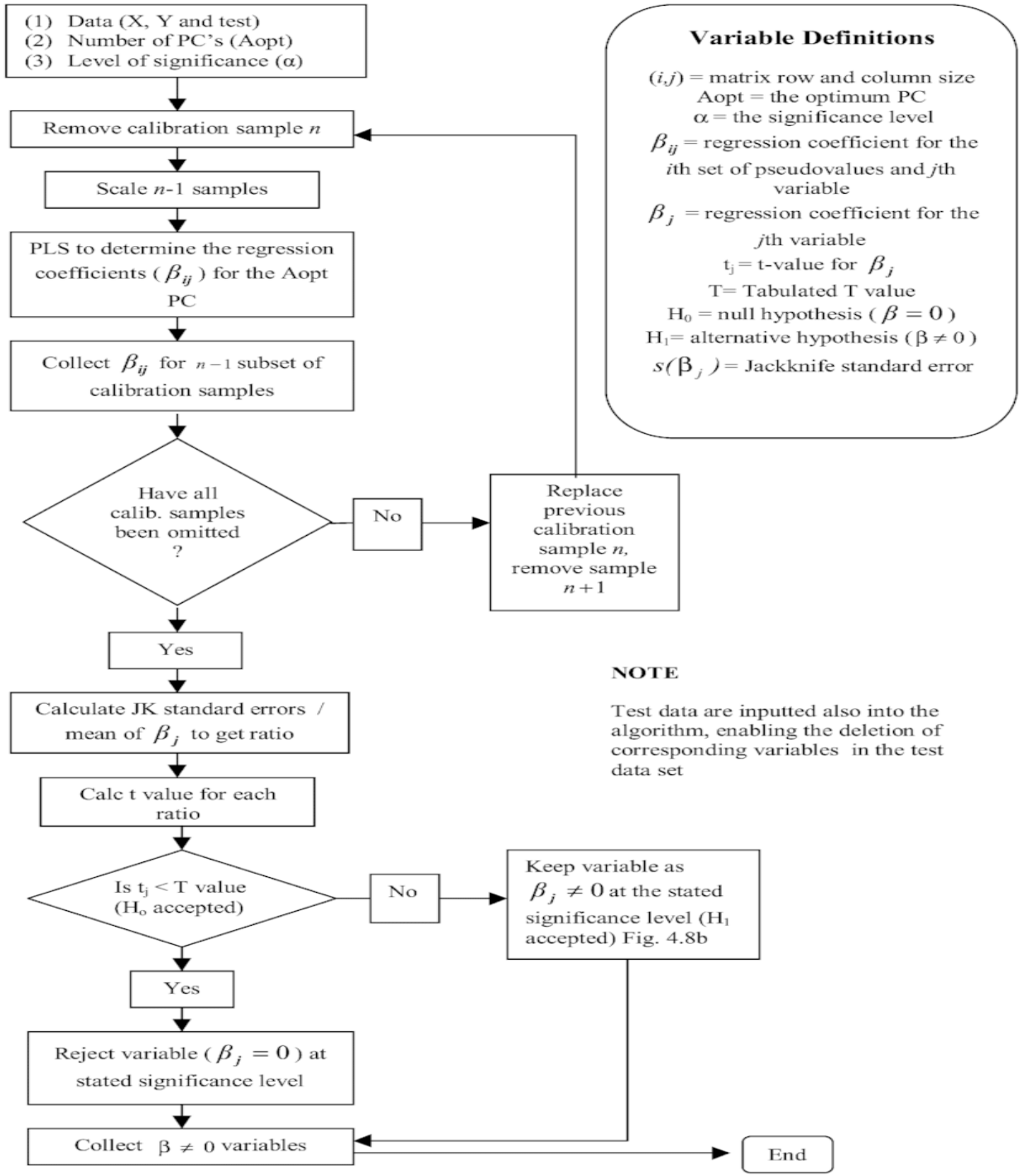
7. Elimination from **X** of the experimental variables for which $\text{abs}(c_j) < \text{abs}(\max(c_{\text{artif}}))$, for $j = 1, \ldots, p$. The remaining variables constitute the new **X** matrix, $\mathbf{X}_{\text{new}}$.

8. Building of the final PLS leave-one-out cross-validated models on $\mathbf{X}_{\text{new}}$ and prediction $\hat{y}$ with A factors.

9. Quantification of the predictive ability of the new model as the cross-validated $\text{RMSEP}_{\text{new}}$ according to eq 4.

10. If (a) $\text{RMSEP}_{\text{new}} >$ RMSEP one concludes that the elimination of uninformative variables did not improve modeling and the algorithm is terminated. Otherwise if (b) $\text{RMSEP}_{\text{new}} <$ RMSEP, one will first wonder whether A was not too large (overfitting), due to the uninformative variables which could have influenced the selection (it is extremely improbable that A was too small due to uninformative variables). In order to check this possibility, the algorithm starting with a new selection on **XR** (point 2) is repeated again for $A = A - 1$ and the original RMSEP is replaced by the $\text{RMSEP}_{\text{new}}$. When the reduction of A to $A - 1$ does not improve modeling ($\text{RMSEP}_{\text{new}} >$ RMSEP), the algorithm terminates in 10 (a).

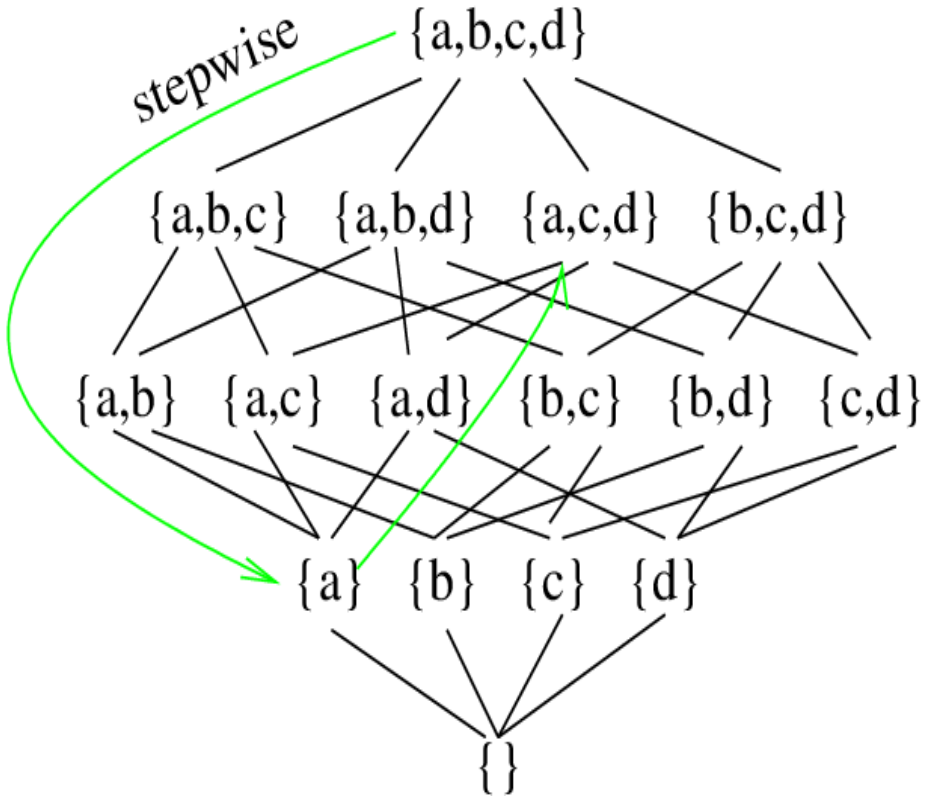
WRAPPER



Analytical Chemistry, Vol. 68, No. 21, November 1, 1996

WRAPPER

Backward/forward variable elimination (BVE)

