Cluster Analysis

- è un insieme di tecniche esplorative che mirano a raggruppare le unità statistiche di una popolazione sulla base della loro similarità in termini di valori assunti dalle variabili osservate
- idealmente si vorrebbe partizionare la popolazione in modo che unità facenti parte dello stesso gruppo siano fra loro molto simili mentre unità facenti parte di gruppi diversi siano fra loro molto dissimili

Public utility data (1975)

Johnson & Wichern (2002, Chapter 12)

X_1: fixed-charge coverage ratio (income/debt)

X_2: rate of return on capital

X_3: cost of KW capacity in place

X_4: annual load factor

X_5: peak kWh demand growth from 1974 to 1975

X_6: sales (kWh use per year)

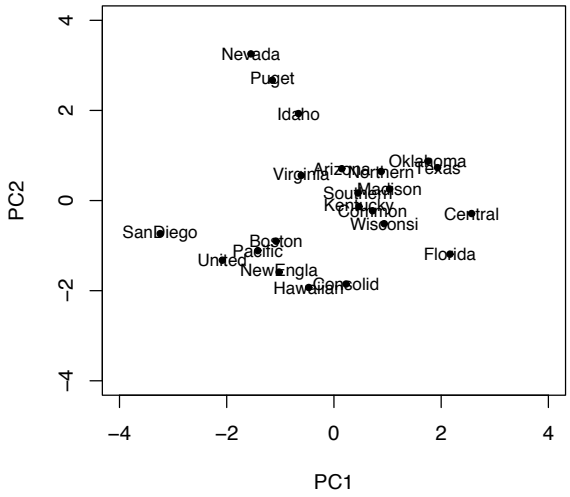
X_7: percent nuclear

X_8: total fuel costs (cents per kWh)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Arizona	1.06	9.2	151	54.4	1.6	9077	0.0	0.628
Boston	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
Central	1.43	15.4	113	53.0	3.4	9212	0.0	1.058
Common	1.02	11.2	168	56.0	0.3	6423	34.3	0.700
Consolid	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
Florida	1.32	13.5	111	60.0	−2.2	11127	22.5	1.241
Hawaiian	1.22	12.2	175	67.6	2.2	7642	0.0	1.652
Idaho	1.10	9.2	245	57.0	3.3	13082	0.0	0.309
Kentucky	1.34	13.0	168	60.4	7.2	8406	0.0	0.862
Madison	1.12	12.4	197	53.0	2.7	6455	39.2	0.623
Nevada	0.75	7.5	173	51.5	6.5	17441	0.0	0.768
NewEngla	1.13	10.9	178	62.0	3.7	6154	0.0	1.897
Northern	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
Oklahoma	1.09	12.0	96	49.8	1.4	9673	0.0	0.588
Pacific	0.96	7.6	164	62.2	−0.1	6468	0.9	1.400
Puget	1.16	9.9	252	56.0	9.2	15991	0.0	0.620
SanDiego	0.76	6.4	136	61.9	9.0	5714	8.3	1.920
Southern	1.05	12.6	150	56.7	2.7	10140	0.0	1.108
Texas	1.16	11.7	104	54.0	−2.1	13507	0.0	0.636
Wisconsi	1.20	11.8	148	59.9	3.5	7287	41.1	0.702
United	1.04	8.6	204	61.0	3.5	6650	0.0	2.116
Virginia	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

