

5. COMPONENTI PRINCIPALI

5.1 Introduzione

L'analisi in componenti principali (CP) si pone l'obiettivo di determinare quelle trasformazioni lineari delle h variabili originarie X_i ($i = 1, 2, \dots, h$), linearmente indipendenti fra loro, che racchiudano in sé la maggior parte delle informazioni raccolte sulle n unità. Se questa sintesi è possibile, i valori delle nuove variabili possono essere utilizzati in successive elaborazioni al posto di quelli originari, senza una significativa perdita di informazione e con i vantaggi derivanti dalla minore dimensione della matrice dei dati e dalla indipendenza lineare fra le variabili.

Indicata con Y_j la j -esima componente principale ($j = 1, 2, \dots, k \leq h$)

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \dots + a_{hj}X_h$$

questa variabile corrisponde alla somma ponderata delle variabili originarie in cui i pesi sono costituiti dai coefficienti a_{ij} .

La determinazione delle trasformazioni lineari può essere effettuata in modi diversi che dipendono dal criterio con cui è misurata la quantità complessiva d'informazione contenuta nelle variabili originarie.

Un primo criterio utilizza come prima componente principale la combinazione Y_1 che ha la **massima varianza** fra tutte le possibili combinazioni lineari, la seconda componente è invece la combinazione Y_2 , incorrelata con Y_1 , che ha la massima varianza fra tutte le possibili combinazioni lineari rimanenti e così via, per cui le prime k componenti derivate dalla matrice di covarianza \mathbf{S} delle variabili originarie sono quelle combinazioni lineari, incorrelate fra di loro, per le quali risulta massima la somma delle varianze. Se si considerassero tutte le h possibili componenti principali, la somma delle loro varianze è sempre uguale alla somma delle varianze delle h variabili originarie.

Usualmente il numero di componenti che vengono considerate è però minore di h , dato che una componente con una bassa varianza ha meno contenuto informativo delle precedenti. La quantità che viene presa in considerazione per valutare se le prime k componenti contengono una quantità sufficiente dell'informazione originaria è il rapporto fra la loro variabilità complessiva e la variabilità totale delle variabili originarie.

Si comprende quindi come la valutazione preliminare delle correlazioni tra le X_i è indispensabile. Se le variabili presentassero correlazioni molto basse non avrebbe senso operare alcuna sintesi. Questa analisi

è tanto più valida quanto più tra le variabili esiste un'apprezzabile quota di varianza in comune, così che poche CP basteranno per ottenere una buona approssimazione delle informazioni originarie.

Le componenti derivate dalla matrice di covarianza, determinate in base ai valori degli autovalori e autovettori di \mathbf{S} , non sono invarianti rispetto a trasformazioni di scala delle variabili originarie per cui, a parità di altre condizioni, le prime componenti tenderanno a ricostruire meglio quelle X_i che hanno varianza maggiore.

Quando le variabili sono espresse in unità di misura diverse o hanno diverso ordine di grandezza si preferisce costruire le CP sulla base degli autovalori e degli autovettori della matrice di correlazione \mathbf{R} anziché lavorare sulla matrice \mathbf{S} .

Le CP ottenute da \mathbf{S} e \mathbf{R} non sono in genere le stesse, né è possibile passare dalle une alle altre mediante trasformazione dei coefficienti.

5.2 Proc princomp sulla matrice di covarianza

Si può dimostrare che

- 1) le CP hanno media pari a zero;
- 2) la varianza della j -esima componente principale è pari all'autovalore λ_j ;
- 3) L'autovalore λ_j rappresenta anche il contributo della j -esima componente principale alla varianza complessiva di tutte le h variabili originarie.

Le prime k componenti Y_j derivate dalla matrice di covarianza \mathbf{S} sono quindi quelle combinazioni lineari, incorrelate fra loro, per le quali risulta massima la somma delle loro varianze

$$\sum_{j=1}^k \lambda_j = \max$$

mentre la somma delle varianze di tutte le h CP è uguale alla somma delle varianze delle h variabili X_i .

