

PCA: Applicazioni

Analisi Esplorativa

Aldo Solari



① Dati Marks

② Dati Wine

③ Dati Face



Outline

① **Dati Marks**

② Dati Wine

③ Dati Face



Dati Marks

Studente	Mechanics CB	Vectors CB	Algebra OB	Analysis OB	Statistics OB
1	77	82	67	67	81
2	63	78	80	70	81
3	75	73	71	66	81
4	55	72	63	70	68
5	63	63	65	70	63
6	53	61	72	64	73
7	51	67	65	65	68
8	59	70	68	62	56
9	62	60	58	62	70
⋮	⋮	⋮	⋮	⋮	
88	0	40	21	9	14



Dati Marks: Analisi delle componenti principali

- Domanda di interesse: come descrivere in maniera sintetica (i.e. in $q < p = 5$ dimensioni) i voti di ciascun studente?
- Calcolare la matrice degli autovettori standardizzati $V_{5 \times 5}$ di $S_{5 \times 5}$
- Calcolare le $p = 5$ componenti principali

$$Y_{88 \times 5} = \tilde{X}_{88 \times 5} V_{5 \times 5}$$



Dati Marks: matrice dei pesi V

	v_1	v_2	v_3	v_4	v_5
Mechanics	-0.51	-0.75	0.30	-0.30	-0.08
Vectors	-0.37	-0.21	-0.42	0.78	-0.19
Algebra	-0.35	0.08	-0.15	0.00	0.92
Analysis	-0.45	0.30	-0.60	-0.52	-0.29
Statistics	-0.53	0.55	0.60	0.18	-0.15

- I pesi (*loadings*) v_1 della prima componente principale sono più o meno omogenei, quindi il vettore dei punteggi (*scores*)
 $y_1 = \tilde{X} v_1$ della prima componente principale sarà più o meno la media dei voti (centrati)
 $88 \times 1 \quad 88 \times 5 \quad 5 \times 1$
- I pesi v_2 della seconda componente principale sono concentrati sulle variabili Mechanics (-0.75) e Statistics (0.55). Hanno segno negativo gli esami *closed book* (CB), segno positivo gli esami *open book* (OB)



Dati Marks: matrice dei punteggi Y

Studente	y_1	y_2	y_3	y_4	y_5
1	-66.32	-6.45	7.07	9.65	-5.46
2	-63.62	6.75	0.86	9.15	7.57
3	-62.93	-3.08	10.23	3.72	0.38
4	-44.54	5.58	-4.38	4.48	-4.41
5	-43.28	-1.13	-1.53	-5.81	-0.74
6	-42.55	10.97	4.87	0.48	7.10
7	-39.11	8.26	-0.81	4.35	0.13
8	-37.53	-5.60	-5.50	3.78	4.37
9	-39.39	1.13	9.41	-2.51	-5.33
\vdots	\vdots	\vdots	\vdots	\vdots	
88	65.96	2.27	2.52	17.70	-7.22

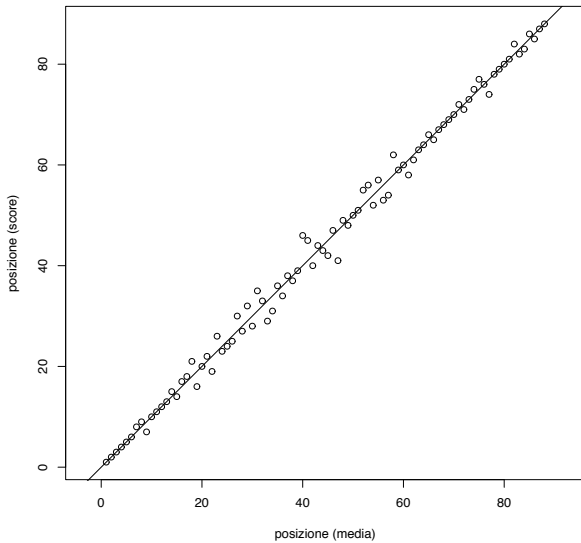


Punteggi della prima componente principale

$$\text{Score}_i = 0.51 \cdot \text{Mechanics}_i + 0.37 \cdot \text{Vectors}_i + 0.35 \cdot \text{Algebra}_i + 0.45 \cdot \text{Analysis}_i + 0.53 \cdot \text{Statistics}_i$$

	media	posizione	score	posizione (score)
1	74.80	1	66.30	1
2	74.40	2	63.60	2
3	73.20	3	62.90	3
4	65.60	4	44.50	4
5	64.80	5	43.30	5
6	64.60	6	42.60	6
7	63.20	7	39.10	8
8	63.00	8	37.50	9
9	62.40	9	39.40	7
10	60.60	10	32.10	10
...				
88	16.80	88	-66.00	88