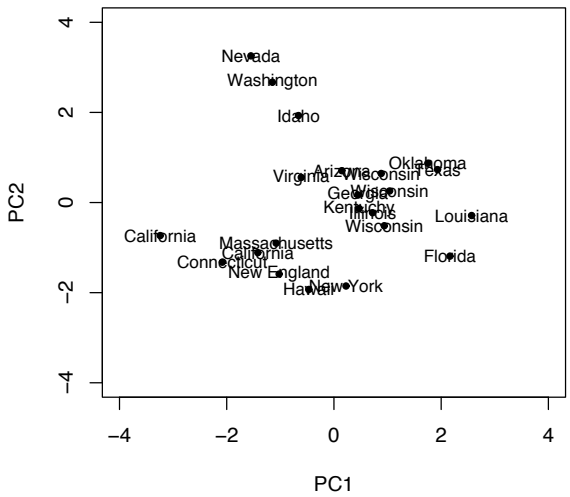
ID	COMPANY	STATE
Arizona	Arizona Public Service	Arizona
Boston	Boston Edison Co.	Massachusetts
Central	Central Louisiana Electric Co.	Louisiana
Common	Commonwealth Edison Co.	Illinois
Consolid	Consolidated Edison Co. (N.Y.)	New York
Florida	Florida Power & Light Co.	Florida
Hawaiian	Hawaiian Electric Co.	Hawaii
Idaho	Idaho Power Co.	Idaho
Kentucky	Kentuchy Utilities Co.	Kentuchy
Madison	Madison Gas & Electric Co.	Wisconsin
Nevada	Nevada Power Co.	Nevada
NewEngla	New England Electric Co.	New England
Northern	Northern States Power Co.	Wisconsin
Oklahoma	Oklahoma Gas & Electric Co.	Oklahoma
Pacific	Pacific Gas & Electric Co.	California
Puget	Puget Sound Power & Light Co.	Washington
SanDiego	San Diego Gas & Electric Co.	California
Southern	The Southern Co.	Georgia
Texas	Texas Utilities Co.	Texas
Wisconsi	Wisconsin Electric Power Co.	Wisconsin
United	United Illuminating Co.	Connecticut
Virginia	Virginia Electric & Power Co.	Virginia

Considerando le prime due componenti principali...



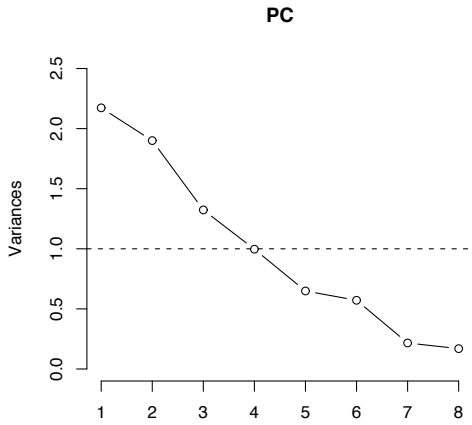
...si individuano "a occhio" tre o quattro cluster...

...e concentrando l'attenzione sugli stati sede...



...si perviene a una spiegazione "geografica" dei cluster

Ma due componenti principali sono sufficienti?




```
Importance of components:
                          PC1   PC2   PC3   PC4    PC5    PC6    PC7    PC8
Standard deviation     1.474 1.379 1.150 0.998 0.8056 0.7561 0.4653 0.4116
Proportion of Variance 0.272 0.238 0.165 0.125 0.0811 0.0715 0.0271 0.0212
Cumulative Proportion  0.272 0.509 0.675 0.799 0.8803 0.9518 0.9788 1.0000
```


Sommario

Se vogliamo considerare quattro componenti principali, in modo da spiegare l'80% della varianza totale, oppure direttamente tutte le variabili osservate, occorre formalizzare la "metrica dell'occhio" mediante la quale abbiamo individuato dei cluster per via grafica

A tal fine, nel seguito, introdurremo e discuteremo

- misure di distanza/dissimlilarità fra unità statistiche
  (e fra cluster di unità statistiche)
- algoritmi per la formazione di cluster
  (e criteri per la scelta del numero di cluster)

Distanza euclidea

Se x e y sono due righe (unità statistiche) di una matrice di dati con p colonne (tutte variabili quantitative) la loro distanza euclidea è definita come

$$d_E(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}$$

Nel caso particolare in cui $p = 2$ si tratta dell'ordinaria distanza nel piano ed è chiaro che la "metrica dell'occhio" si appoggia su di essa; per esempio

	Nevada	Puget	Idaho	Virginia	Arizona
Nevada	0.00				
Puget	0.71	0.00			
Idaho	1.59	0.89	0.00		
Virginia	2.86	2.18	1.37	0.00	
Arizona	3.06	2.35	1.47	0.77	0.00

"Alcuni pensano, o re Gelone, che il numero dei granelli di sabbia sia infinito per grandezza..."