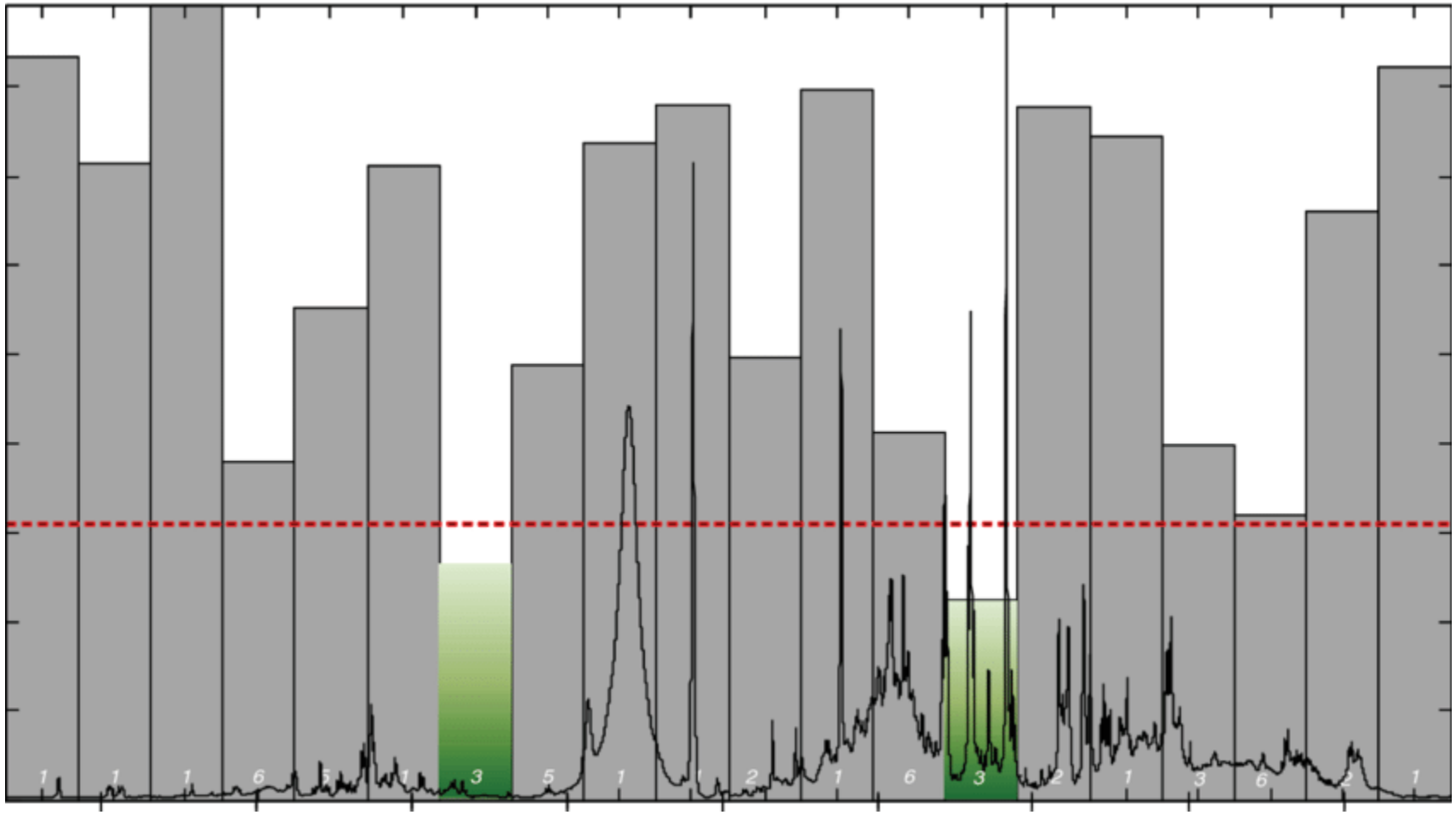



```
Forward stepwise selection
  R^2    Predictors
1 0.3215 ['CRBI']
2 0.4252 ['CRBI', 'Hits']
3 0.4514 ['CRBI', 'Hits', 'PutOuts']
4 0.4754 ['CRBI', 'Hits', 'PutOuts', 'Division_W']
5 0.4908 ['CRBI', 'Hits', 'PutOuts', 'Division_W', 'AtBat']
6 0.5087 ['CRBI', 'Hits', 'PutOuts', 'Division_W', 'AtBat', 'Walks']
7 0.5132 ['CRBI', 'Hits', 'PutOuts', 'Division_W', 'AtBat', 'Walks', 'CWalks']

Backrward stepwise selection
  R^2    Predictors
7 0.5136 ['AtBat', 'Hits', 'Walks', 'CRuns', 'CWalks', 'PutOuts', 'Division_W']
6 0.4997 ['AtBat', 'Hits', 'Walks', 'CRuns', 'PutOuts', 'Division_W']
5 0.4841 ['AtBat', 'Hits', 'Walks', 'CRuns', 'PutOuts']
4 0.4664 ['AtBat', 'Hits', 'CRuns', 'PutOuts']
3 0.4485 ['Hits', 'CRuns', 'PutOuts']
2 0.4148 ['Hits', 'CRuns']
1 0.3166 ['CRuns']
```

WRAPPER

Interval PLS (iPLS)



WRAPPER

Problemi dei metodi Wrapper

1. Costosi da un punto di vista computazionale: è necessario calcolare un modello multivariato per ogni subset di variabili preso in considerazione.

2. Non è possibile effettuare una ricerca di tipo esaustivo (si dovrebbero calcolare troppi modelli)
   → i risultati potrebbero variare a seconda del subset iniziale che viene scelto.

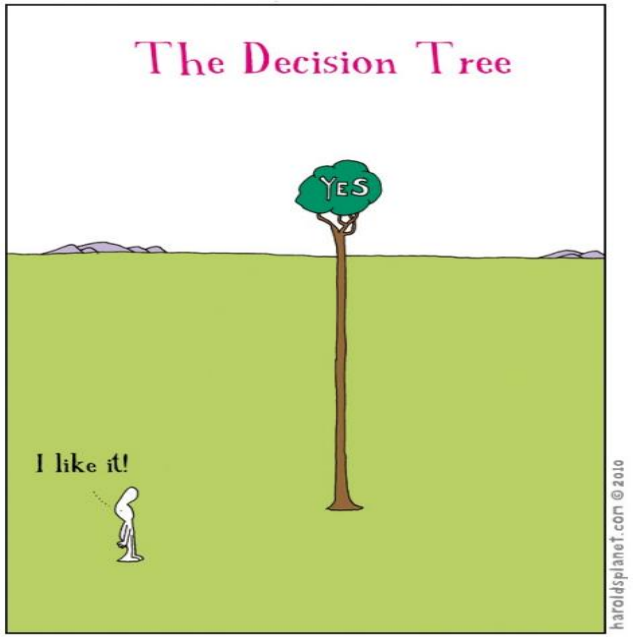
3. Possibilità di over-fitting.

