

Capitolo 2

Analisi in componenti principali

2.1 Introduzione

L'analisi in componenti principali è una tecnica di analisi multivariata tra le più diffuse. Viene utilizzata quando nel dataset osservato sono presenti numerose variabili e si è interessati a ridurre la dimensionalità. L'analisi in componenti principali è assolutamente necessaria quando un certo carattere (o variabile) non è direttamente osservabile e quantificabile, per cui bisogna lavorare e analizzare diversi indicatori legati al quel carattere¹. Nella sostanza l'analisi in componenti principali porta alla creazione di nuove variabili, dette *Componenti Principali* appunto, che sono combinazioni lineari delle originarie, e che godono delle due seguenti proprietà:

1. sono tra loro incorrelate (ortogonali),
2. sono elencate in ordine decrescente della loro varianza,

Questa tecnica viene impiegata esclusivamente per variabili quantitative e chiaramente con un certo grado di correlazione.

Di seguito riportiamo un'analisi in componenti principali (d'ora in avanti abbreviata in acp).

¹Un esempio è la misurazione dell'intelligenza degli individui sulla base dei risultati di specifici test

2.2 Un'analisi in componenti principali

2.2.1 Descrizione del dataset

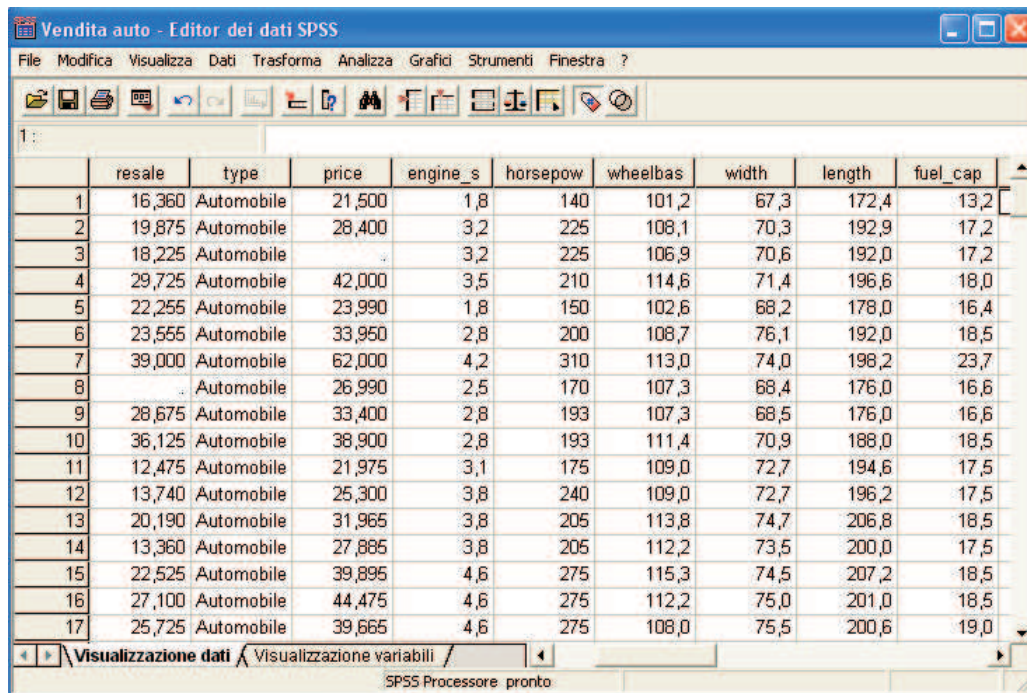
Il dataset che utilizzeremo riguarda l'osservazione di alcune variabili su automobili di diverse case di produzione e diversi modelli. L'analisi consiste nel valutare se è possibile predire le vendite delle vetture in base ad alcuni indicatori. Le variabili contenute nel dataset sono:

- *manufact*: il costruttore della vettura,
- *model*: il modello della vettura,
- *sales*: numero di vetture vendute,
- *resale*: numero di vetture vendute dopo quattro anni,
- *type*: tipo di vettura,
- *price*: prezzo di vendita
- *engine_s*: misura del motore,
- *horsepow*: potenza (in cavalli),
- *wheelbas*: il passo, distanza tra gli assi delle ruote,
- *width*: la larghezza della vettura,
- *length*: la lunghezza della vettura,
- *fuel_cap*: capacità del serbatoio,

Il dataset si trova nel file in formato *.sav* denominato *vendite auto*².