...scriveva Archimede (circa 287–212 a.C.) al sovrano di Siracusa nell'*Arenario*... per poi mostrare "attraverso dimostrazioni geometriche che tu potrai seguire" come 10^{63} granelli di sabbia riempissero l'Universo...e soprattutto come si potesse "esprimere" tale numero (`http://www2.polito.it/didattica/polymath/`); più recentemente Everitt (2005, Chapter 6) ha calcolato...

Unità	Cluster	Possibili partizioni
15	3	2 375 101
20	4	45 232 115 901
25	8	690 223 721 118 368 580
100	5	10^{68}

...numeri che mostrano la necessità di algoritmi "un po' furbi"

Clustering gerarchico agglomerativo

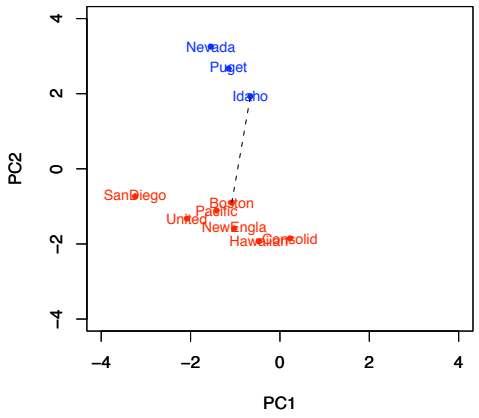
Se iniziamo formando un cluster con le due unità statistiche fra loro più vicine e mettendo ogni altra unità in un cluster a sé...

...possiamo pensare di procedere aggregando via via i cluster fra loro più vicini, finché non siamo "soddisfatti" del risultato...

...ma per farlo abbiamo bisogno innanzi tutto di definire una distanza fra cluster che riassuma le distanze fra i loro elementi:

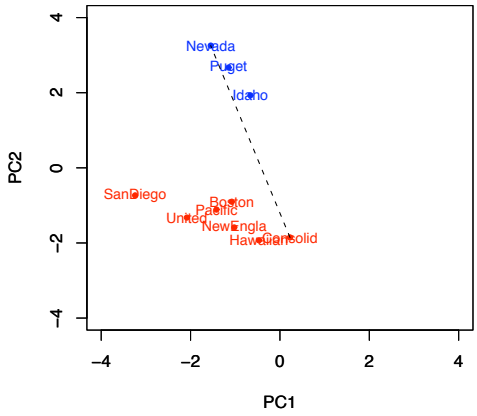
si parla di LINKAGE METHOD

Single linkage



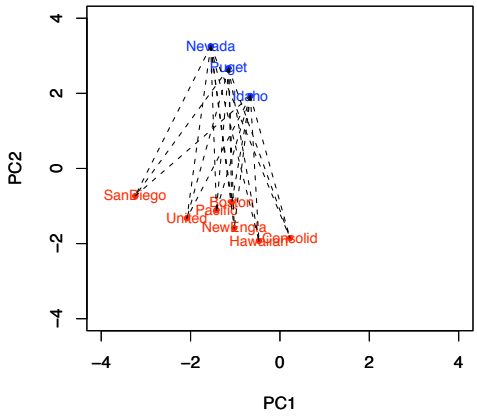
$$d(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y)$$

Complete linkage



$$d(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x, y)$$

Average linkage



$$d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{x \in C_1, y \in C_2} d(x, y)$$

Quanti cluster?

Non vi è una risposta univoca a questa domanda...

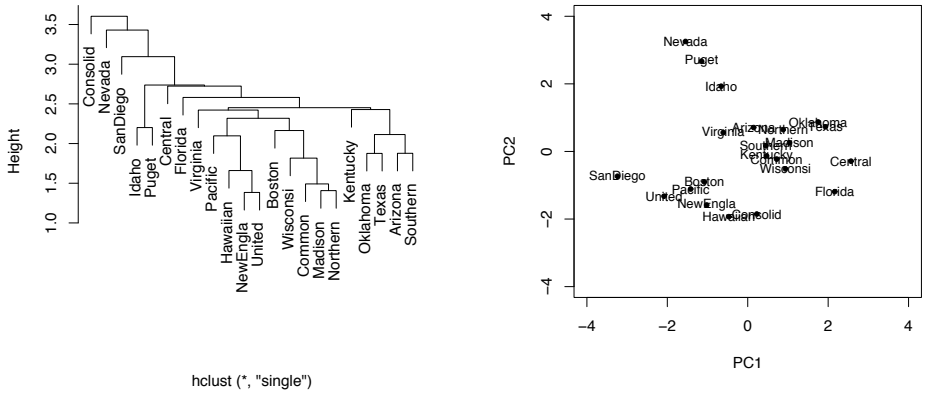
...infatti l'output di un algoritmo di clustering gerarchico agglomerativo è "semplicemente" un dendogramma che mostra la successione delle aggregazioni, con le distanze alle quali esse avvengono, fino al raggruppamento di tutte le unità statistiche in un unico cluster...