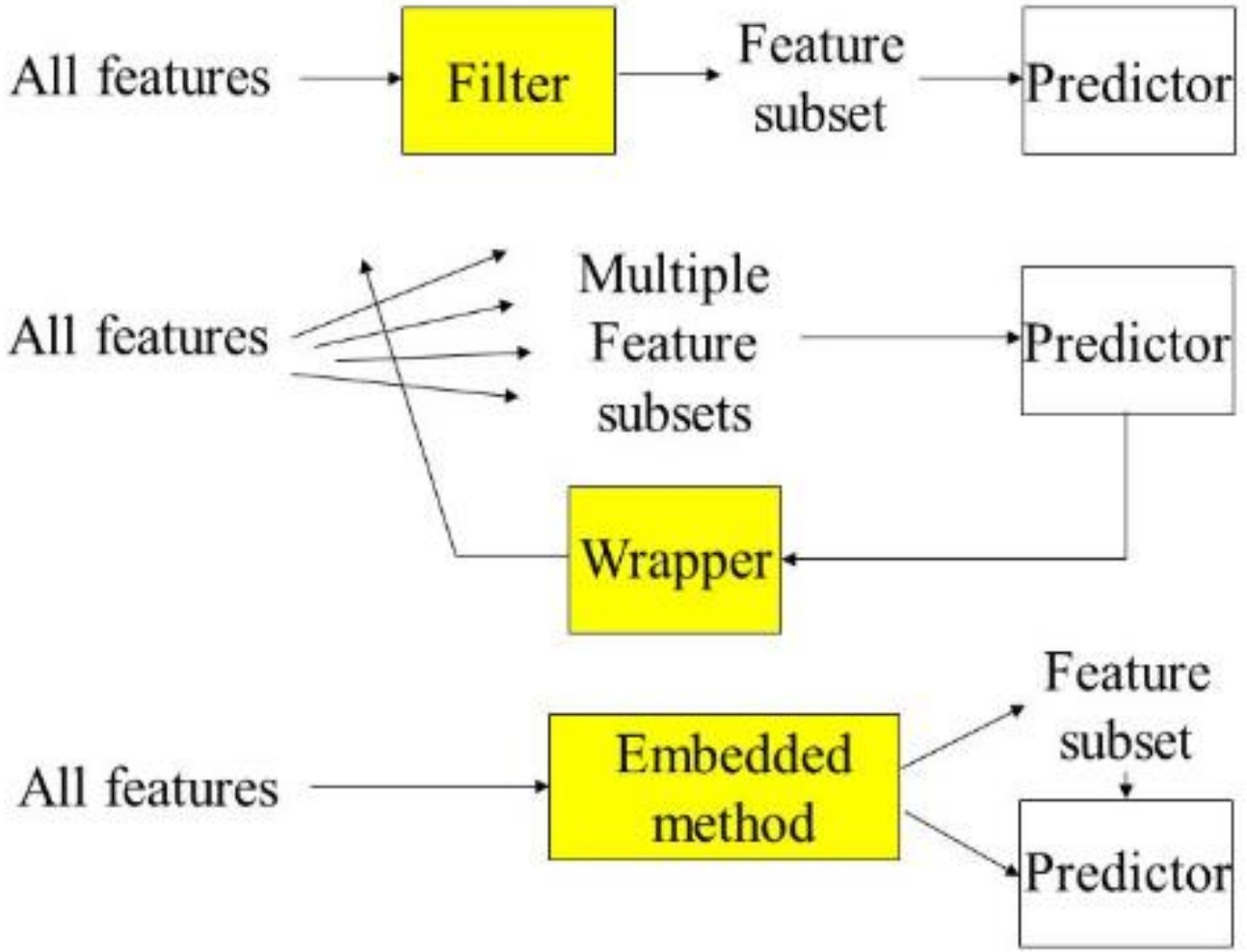
EMBEDDED

Embedded methods

Moltissimi metodi!

Esempi: decision tree, random forests



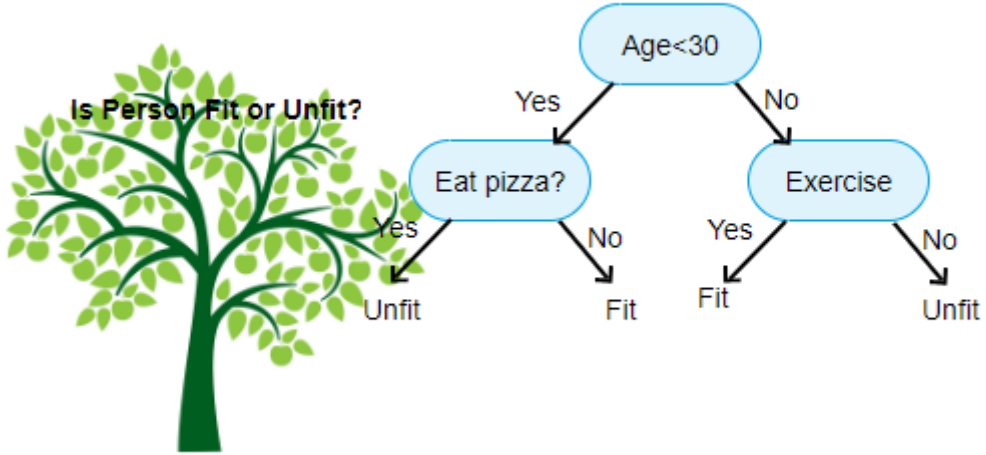


EMBEDDED

Embedded methods

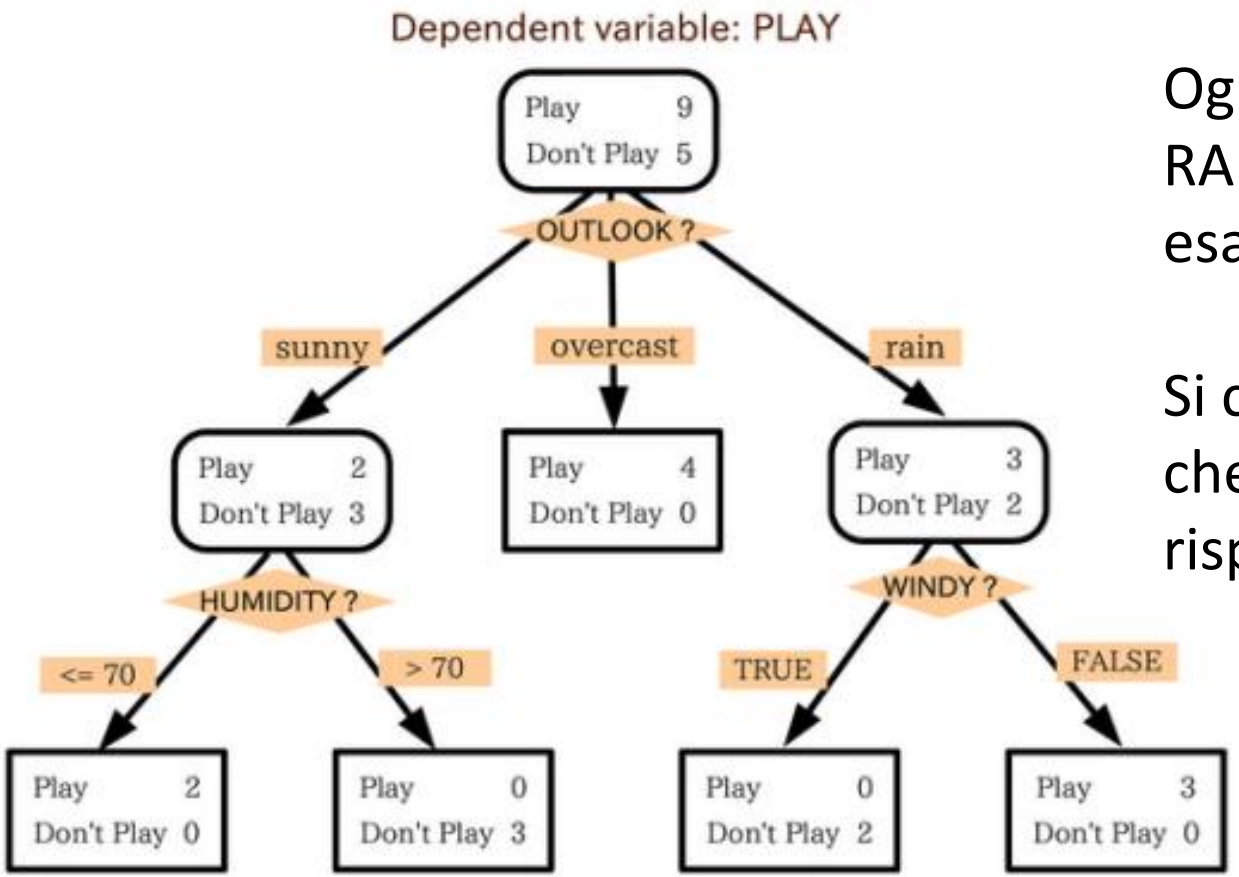
La selezione delle variabili è parte integrante della stessa costruzione del modello.
Questi metodi danno difficilmente over-fitting ma sono computazionalmente complessi (time-consuming).

I metodi più noti sono quelli di tipo Decision Trees e Random Forests.



EMBEDDED

Metodo Decision Tree



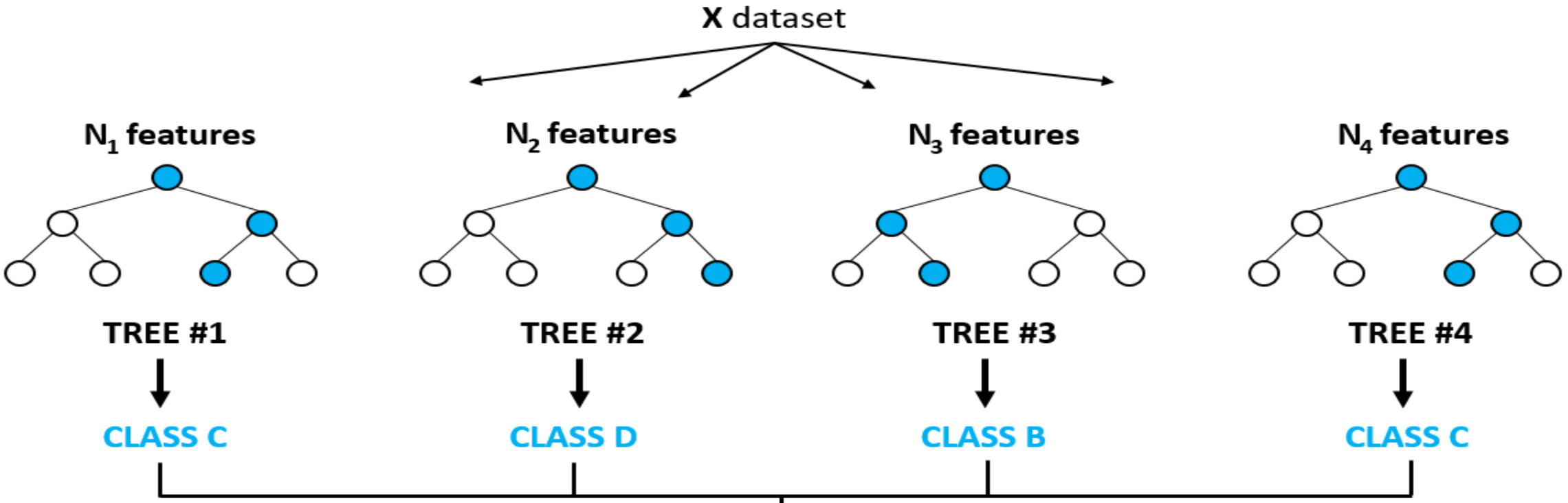
Ogni albero è costituito da NODI (variabili) e RAMI (livelli/categorie delle variabili prese in esame).