Una volta aperto, il file si presenta come in Figura 2.1,

²Il file è reperibile nella sezione 'Materiale Didattico' sul sito del corso di laurea in Statistica (www.economia.unical.it/statistica/)



	resale	type	price	engine_s	horsepow	wheelbas	width	length	fuel_cap
1	16,360	Automobile	21,500	1,8	140	101,2	67,3	172,4	13,2
2	19,875	Automobile	28,400	3,2	225	108,1	70,3	192,9	17,2
3	18,225	Automobile	.	3,2	225	106,9	70,6	192,0	17,2
4	29,725	Automobile	42,000	3,5	210	114,6	71,4	196,6	18,0
5	22,255	Automobile	23,990	1,8	150	102,6	68,2	178,0	16,4
6	23,555	Automobile	33,950	2,8	200	108,7	76,1	192,0	18,5
7	39,000	Automobile	62,000	4,2	310	113,0	74,0	198,2	23,7
8	.	Automobile	26,990	2,5	170	107,3	68,4	176,0	16,6
9	28,675	Automobile	33,400	2,8	193	107,3	68,5	176,0	16,6
10	36,125	Automobile	38,900	2,8	193	111,4	70,9	188,0	18,5
11	12,475	Automobile	21,975	3,1	175	109,0	72,7	194,6	17,5
12	13,740	Automobile	25,300	3,8	240	109,0	72,7	196,2	17,5
13	20,190	Automobile	31,965	3,8	205	113,8	74,7	206,8	18,5
14	13,360	Automobile	27,885	3,8	205	112,2	73,5	200,0	17,5
15	22,525	Automobile	39,895	4,6	275	115,3	74,5	207,2	18,5
16	27,100	Automobile	44,475	4,6	275	112,2	75,0	201,0	18,5
17	25,725	Automobile	39,665	4,6	275	108,0	75,5	200,6	19,0

Figura 2.1: Il dataset

2.2.2 La procedura in SPSS

Il pacchetto statistico SPSS non prevede una procedura specifica per effettuare l'acp, ma la racchiude tra la famiglia dei metodi fattoriali.

Il primo passo da compiere è un'attenta analisi della matrice dei dati iniziali che ci permette di avere una visione più ampia del dataset in esame. L'acp prevede una prima analisi di tipo descrittivo sulle variabili per cui si può evitare di effettuare queste analisi separatamente.

Dal menù *Analizza* selezionare *Riduzione dei dati* e di seguito *Fattoriale* (*Analizza* → *Riduzione dei dati* → *Fattoriale*).

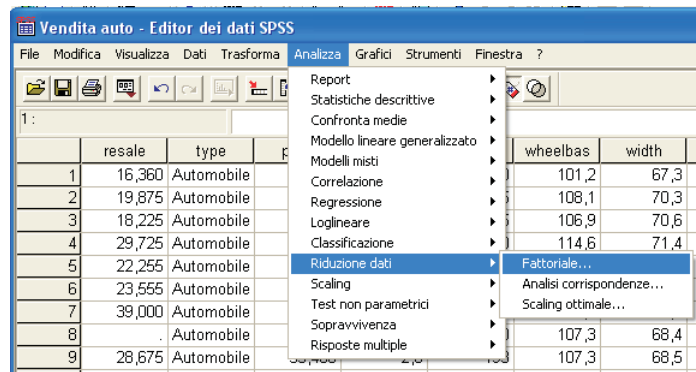


Figura 2.2: La procedura per avviare un'acp

Si apre la finestra mostrata in Figura 2.3, in cui vengono visualizzate le variabili che possono essere selezionate per effettuare l'analisi³.