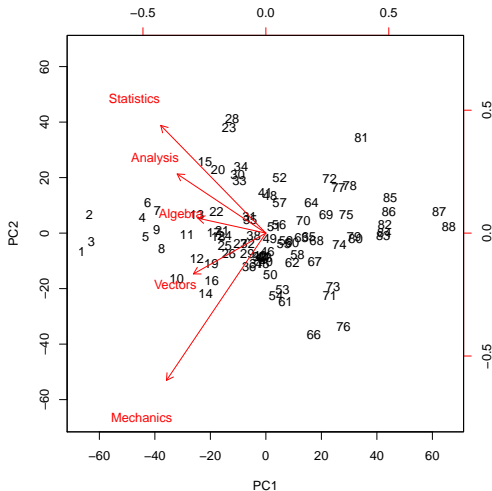


Rappresentazione grafica: *biplot*

- Il *biplot* è una rappresentazione grafica bidimensionale dei punteggi (n punti) e dei pesi (p autovettori non standardizzati) della prime due componenti principali
- Permette l'ispezione visiva della posizione di ciascuna unità statistica e di ciascuna variabile nello spazio delle prime due componenti principali



Dati Marks: *biplot*



Punto "Studente 3" : $(y_{31}, y_{32}) = (-62.93, -3.08)$
Vettore "Algebra" : direzione $(v_{31}, v_{32})' = (-0.35, 0.08)'$,
lunghezza \propto alla varianza



Scelta del numero di componenti principali

- Scegliere le prime q componenti con q pari al valore minimo tale per cui la proporzione di varianza spiegata cumulata dalle prime $q \leq p$ componenti principali

$$\frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j}$$

sia superiore a una prefissata percentuale, generalmente dell'ordine di 70, 80%, dove la soglia può essere diminuita qualora p sia molto grande



Scelta del numero di componenti principali

- Ignorare le componenti principali che spiegano un ammontare di varianza inferiore a un livello prefissato c
- Una scelta tipica è

$$c = \frac{1}{p} \sum_{j=1}^p \lambda_j$$

- Procedendo in questo modo non è predeterminata la percentuale di varianza spiegata in totale



Scree plot

- Scegliere q esaminando il diagramma *scree* (*scree plot*), cioè la rappresentazione sul piano cartesiano di (j, λ_j)
- Si cerca di selezionare q in corrispondenza a un gomito del grafico, cioè un punto tale per cui gli autovalori precedenti sono 'grandi' e quelli successivi 'piccoli'.
- Chiaramente è ben possibile che tale grafico non offra alcuna indicazione (se ad esempio gli autovalori λ_j decrescono linearmente con j).



Dati Marks: varianza spiegata

	PC1	PC2	PC3	PC4	PC5
Varianza spiegata λ_j	679.18	199.81	102.57	83.67	31.79
Prop. di var. spiegata	0.6191	0.1821	0.0935	0.0763	0.0290
Proporzione cumulata	61.91%	80.13%	89.48%	97.10%	100%

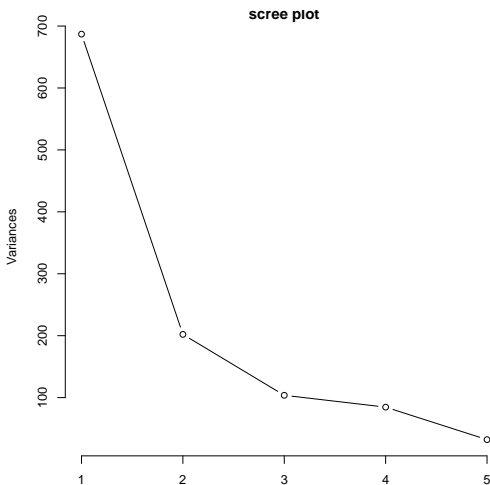
- Se vogliamo spiegare almeno l'80% della variabilità
→ Prime due componenti principali
- Se ignoriamo le componenti con varianza spiegata inferiore a