Si cercano, per ogni nodo, quali sono le variabili che meglio separano i campioni presi in esame, rispetto ad una specifica classe/responso.

Rischio di over-fitting!
Buona interpretazione dei risultati

EMBEDDED

Metodo Random Forests

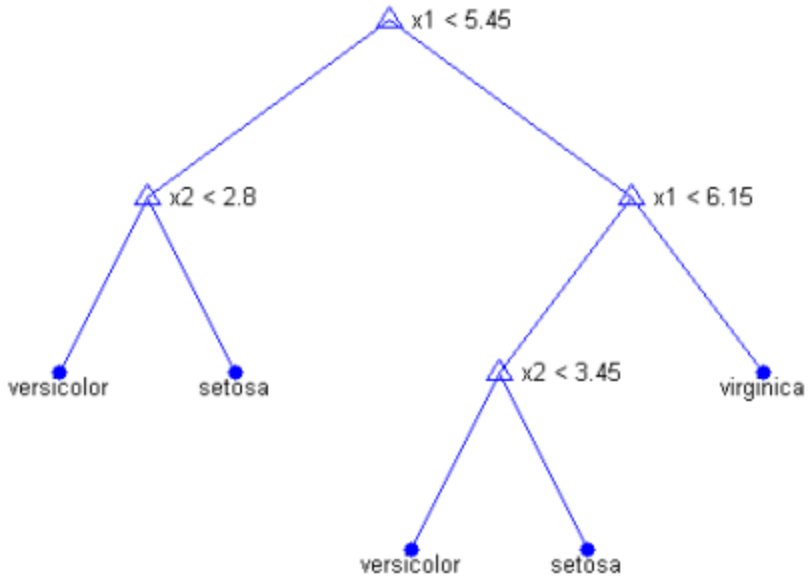


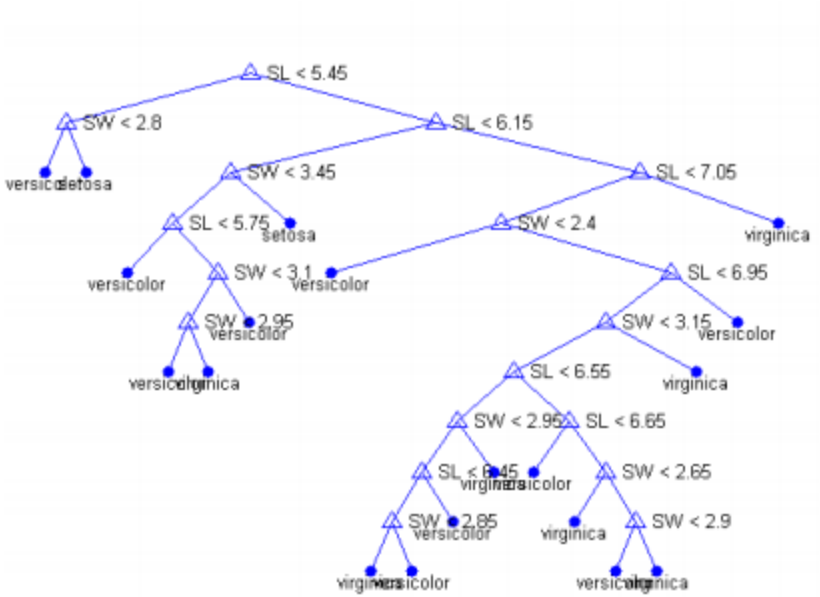
Costruisco molti decision trees e valuto la maggioranza delle classificazioni.

Bassissimo rischio di over-fitting ma possibilità di interpretazione molto limitata.

EMBEDDED

Metodi Decision Tree





Possono dare risultati diversi tra una prova e l'altra

EMBEDDED

Metodi Decision Tree

For each feature X_k
  For each tree t
    If X_k was used, examine the split
      Record the information-gain $gain(k,t)$
    End
  End
End

Feature score: $J(X_k) = \frac{1}{T}\sum_{t=1}^{T} gain(k,t)$

(Averaged gain in the forest)

Ampiamente usati in bio-informatica

Un parametro di informazione (information-gain) ci dice quanto la variabile selezionata per un certo nodo è significativa