...dopo di che si sceglie il numero di cluster analizzando il dendogramma alla luce di considerazioni diverse

Nella scelta del numero di cluster giocano un ruolo:

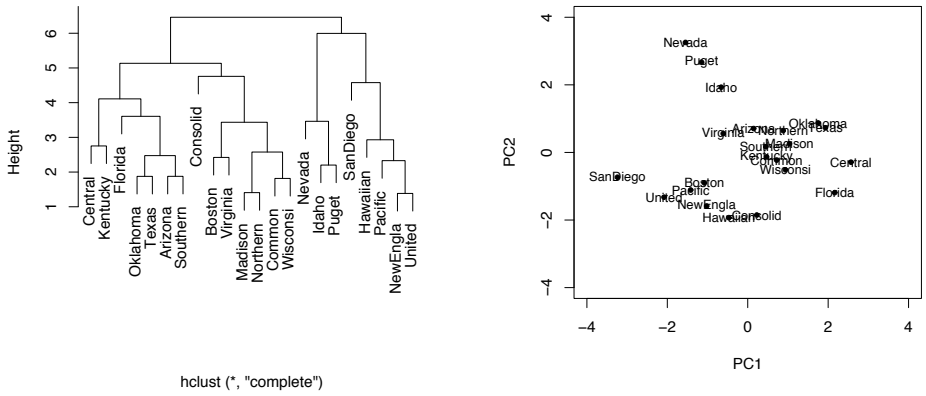
- le conoscenze specifiche sulla popolazione studiata e in particolare la possibile interpretazione dei cluster
- l'uso che dei cluster si vuole fare e in particolare la loro accessibilità da un punto di vista operativo
- l'esigenza di evitare cluster di numerosità troppo esigua (che tuttavia possono essere rivelatori di unità anomale)
- l'aspirazione ad avere omogeneità all'interno dei cluster e disomogeneità fra cluster diversi (che porta a "tagliare" il dendogramma a un'altezza intorno alla quale, per quanto possibile, non si realizzino aggregazioni)

Single linkage
(distanza euclidea)



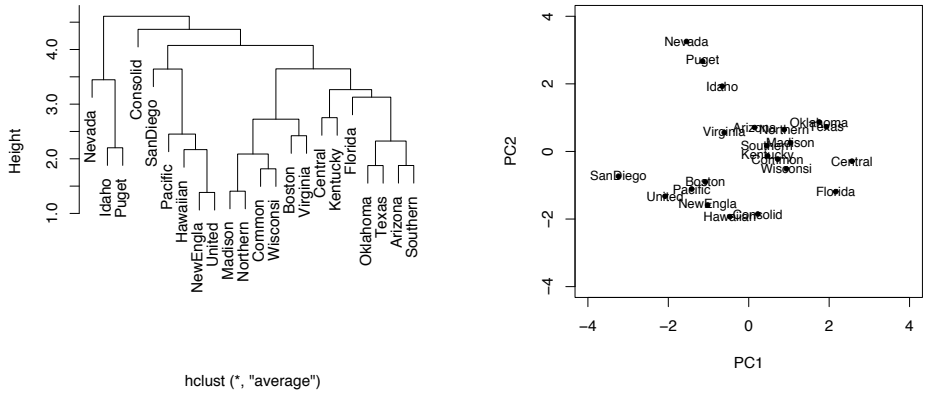
clustering gerarchico agglomerativo
su tutte le variabili standardizzate