$$c = \frac{1}{5} \sum_{j=1}^5 \lambda_j = 219.4$$

→ Prima componente principale



Dati Marks: *scree plot*



Il 'gomito' indica le prime due componenti principali



Outline

① Dati Marks

② **Dati Wine**

③ Dati Face



Dati Wine

	Alcohol	MalicAcid	Ash	AlcAsh	Mg	Phenols	Flav	:
1	14.23	1.71	2.43	15.60	127	2.80	3.06	:
2	13.20	1.78	2.14	11.20	100	2.65	2.76	:
3	13.16	2.36	2.67	18.60	101	2.80	3.24	:
4	14.37	1.95	2.50	16.80	113	3.85	3.49	:
5	13.24	2.59	2.87	21.00	118	2.80	2.69	:
6	14.20	1.76	2.45	15.20	112	3.27	3.39	:
:	:	:	:	:	:	:	:	:



Dati Wine: varianze

j	s_{jj}
Alcohol	0.66
MalicAcid	1.24
Ash	0.07
AlcAsh	11.09
Mg	202.84
Phenols	0.39
Flav	0.99
NonFlavPhenols	0.02
Proa	0.33
Color	5.34
Hue	0.05
OD	0.50
Proline	98609.60

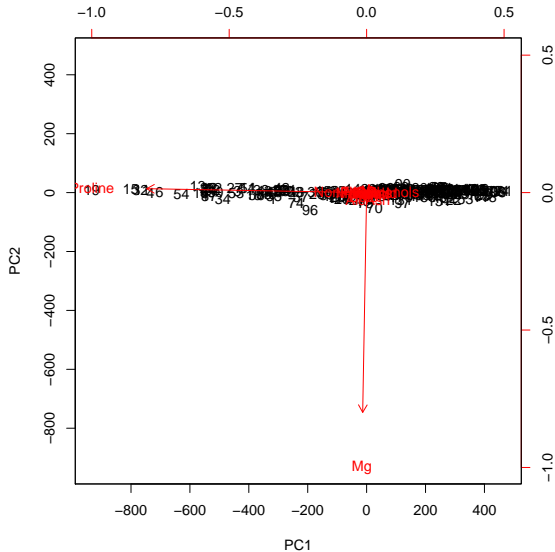


Dati Wine: pesi per PCA(\tilde{X})

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	
Alcohol	-0.00	-0.00	0.02	-0.14	0.02	-0.19	0.92	0
MalicAcid	0.00	-0.00	0.12	-0.16	-0.61	-0.74	-0.15	0
Ash	-0.00	-0.00	0.05	0.01	0.02	-0.04	0.05	0
AlcAsh	0.00	-0.03	0.94	0.33	0.06	0.02	0.03	-0
Mg	-0.02	-1.00	-0.03	0.01	-0.01	0.00	0.00	0
Phenols	-0.00	-0.00	-0.04	0.07	0.32	-0.28	-0.02	0
Flav	-0.00	0.00	-0.09	0.17	0.52	-0.43	-0.04	0
NonFlavPhenols	0.00	0.00	0.01	-0.01	-0.03	0.02	-0.00	-0
Proa	-0.00	-0.01	-0.02	0.05	0.25	-0.24	-0.31	-0
Color	-0.00	-0.02	0.29	-0.88	0.33	-0.00	-0.11	0
Hue	-0.00	0.00	-0.03	0.06	0.05	0.02	0.03	0
OD	-0.00	0.00	-0.07	0.18	0.26	-0.29	0.10	0
Proline	-1.00	0.02	0.00	0.00	-0.00	0.00	-0.00	-0



Dati Wine: *biplot* per PCA(\tilde{X})

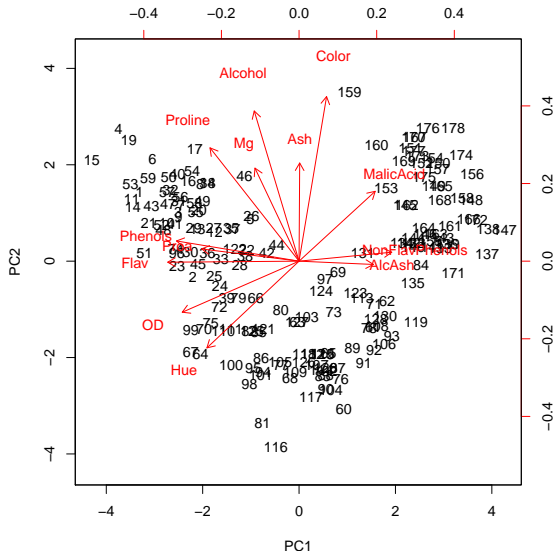


Dati Wine: pesi per PCA(Z)