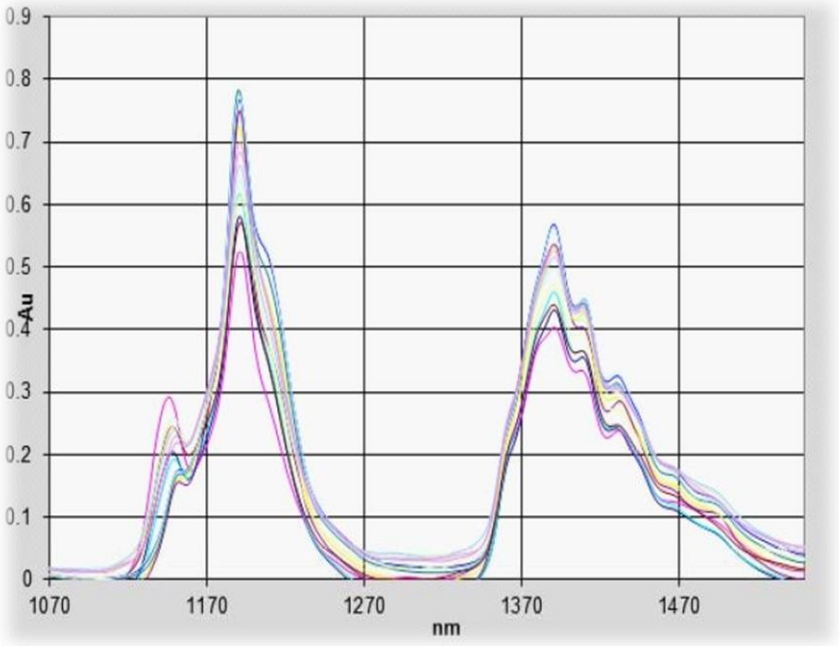
EMBEDDED

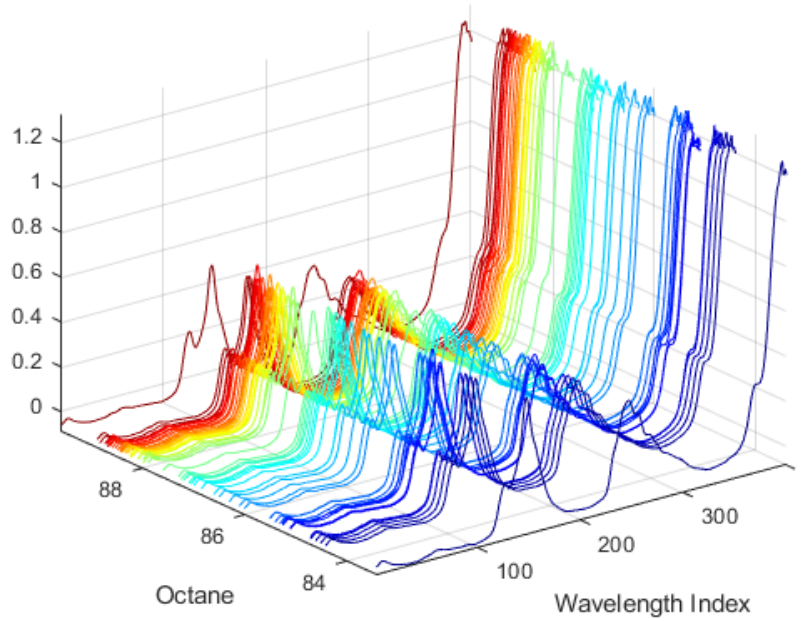
Problemi dei metodi Embedded

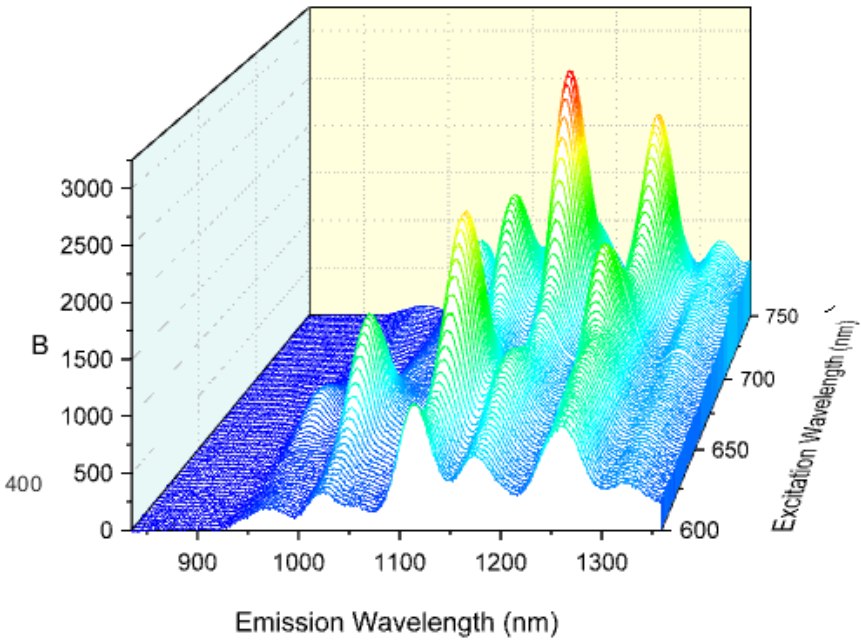
1. Estremamente costosi da un punto di vista computazionale.
2. Vanno testati molteplici volte.
3. Difficilmente interpretabili.

Metodi di Regressione Multivariata

Esempi







The selectivity problem – troppa informazione!

Per questo motive è necessario utilizzare un approccio multivariato

OLS: Ordinary Least Squares Regression

Tipologie possibili di regression OLS

All'interno dei metodi di **tipo OLS** ritroviamo le regressioni:

- **Lineare semplice** ($y = b + mx$);
- **Polinomiale** ($y = b + mx + nx^2$);
- **Lineare multipla** ($y = b + m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_nx_n$).

OLS: Ordinary Least Squares Regression

Tipologie possibili di regression OLS

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki} \quad i = 1 \ldots n$$

Il valore predetto (la variabile dipendente)

L'intercetta (passa per X=0)

I predittori (le variabili misurate, indipendenti)

I coefficienti di regressione (pendenze) per ogni specifico predittore

OLS: Ordinary Least Squares Regression

Obiettivi della regressione OLS

Minimizzare gli errori (residui, E)

$$\sum_{i=1}^{n}(\mathrm{Y_i}-\hat{\mathrm{Y}}_\mathrm{i})^2=\sum_{i=1}^{n}(\mathrm{Y_i}-\hat{\beta}_0+\hat{\beta}_1\mathrm{X_{1i}}+\ldots+\hat{\beta}_\mathrm{k}\mathrm{X_{ki}})^2=\sum_{i=1}^{n}\varepsilon_i^2$$

Il modello…

$$\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{e}$$

…viene calcolato mediante i seguenti parametri

$$\mathbf{b}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

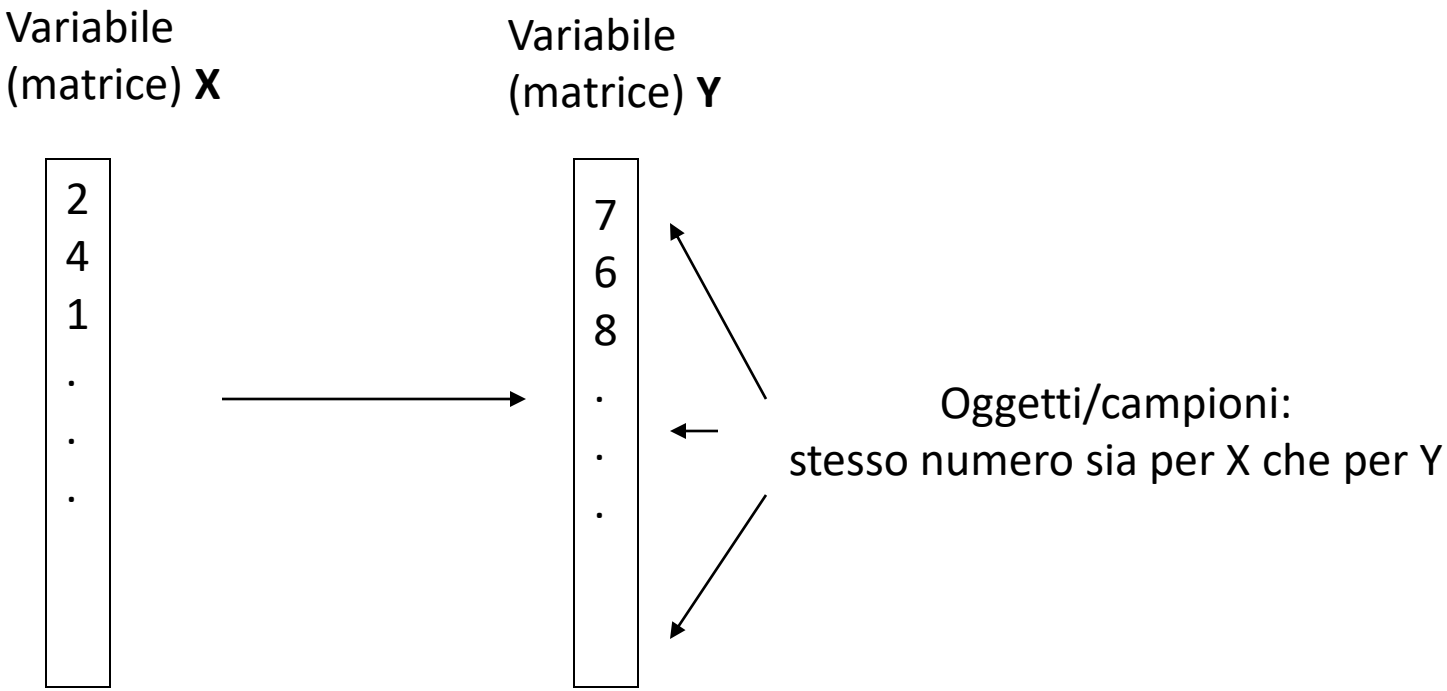
OLS: Ordinary Least Squares Regression

Obiettivi della regressione OLS

- Una Y e una X. Utilizza X per predire Y.
- Utilizza un modello/equazione lineare per trovare un approssimazione (*fit*) mediante il calcolo dei minimi quadrati.
- Le variabili devono presentare una distribuzione gaussiana.