Complete linkage
(distanza euclidea)



clustering gerarchico agglomerativo
su tutte le variabili standardizzate

Average linkage
(distanza euclidea)



clustering gerarchico agglomerativo
su tutte le variabili standardizzate

Tre o quattro cluster

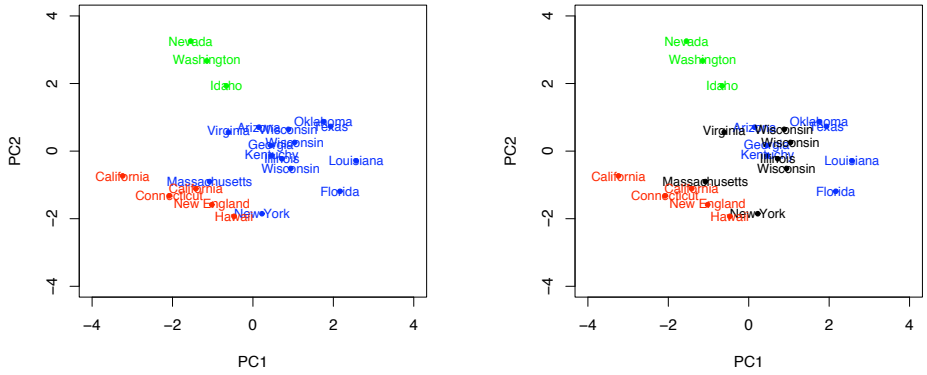
Il dendogramma relativo al metodo complete linkage offre dei "punti di taglio" migliori rispetto agli altri due, sia per la lunghezza dei tratti verticali (cluster disomogenei), sia per il bilanciamento (delle numerosità) dei cluster che si ottengono

Possiamo optare per tre cluster ben distanziati fra loro oppure preferire una soluzione con quattro cluster meglio bilanciati:

```
[[1]] "Arizona"  "Boston"   "Central"  "Common"   "Consolid" "Florida"  "Kentucky"
      "Madison"  "Northern" "Oklahoma" "Southern" "Texas"    "Wisconsi" "Virginia"
[[2]] "Hawaiian" "NewEngla" "Pacific"  "SanDiego" "United"
[[3]] "Idaho"    "Nevada"   "Puget"

[[1]] "Arizona"  "Central"  "Florida"  "Kentucky" "Oklahoma" "Southern" "Texas"
[[2]] "Boston"   "Common"   "Consolid" "Madison"  "Northern" "Wisconsi" "Virginia"
[[3]] "Hawaiian" "NewEngla" "Pacific"  "SanDiego" "United"
[[4]] "Idaho"    "Nevada"   "Puget"
```

Per quanto riguarda l'interpretazione...



...con 4 cluster distinguiamo tra utility "del nord-est" e utility "del sud", ferme restando quelle "del nord-ovest" e "sulla costa"

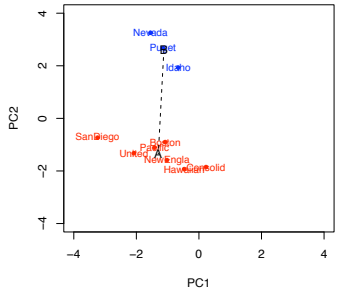
Pro e contro...

...dei diversi metodi di linkage:

- il single linkage ha il pregio di riuscire a individuare dei cluster di forma allungata invece che ellittica (tendenza al chaining)
- d'altra parte, per lo stesso motivo, il single linkage ha il difetto di non riuscire a individuare cluster ellittici che siano "contigui"
- sia il single linkage che il complete linkage forniscono risultati invarianti rispetto a trasformazioni monotone delle distanze
- di questa proprietà non gode l'average linkage
- l'average linkage è una soluzione di compromesso