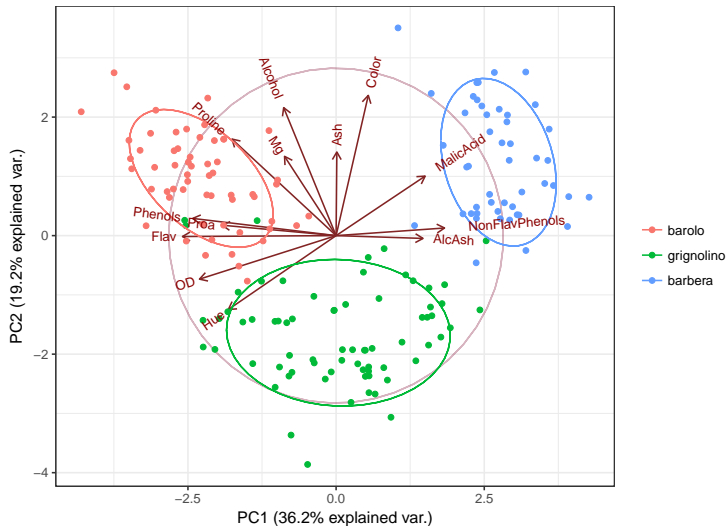
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	
Alcohol	-0.14	-0.48	-0.21	-0.02	0.27	0.21	0.06	0
MalicAcid	0.25	-0.22	0.09	0.54	-0.04	0.54	-0.42	0
Ash	0.00	-0.32	0.63	-0.21	0.14	0.15	0.15	-0
AlcAsh	0.24	0.01	0.61	0.06	-0.07	-0.10	0.29	0
Mg	-0.14	-0.30	0.13	-0.35	-0.73	0.04	-0.32	-0
Phenols	-0.39	-0.07	0.15	0.20	0.15	-0.08	0.03	-0
Flav	-0.42	0.00	0.15	0.15	0.11	-0.02	0.06	-0
NonFlavPhenols	0.30	-0.03	0.17	-0.20	0.50	-0.26	-0.60	-0
Proa	-0.31	-0.04	0.15	0.40	-0.14	-0.53	-0.37	0
Color	0.09	-0.53	-0.14	0.07	0.08	-0.42	0.23	-0
Hue	-0.30	0.28	0.09	-0.43	0.17	0.11	-0.23	0
OD	-0.38	0.16	0.17	0.18	0.10	0.27	0.04	-0
Proline	-0.29	-0.36	-0.13	-0.23	0.16	0.12	-0.08	0



Dati Wine: *biplot per PCA(Z)*



Dati Wine: tipologia di vino



Outline

① Dati Marks

② Dati Wine

③ **Dati Face**



Dati Face



X
 243×220



Dati Face: PCA

- Dati centrati: $\tilde{X}_{n \times p} = X_{n \times p} - \frac{1}{n} \bar{x}'_{1 \times p}$
- PCA: $Y_{n \times p} = \tilde{X}_{n \times p} V_{p \times p}$
- Scelta di $q \leq \text{rango}(\tilde{X}_{n \times p})$
- Ricostruzione dell'immagine: migliore approssimazione di rango q di \tilde{X} più vettore delle medie di X :

$$Y_q V_q' + \frac{1}{n} \bar{x}'_{1 \times p}$$



Immagine compressa



$$Y_q V_q' + \frac{1}{n \times 11 \times p} \bar{x}'$$

con $q = 10$



Pixels e bytes

Immagine originale

- $X_{243 \times 220}$: $243 \times 220 = 53460$ pixels
- Memoria richiesta: 427880 bytes

Immagine compressa

- $Y_{243 \times 10}, V_{220 \times 10}, \bar{x}_{220 \times 1}$: $243 \times 10 + 220 \times 10 + 220 = 4850$ pixels
- Memoria richiesta: 40872 bytes
- Fattore di riduzione = $427880 \text{ bytes} / 40872 \text{ bytes} = 10.47$