Figura 2.3: La finestra di avvio della procedura

A questo punto, si procede selezionando le variabili⁴ presenti nell'elenco posto a sinistra e 'inviandole' nella parte destra della finestra (vedi Figura 2.4,

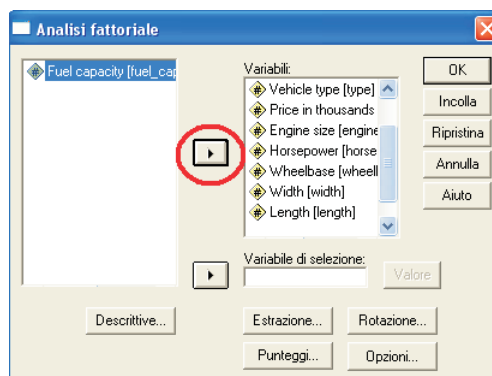


Figura 2.4: Selezione delle variabili

³Si tratta delle sole variabili di tipo quantitativo

⁴ATTENZIONE! Quali delle variabili presenti nel dataset vanno utilizzate per l'analisi?

Questa finestra mostra alcuni titoli, *Statistiche...*, *Estrazione...*, *Rotazione...*, *Punteggi...*, *Opzioni...* che andiamo a considerare nel dettaglio:

i. selezionando *Statistiche* si apre una nuova finestra che contiene un elenco di statistiche descrittive e selezioniamo

- Statistiche descrittive,
- Coefficienti per la matrice di correlazione,

clickare su *Continua*;

ii. selezionando *Estrazione* si apre una nuova finestra in cui selezioniamo come metodo di estrazione dei fattori *componenti principali*; inoltre scegliamo di condurre l'indagine dalla matrice di correlazione e di ottenere il grafico decrescente degli autovalori, clickare su *Continua*;

iii. selezionando *Rotazione* si ha la possibilità di scegliere il metodo di rotazione, nel nostro caso lasciamo la scelta di *default*, clickare su *Continua*;

iv. selezionando *Punteggi* si sceglie di salvarli come punteggi, clickare su *Continua* e infine su *Ok*.

Viene eseguita la procedura dell'acp e aperto il file di *output* su cui vengono visualizzate le tabelle che riportano i risultati dell'analisi (Figura 2.5).

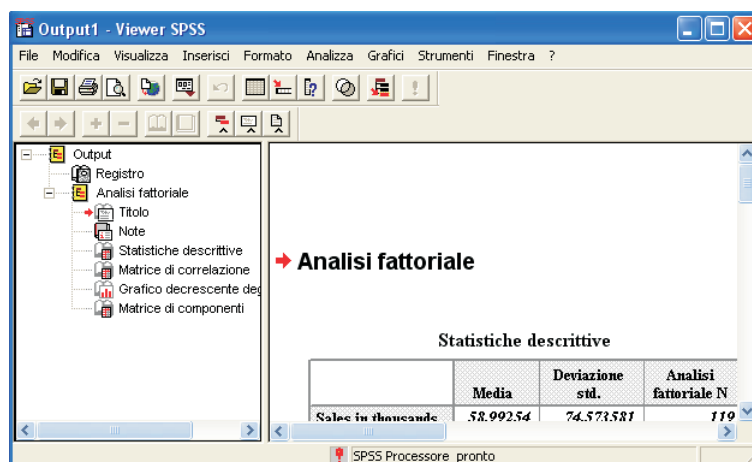


Figura 2.5: L'output

2.3 L'interpretazione dei risultati

Nelle Figure 2.6 e 2.7 sono riportate, rispettivamente le statistiche descrittive e la matrice di correlazione delle variabili in esame. Si lascia allo studente l'analisi accurata delle due tabelle e le osservazioni e riflessioni in merito⁵.