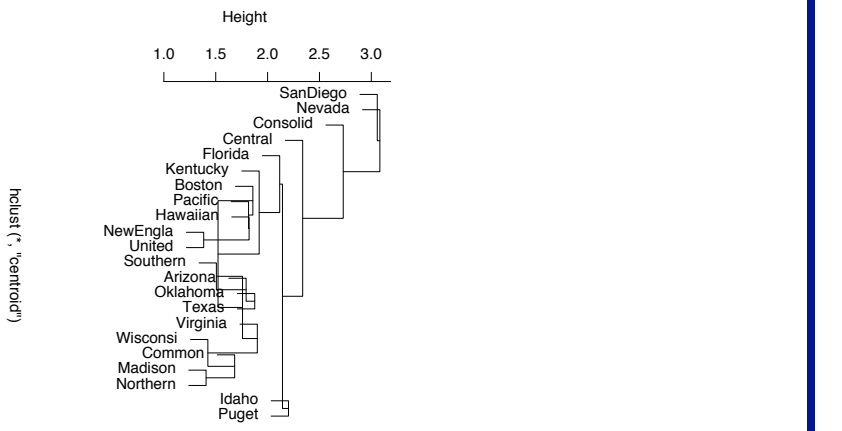
Metodo del "centroide"

Si tratta di un metodo di linkage alternativo nel quale ogni cluster è rappresentato del proprio baricentro, ovvero dal vettore delle medie calcolate nel cluster...



...dopo di che la distanza fra due cluster è la distanza fra i baricentri che li rappresentano (fra i rispettivi centroidi)

Il problema delle "inversioni"



Con il metodo del centroide non è più vero che le aggregazioni avvengono a distanza via via maggiore e quindi non si può più dire che si "taglia" il dendogramma a una certa altezza

Altre distanze

Una generalizzazione della distanza euclidea
è la distanza di Minkowski di ordine m:

$$d_m(x,y) = [(x_1 - y_1)^m + \cdots + (x_p - y_p)^m]^{\frac{1}{m}}, \quad m \geq 1$$

Per $m = 1$ si parla anche di distanza di Manhattan
(o "city-block" distance)

$$d_1(x,y) = |x_1 - y_1| + \cdots + |x_p - y_p|$$

mentre la distanza euclidea si ritrova per $m = 2$ ($d_2 = d_E$)

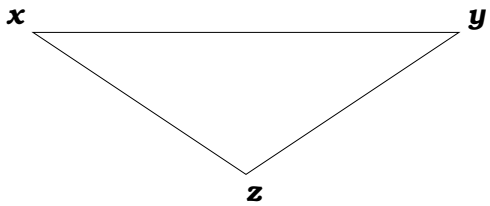
Al limite per $m \to \infty$ si trova invece la distanza del massimo:

$$d_\infty(x,y) = \max\{|x_1 - y_1|, \ldots, |x_p - y_p|\}$$

Proprietà delle distanze

In generale si dice distanza una qualsiasi funzione d tale che

- $d(x,x) = 0$ & $d(x,y) > 0$ per $x \neq y$ (definita positiva)
- $d(x,y) = d(y,x)$ (simmetrica)
- $d(x,y) \leq d(x,z) + d(z,y)$ (disuguaglianza triangolare)



Misure di dissimilarità

Una funzione d che sia simmetrica e definita positiva, senza necessariamente soddisfare la disuguaglianza triangolare, si dice dissimilarità (o "distanza non metrica")

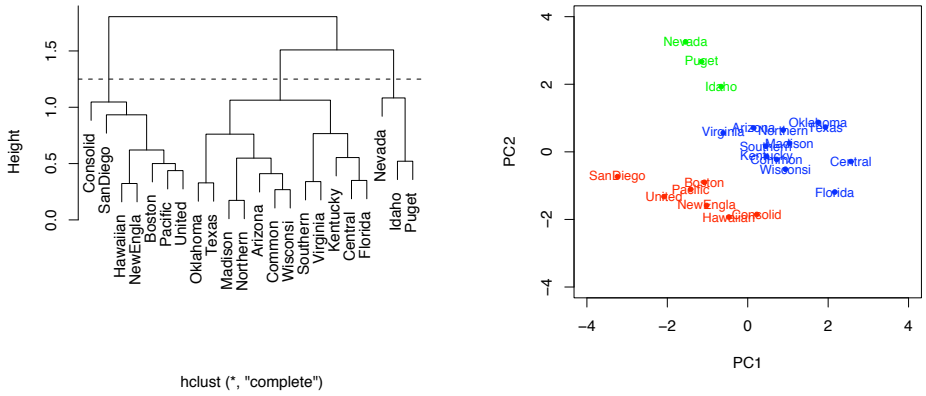
Un esempio è la cosiddetta distanza di Canberra:

$$d_C(x, y) = \frac{|x_1 - y_1|}{x_1 + y_1} + \cdots + \frac{|x_p - y_p|}{x_p + y_p}$$