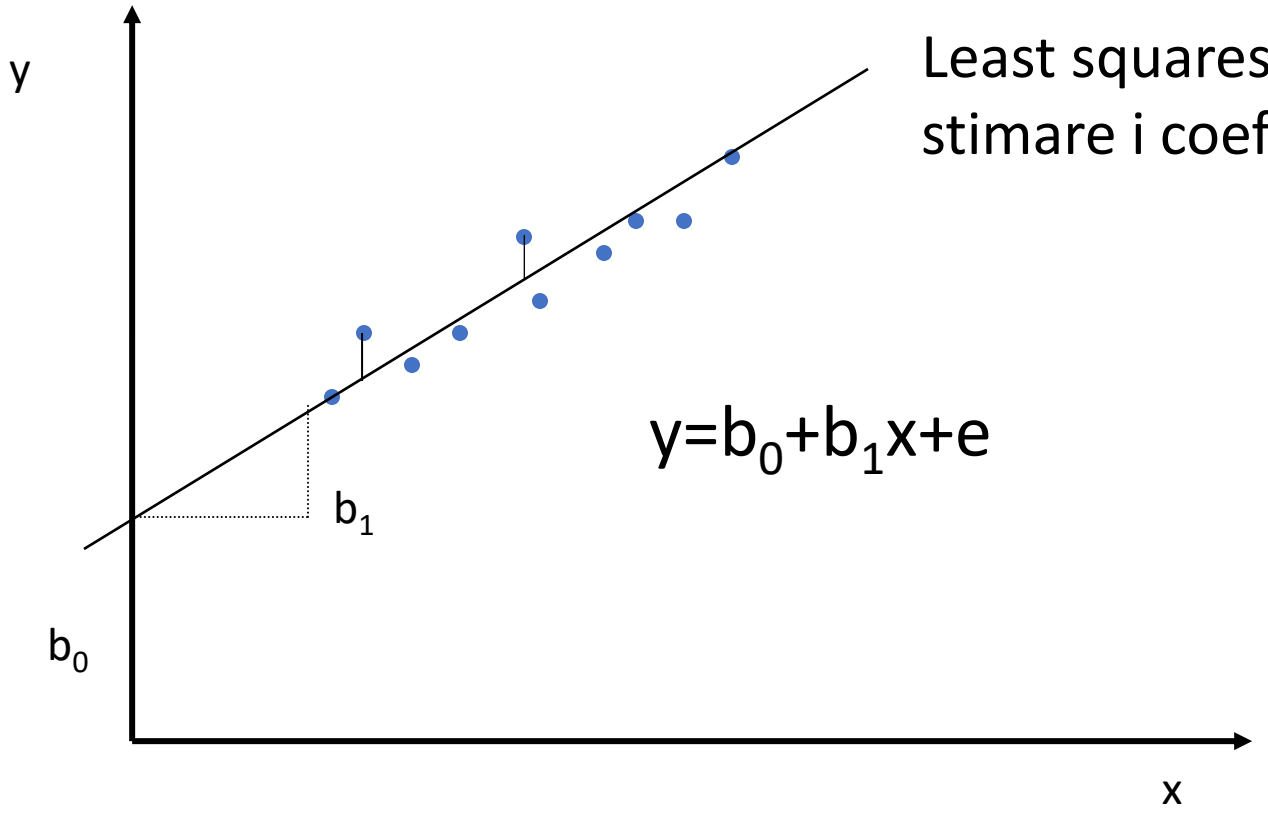
OLS: Ordinary Least Squares Regression

La struttura dei dati



OLS: Ordinary Least Squares Regression

Il modello



Least squares (LS) viene utilizzato per stimare i coefficient lineari

$$y=b_0+b_1x+e$$

Regressione lineare semplice

OLS: Ordinary Least Squares Regression

Assunzioni

Normalità — Le variabili in esame presentano distribuzioni di tipo gaussiane.

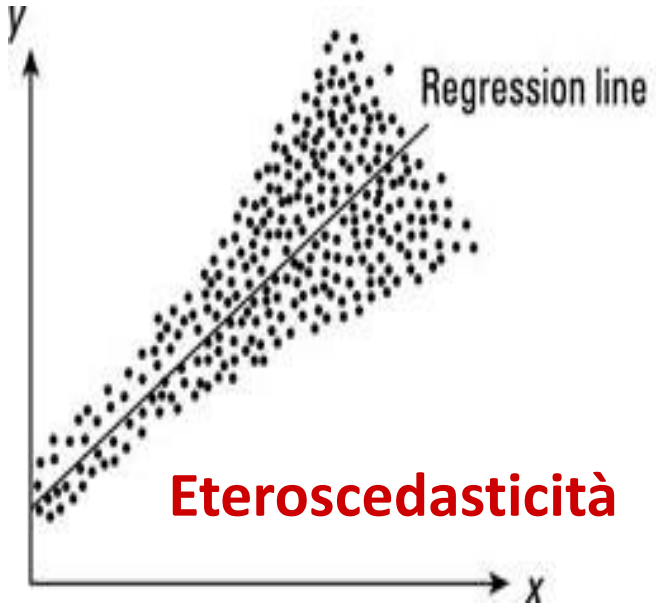
Indipendenza —I valori delle Y sono uno indipendente dall'altro.

Linearità —La variabile dipendente è correlata in maniera lineare alla variabile indipendente.

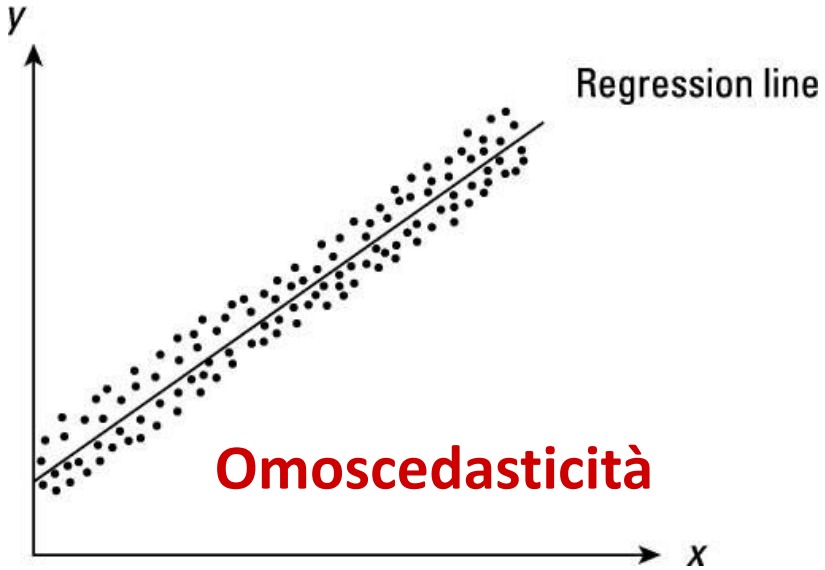
Omoscedasticità —La varianza della variabile dipendente non cambia al variare della variabile indipendente.

OLS: Ordinary Least Squares Regression

Omoscedasticità vs eteroscedasticità



Eteroscedasticità



Omoscedasticità

We could call this constant variance, but saying homoscedasticity makes me feel smarter (R.I. Kabacoff).

Metodi di Regressione Multivariata

Introduzione

L'approccio tradizionale OLS…

LS - MLR

…così come la Regressione Multipla Lineare…

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

…non si può utilizzare quando si hanno molte variabili!

Metodi di Regressione Multivariata

Introduzione

- MLR fornisce:
  - valori predetti
  - coefficienti di regressione
  - grafici “diagnostici”

- Ma se ci sono molte variabili correlate/collineare (comb. lin. di variabili)
  - fornisce equazioni di regressione **instabili**
  - risultano molto **complesse** da elaborare → viene meno il loro utilizzo

Metodi di Regressione Multivariata

Approcci possibili

- Selezionare le variabili che risultano più significative (stepwise methods)

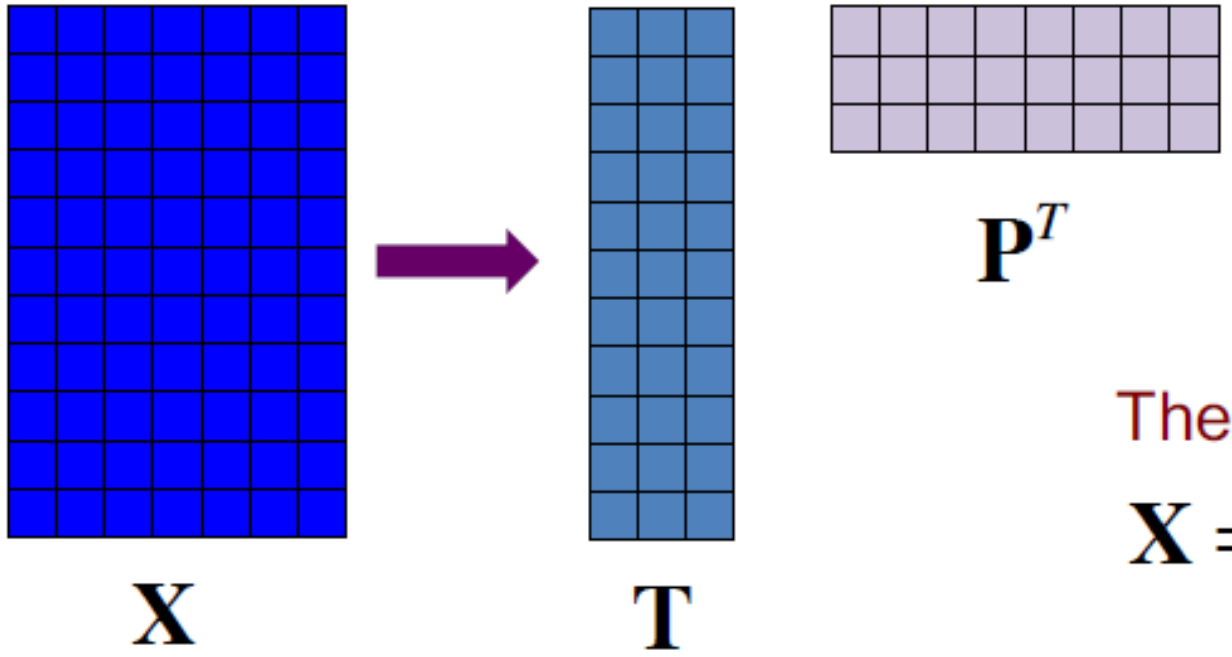
- Comprimere le variabili e direzionarle verso le “zone” dei dati più rilevanti (PCR, PLS)