Statistiche descrittive			
	Media	Deviazione std.	Analisi fattoriale N
Sales in thousand	59,0	74,6	119
4-year resale value	18,1	11,5	119
Price in thousand	26,1	14,1	119
Engine size	3,1	1,1	119
Horsepower	182,2	58,8	119
Wheelbase	107,4	8,0	119
Width	71,3	3,5	119
Length	188,0	13,9	119
Fuel capacity	17,8	3,8	119

Figura 2.6: Tabella delle Statistiche Descrittive

Matrice di correlazione									
	Sales in thousands	4-year resale value	Price in thousands	Engine size	Horsepower	Wheelbase	Width	Length	Fuel capacity
Sales in thousands	1,000	-,279	-,257	,029	-,157	,404	,174	,268	,136
4-year resale value	-,279	1,000	,954	,531	,771	-,052	,179	,027	,326
Price in thousands	-,257	,954	1,000	,655	,854	,071	,306	,188	,406
Engine size	,029	,531	,655	1,000	,862	,410	,670	,536	,614
Horsepower	-,157	,771	,854	,862	1,000	,233	,514	,409	,476
Wheelbase	,404	-,052	,071	,410	,233	1,000	,678	,854	,656
Width	,174	,179	,306	,670	,514	,678	1,000	,748	,666
Length	,268	,027	,188	,536	,409	,854	,748	1,000	,556
Fuel capacity	,136	,326	,406	,614	,476	,656	,666	,556	1,000

Figura 2.7: La matrice di correlazione

I primi risultati che l'acp effettuata con SPSS produce sono mostrati nella tabella delle comunalità (Figura 2.8).

⁵Vedi la sezione Esercizi

Comunalità

	Iniziale	Estrazione
Sales in thousand	1,000	,422
4-year resale valu	1,000	,879
Price in thousand	1,000	,924
Engine size	1,000	,805
Horsepower	1,000	,901
Wheelbase	1,000	,863
Width	1,000	,772
Length	1,000	,813
Fuel capacity	1,000	,662

Figura 2.8: Tabella delle comunalità

In essa troviamo le quote di varianza di ciascuna variabile spiegate dalle *componenti principali* appena estratte. Per cui osserviamo che esse spiegano più del 90% della varianza delle variabili PREZZO,POTENZA , una percentuale superiore all'80% per le variabili PREZZO DOPO ANNI, DIMENSIONE DEL MOTORE, PASSO, LUNGHEZZA, mentre la variabile VENDITE è quella riprodotta meno bene dalle *componenti principali*, solo il 42% della varianza spiegata. Ma quante *componenti principali* abbiamo estratto? Questa informazione la troviamo nella tabella intitolata 'Totale varianza spiegata'(Figura 2.9) Questa tabella ci dice che la prima componente estratta, il cui autovalore è pari a 4,591, spiega il 51,014% della varianza totale e la seconda, il cui autovalore è pari a 2,450, un'ulteriore 27,227%, per in totale esse spiegano il 78,24% della varianza totale⁶.

Varianza totale spiegata

		Autovalori iniziali			Pesi dei fattori non ruotati		
		Totale	% di varianza	% cumulata	Totale	% di varianza	% cumulata
Componente	1	4,591	51,014	51,014	4,591	51,014	51,014
	2	2,450	27,227	78,241	2,450	27,227	78,241
	3	,713	7,925	86,165			
	4	,467	5,184	91,349			
	5	,374	4,159	95,508			
	6	,225	2,498	98,007			
	7	,090	1,001	99,008			
	8	,061	,674	99,682			
	9	,029	,318	100,000			

Metodo di estrazione: Analisi componenti principali.

Figura 2.9: Tabella della Varianza Spiegata

⁶Vedi Esercizio 1.2

Il numero di *componenti principali* che vengono ritenute sufficienti e utilizzabili può essere osservato nel grafico che riporta gli autovalori in funzione del numero delle componenti (da 1 a 9 poichè sono nove le variabili originarie). Il grafico, mostrato in Figura 2.10, si presenta come una spezzata sempre decrescente e ad un certo punto avviene una brusca variazione della pendenza (il gomito) che ci segnali il numero delle componenti da utilizzare. Nel nostro caso il gomito avviene tra la seconda e la terza componente, per cui sono le prime due le componenti da considerare⁷.