(così) definita per variabili (strettamente) positive

La dissimilarità d_C interessa per la sua caratteristica di essere sensibile alle differenze relative piuttosto che a quelle assolute

Canberra "distance"



Clustering gerarchico agglomerativo sulle variabili X_1, X_2, X_3, X_4, X_6, X_8 (complete linkage)

Dissimilarità nel caso di variabili categoriali

Considerando per semplicità variabili dicotomiche, diciamo a valori in $\{0, 1\}$, conviene fare riferimento alla tabella di match/mismatch

		y		
		1	0	Totals
x	1	a	b	a+b
	0	c	d	c+d
Totals		a+c	b+d	p=a+b+c+d

Una possibilità è prendere la frequenza di mismatch $d = (b + c)/p$ (proporzionale al quadrato della distanza euclidea)

Oppure si può prendere la distanza binaria asimmetrica $d = (b + c)/(p - d)$ (due unità sono simili quando hanno una caratteristica in comune e si ignora il caso in cui questa manchi a tutte e due (es. conoscenza di una lingua straniera)

Altre misure di dissimilarità

- può capitare che i dati consistano direttamente in una matrice di dissimilarità (Everitt, 2005, Chapter 5, presenta un esempio in cui le dissimilarità tra 10 bevande sono state espresse da un intervistato su una scala da 0 a 100)
- se si vogliono raggruppare le variabili invece che le unità statistiche, si può costruire una matrice di dissimilarità prendendo i complementi a uno dei valori assoluti dei coefficienti di correlazione lineare (potrebbe essere un'utile analisi preliminare all'analisi fattoriale)

Altri algoritmi di clustering

- clustering gerarchico agglomerativo con il metodo di Ward
- clustering gerarchico divisivo
- clustering non gerarchico
  - metodo delle K-medie
  - model based clustering (dove si specifica un modello mistura per le variabili osservate, se ne stimano i parametri mediante il cosiddetto "algoritmo EM" e si ottengono "gratis" le probabilità di appartenenza di ogni unità ai vari cluster; si rinvia a Fraley & Raftery, 2002, per approfondimenti)

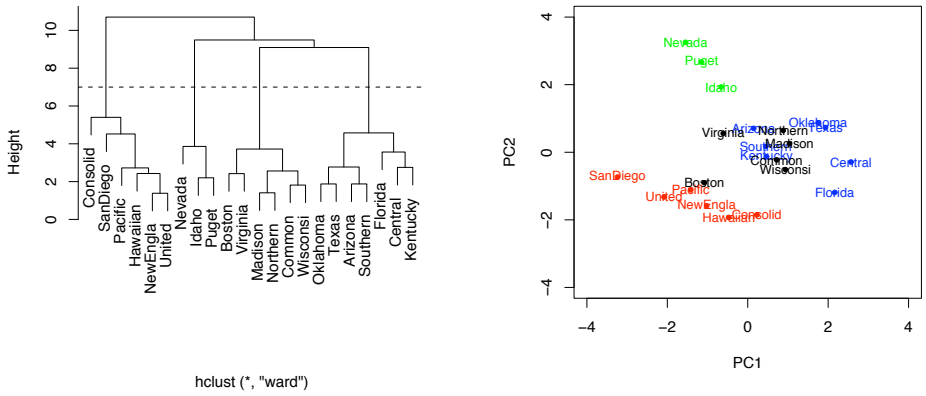
Clustering con il metodo di Ward

- si inizia formando tanti cluster quante sono le unità statistiche
- per ogni cluster c si considera la somma SS_c dei quadrati degli scarti delle unità di c dal centroide di c (all'inizio 0)
- si procede unendo due cluster alla volta in modo da minimizzare l'incremento della quantità $WSS = \sum_c SS_c$ (Within cluster Sum of Squares)
- si termina quando tutte le unità formano un unico cluster