Metodi di Regressione Multivariata

Approcci possibili – PCR (Principal Component Regression)

Procedura a 2 step:

1. Costruisco un modello di **PCA**;
2. Faccio **MLR** sugli scores.

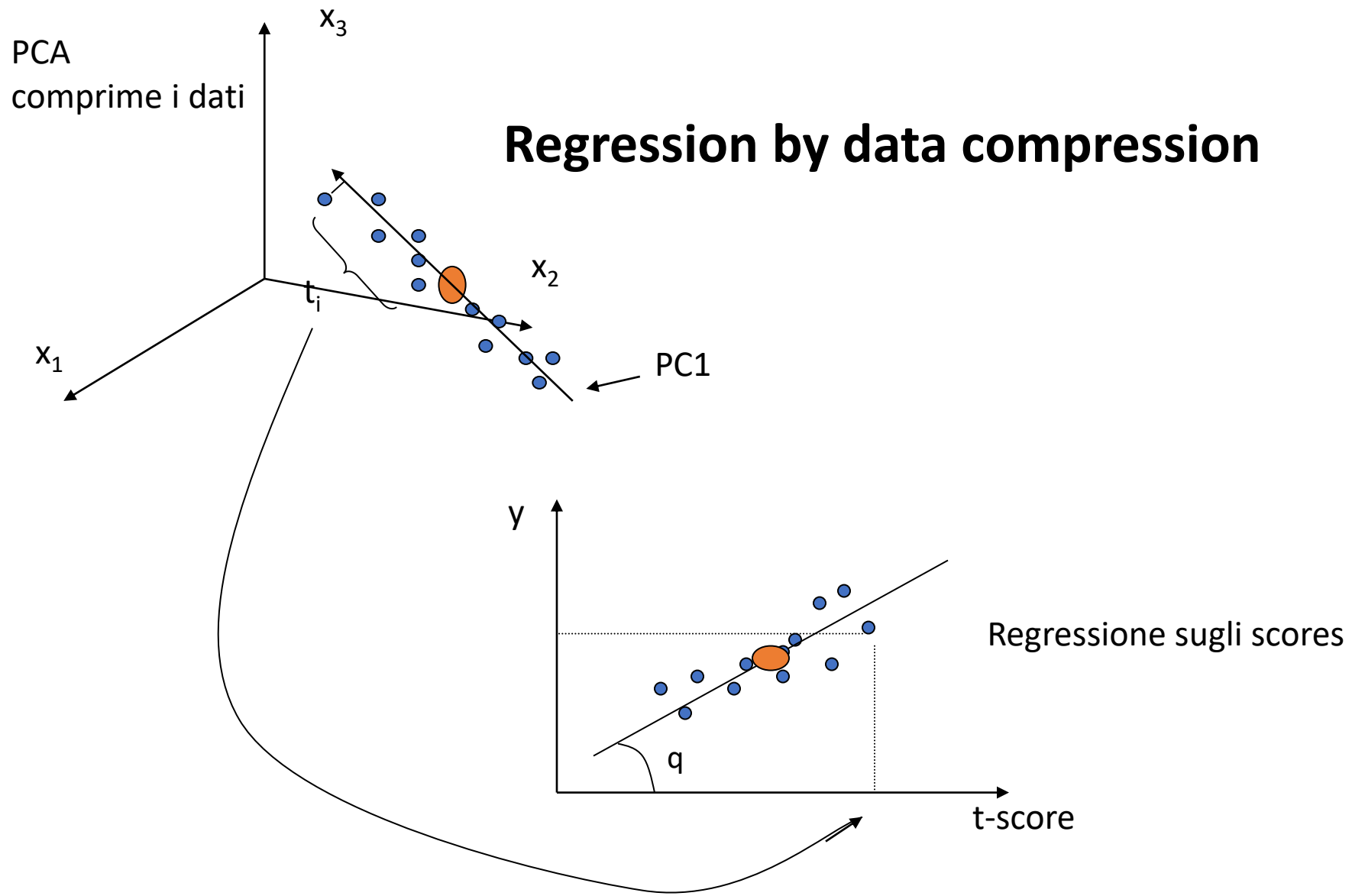


The PCA model:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_X$$

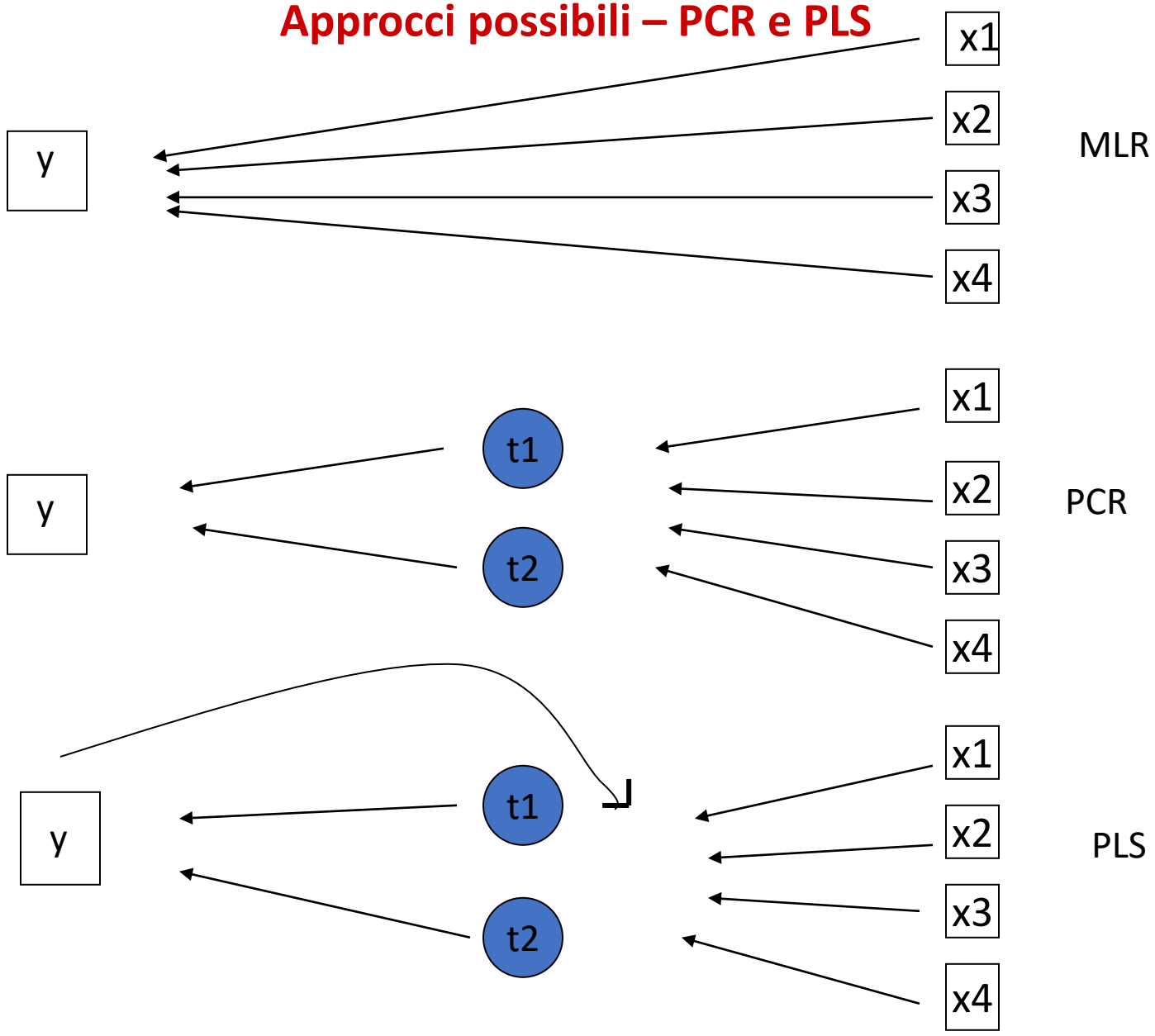
Metodi di Regressione Multivariata

Approcci possibili – PCR (Principal Component Regression)



Metodi di Regressione Multivariata

Approcci possibili – PCR e PLS



Metodi di Regressione Multivariata

Approcci possibili – PCR e PLS

Per ogni fattore/componente

- **PCR**
  - Massimizza la varianza delle combinazioni lineari di X

- **PLS**
  - Massimizza la covarianza delle combinazioni lineari di X e Y

Ogni fattore viene sottratto prima di passare al calcolo del fattore successivo

Metodi di Regressione Multivariata

Approcci possibili – PCR (Principal Component Regression)

- Usa le component principali;
- Risolve il problema della collinearità, fornendo risultati stabili;
- Fornisce grafici per l'interpretazione (scores e loadings);
- Facilmente spiegabile;
- Diagnostiche sugli outlier;
- Facilmente adattabile e modificabile.