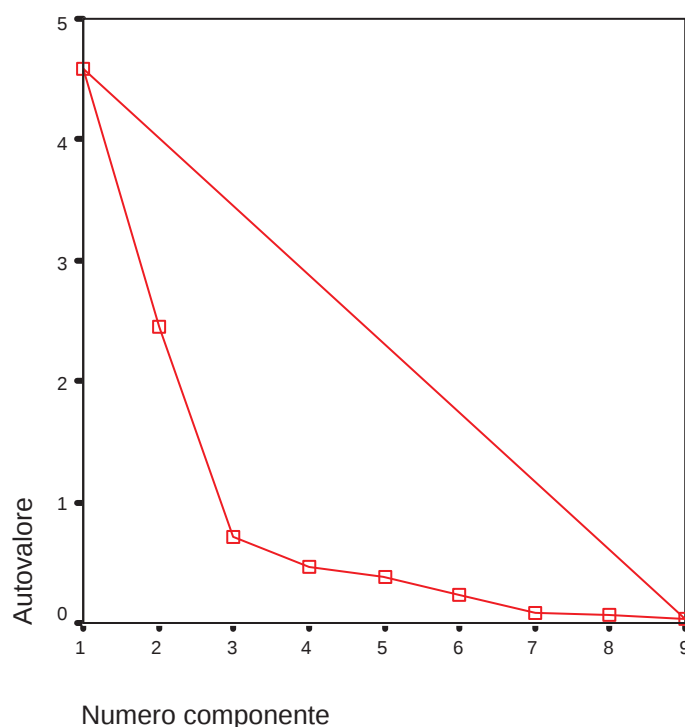


Figura 2.10: Grafico decrescente degli autovalori

La procedura dell'acp presente in SPSS non restituisce in maniera immediata gli autovettori corrispondenti alle *componenti principali*, ma fornisce in una tabella intitolata 'matrice delle comunalità', i coefficienti di correlazione tra ogni variabili e ogni componente estratta.

Per cui osserviamo che la prima componente risulta correlata positivamente (lo si deduce dal segno dei coefficienti) con tutte le variabili, ma questa relazione lineare è più *forte*

⁷Vedi Esercizio 1.3

con le variabili DIMENSIONE DEL MOTORE, POTENZA, mentre la seconda componente risulta correlata in maniera negativa e consistente con la variabile PREZZO DOPO 4 ANNI⁸.

Matrice di componenti

	Componente	
	1	2
Engine size	,890	-,110
Horsepower	,856	-,410
Width	,797	,370
Fuel capacity	,780	,233
Price in thousand	,730	-,626
Length	,715	,550
4-year resale valu	,611	-,711
Wheelbase	,633	,680
Sales in thousand	,063	,647

Figura 2.11: Tabella delle componenti

Per ottenere i coefficienti della combinazione lineare che determina le componenti possiamo ricorrere ai valori contenuti nella matrice delle componenti e ricordando che:

$$r(Y_i, X_j) = a_{i,j} \sqrt{\lambda_i}$$

dove Y_i è l' i -esima componente, X_j è la j -esima variabile e λ_i è l'autovettore corrispondente all' i -esima componente, possiamo determinare gli autovettori⁹.

Otteniamo che gli autovettori corrispondenti sono:

⁸Vedi Esercizio 1.4

⁹Si consiglia di calcolare gli autovettori utilizzando l'applicativo Excel di Microsoft

Variabile	a1j	12j
Sales in thousands	0,0294	0,4132
4-year resale value	0,2851	-0,4542
Price in thousands	0,3406	-0,3997
Engine size	0,4156	-0,0701
Horsepower	0,3994	-0,2622
Wheelbase	0,2954	0,4342
Width	0,3721	0,2361
Length	0,3337	0,3510
Fuel capacity	0,3640	0,1486

Tabella 2.1: Gli autovettori

Gli elementi di tali autovettori sono i coefficienti della combinazione lineare che definisce rispettivamente la prima e la seconda componente in funzione degli scostamenti standardizzati delle 9 variabili¹⁰.