$\Rightarrow$ dendogramma

il metodo di Ward è per sua natura particolarmente adatto per individuare cluster di forma ellittica

Metodo di Ward



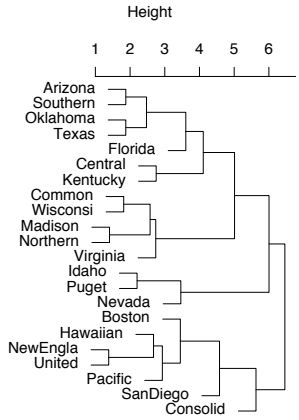
clustering gerarchico agglomerativo
su tutte le variabili standardizzate

Clustering gerarchico divisivo

- si inizia formando un unico cluster con tutte le unità statistiche
- il cluster iniziale viene diviso in due cluster quanto più possibile "distanti" fra loro (non approfondiamo)
- si procede dividendo in due un cluster alla volta (come sopra)
- si termina quando ogni unità statistica forma un cluster a sé

$\Rightarrow$ dendogramma (da leggere "al roverscio")

Dendogramma gerarchico divisivo



(distanza euclidea, tutte le variabili standardizzate)

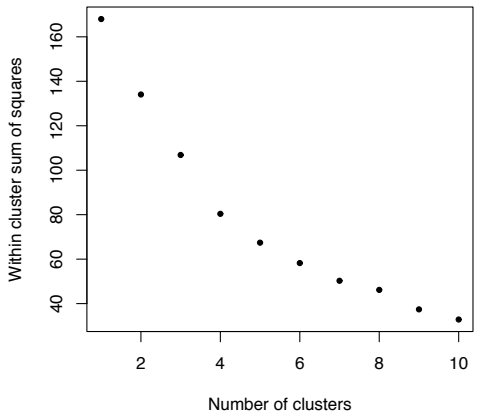
Clustering con il metodo delle K-medie

Si cerca di raggruppare le unità statistiche in K cluster di modo che risulti minima la somma WSS dei quadrati degli scarti di ogni unità dal centroide del proprio cluster

Nella configuarzione cercata, ogni cluster è formato dalle unità statistiche più vicine (in distanza euclidea) al proprio centroide

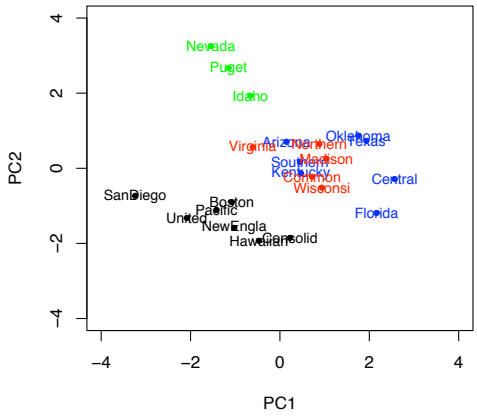
Il numero K di cluster è fissato a priori, ma nulla vieta e anzi è sempre consigliabile provare con diversi valori di K; un grafico del minimo di WSS in funzione di K può poi aiutare a scegliere il numero effettivo di cluster...

WSS minimizzata in funzione di K



il quarto cluster individua un "gomito" e con quattro cluster il minimo di WSS è dimezzato

Metodo delle "4-medie"



clustering su tutte le variabili standardizzate

A proposito delle K-medie

- l'algoritmo classico per le K-medie è un algoritmo iterativo che, scelti "a caso" K (candidati) centroidi, alterna i passi
  - assegna ogni unità al centroide più vicino
  - ricalcola i centroidi come medie di cluster

  finché nessuna unità cambia più cluster di appartenenza
- il metodo di Ward può vedersi come una variante gerarchica del metodo delle K-medie; entrambi, in quanto basati su WSS, sono particolarmente adatti per individuare cluster ellittici

Riferimenti

R. A. Johnson & D. W. Wichern (2002). Applied Multivariate Statistical Analysis. Prentice-Hall, Upper Saddle River, NJ.

B. Everitt (2005). An R and S-PLUS® Companion to Multivariate Analysis. Springer-Verlag, London.

C. Fraley & A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97, 611–631.

http://www-dimat.unipv.it/luca/

mailto://larocca.luca@unimore.it