Metodi di Regressione Multivariata

Approcci possibili – PLS-R (Partial Least Squares Regression)

PLS

$$X = T P^T + E$$
$$y = T_q + f$$

CENTERED X AND y

SET a=1

(i) MAXIMIZE COVARIANCE BETWEEN $t_a = X w_a$ AND y

(ii) FIND p_a AND q_a AND SUBTRACT THE FACTOR

$$X - t_a p_a^T \rightarrow X$$
$$y - t_a q_a \rightarrow y$$

INCREASE a BY 1 AND CONTINUE UNTIL a = A

Single y PLS-R

$$\max_{\mathbf{w}} \left(\mathrm{cov}(\mathbf{t}, \mathbf{y}) \right) \quad \text{with:} \quad \mathbf{t} = \mathbf{X}\mathbf{w}$$

Metodi di Regressione Multivariata

Approcci possibili – PLS-R (Partial Least Squares Regression)

- Facile da calcolare;
- Fornisce soluzioni stabili;
- Fornisce scores e loadings;
- Spesso il numero di componenti (***variabili latenti***) è inferiore a quelle calcolate con la PCR;
- Si possono ottenere migliori predizioni.

Metodi di Regressione Multivariata

Approcci possibili – PCR e PLS

PCR e PLS possono funzionare con più Y-variabili

- PCR è calcolata per ogni Y. Ciascun Y è valutata rispetto alle componenti principali;
- PLS: l'algoritmo è facilmente adattabile. Massimizza la combinazione lineare tra X e Y.
- Per entrambi i metodi: equazione di regressione e grafici.

Metodi di Regressione Multivariata

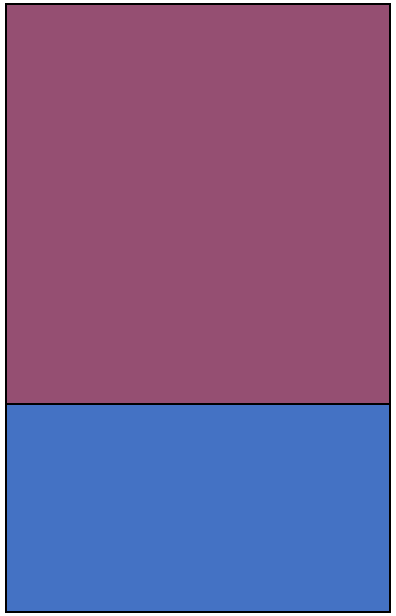
Approcci possibili – PCR e PLS – validazione

La validazione dei modelli è fondamentale

- Misura la qualità dei predittori;
- Determina il numero delle componenti;
- Consente di confrontare i metodi.

Metodi di Regressione Multivariata

Approcci possibili – PCR e PLS – validazione



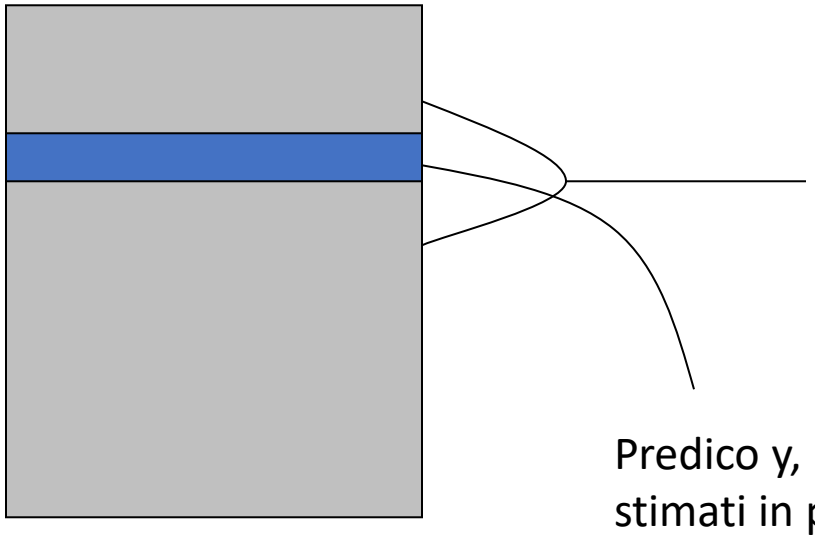
Calibrazione/training
Stima i coefficienti

Test/validazione
Predico y, utilizzando i coefficient stimati in precedenza

Metodi di Regressione Multivariata

Approcci possibili – PCR e PLS – cross-validazione

Cross-validazione



Calibrazione, trovare y=f(x)
per stimare i coefficienti.

Predico y, utilizzando i coefficienti
stimati in precedenza.

Metodi di Regressione Multivariata

Approcci possibili – PCR e PLS

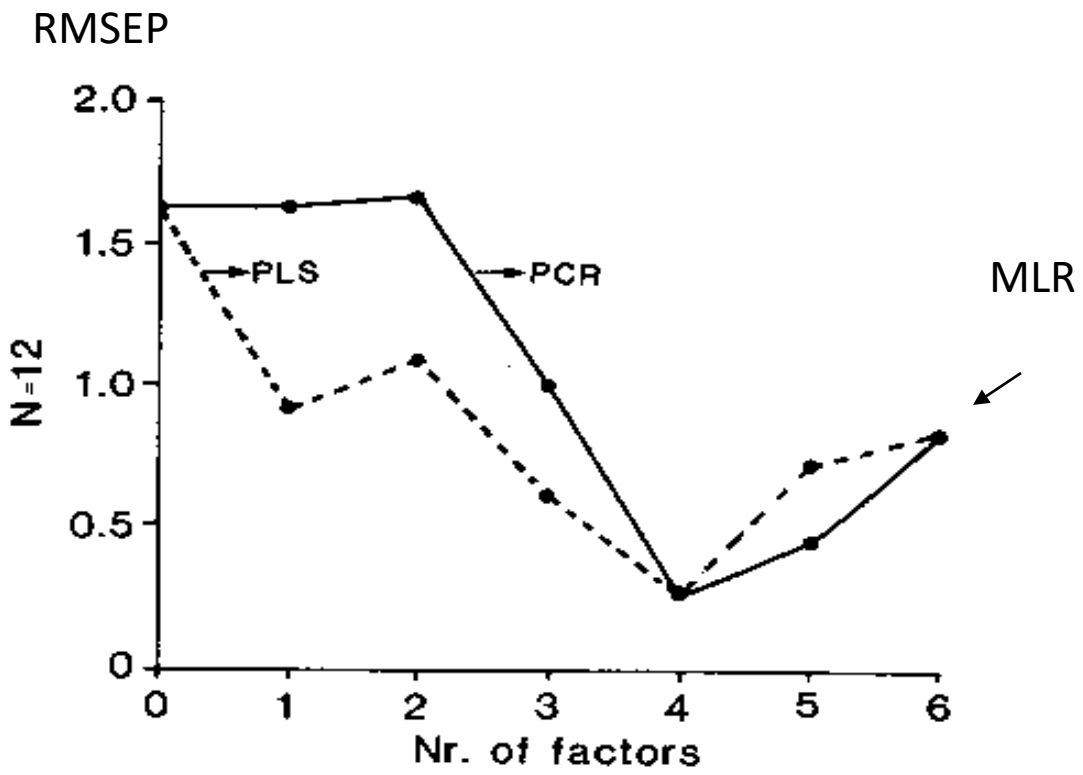
- Calcolo $$RMSEP = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N}$$

- Metto in grafico RMSEP rispetto alle componenti
    - Scelgo il numero di componenti che mi fornisce il miglior RMSEP (compromesso)

- Effettuo una comparazione di metodi diversi

Metodi di Regressione Multivariata

Approcci possibili – PCR e PLS



Calibrazione NIR di proteine nel grano. 6 componenti possibili calcolate.
26 campioni di calibarazione, 12 campioni di test

Metodi di Regressione Multivariata

Approcci possibili – PCR e PLS