Utilizzando questi valori si calcolano i punteggi con media nulla e varianza uguali a λ_1 e λ_2 per la prima e per la seconda componente, per ciascuna vettura presente nel dataset.

Infine SPSS produce il grafico delle componenti in cui i punti sono le variabili e le coordinate di ogni variabile sono i pesi fattoriali di ciascuna componente.

Per leggere in maniera corretta il grafico bisogna considerare che:

- sull'asse orizzontale troviamo la correlazione tra le variabili e la prima componente,
- sull'asse verticale troviamo la correlazione tra le variabili e la seconda componente,
- le variabili presenti nel grafico sono vettori che contengono i coefficienti di correlazione tra le due componenti e le variabili, tutti compresi in un cerchio di raggio unitario il cui centro coincide con l'origine,
- l'angolo compreso tra ciascun vettore e l'asse relativo alla componente che si vuole esaminare individua l'entità della loro correlazione. Per cui se l'angolo è picco-

¹⁰Per esercizio si scrivono le espressioni delle due componenti.

lo vuol dire che tra la variabile e la componente sussiste una forte correlazione
viceversa una debole correlazione.

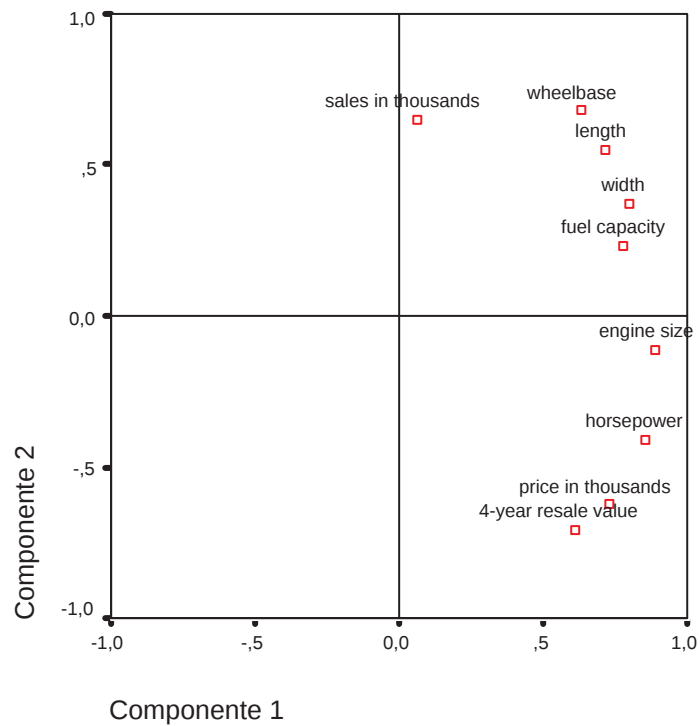


Figura 2.12: Grafico delle componenti

Si lascia allo studente l'interpretazione del grafico relativo all'analisi appena condotta.

2.4 Esercizi

Esercizio 2.1

Effettuare un'analisi di tipo esplorativo sulle variabili contenute nel dataset *Vendite auto*, (tabelle frequenza, statistiche descrittive, grafici rappresentativi di ogni distribuzione, correlazione). Commentare accuratamente i risultati.

Esercizio 2.2

La percentuale di varianza totale spiegata dalle prime due componenti principali risulta sufficiente? Esse tengono conto di una ragionevole quota della varianza totale? Si argomenta la risposta.

Esercizio 2.3

Quali sono i criteri a cui bisogna far riferimento per determinare il numero di componenti principali da utilizzare?

Esercizio 2.4

Commentare il grado di correlazione delle altre variabili con le componenti principali estratte.