Di conseguenza il rapporto

$$\frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^h \lambda_j}$$

rappresenta la quota della variabilità complessiva contenuta nelle prime k CP.

Sono stati proposti diversi criteri empirici per decidere il numero di componenti che contengono una quota sufficiente dell'informazione originaria complessiva. I principali criteri sono:

- 1) utilizzare il grafico degli autovalori o "Screen Plot", scegliendo come numero di componenti quello corrispondente al punto di "gomito" della spezzata. Lo Screen Plot è costruito ponendo sull'asse delle ascisse il numero d'ordine degli autovalori (da 1 ad h) e in ordinata gli autovalori corrispondenti ($\lambda_1, \lambda_2, \dots, \lambda_h$). Si uniscono i punti con dei segmenti e il numero di CP corrisponde al più piccolo k tale che alla sua sinistra l'andamento dei λ_h sia fortemente decrescente, mentre alla sua destra l'andamento deve essere pressoché costante, o comunque debolmente decrescente;
- 2) utilizzare un numero di componenti che contiene circa l'80%-90% della variabilità totale;
- 3) prendere solo quelle componenti che hanno un autovalore maggiore o uguale ad uno, oppure le componenti che hanno varianza maggiore di quella media (ottenuta come media dei λ_j).

L'interpretazione delle CP dipende essenzialmente dalla capacità ed esperienza del ricercatore, per cui non è possibile fornire regole generali, anche se i segni coefficienti a_{ij} e le correlazioni fra la componente Y_j e ciascuna delle X_i , possono fornire interpretazioni a volte originali, suggerendo una qualche "caratteristica generale" degli individui esaminati. In generale, quanto più grande è a_{ij} (considerato in valore assoluto), tanto maggiore sarà il peso della corrispondente variabile X_i nel determinare la j -esima componente principale. Anche il calcolo dei coefficienti di correlazione fra ogni CP e le variabili originarie consente di interpretare le nuove variabili, tenendo conto del segno e del valore assunto da tali coefficienti (un ρ elevato, in valore assoluto, indica che la CP è fortemente influenzata dalla variabile originaria corrispondente).

5.3 Un esempio numerico (matrice di covarianza)

I seguenti dati si riferiscono ad un campione di lettori che hanno valutato 16 scrittori di romanzi esprimendo opinioni sul tipo di contenuto e sul modo in cui sono state scritte alcune delle loro opere. In particolare sono state rilevate le seguenti variabili

X_1 : giudizio

X_2 : leggibilità

X_3 : politica

X_4 : fantasia

X_5 : rilettura

X_6 : thriller

X_7 : attualità

I valori assunti da tali variabili sono riportate nel listato seguente, che riporta anche le istruzioni necessarie per effettuare l'analisi in componenti principali sulla matrice di covarianza, memorizzando l'output in un DSS, e la matrice di correlazione fra variabili originarie e CP. La `proc gplot` riporta invece lo scatter dei valori assunti dalle prime due componenti principali per ciascuno dei 16 autori (ogni simbolo identifica un autore diverso).

```
data a;
input scrittore $ 1-11 x1 x2 x3 x4 x5 x6 x7;
cards;
Fallaci      8.90 9.40 7.50 2.30 7.40 1.50 10.5
Tolkien     7.99 10.30 0.40 9.70 8.90 6.50 1.3
Rowling     8.01 10.56 1.60 10.02 7.59 7.01 1.68
Hardt       9.06 6.25 7.09 2.30 4.50 1.33 8.45
Cornwel     6.75 7.88 6.22 8.42 6.04 6.99 6.05
Camilleri  6.99 7.56 6.78 3.01 7.85 3.14 4.25
Mazzantini 8.33 8.01 2.52 3.50 6.55 2.47 6.70
Allende     5.95 7.73 3.50 5.80 6.45 3.20 2.40
Nasar       7.23 8.21 3.56 1.50 3.24 4.01 5.63
Eco         8.68 9.04 4.56 4.36 8.25 2.32 6.78
Gadda       9.12 7.56 6.89 1.23 5.89 2.32 7.88
Grass       7.56 6.12 6.52 4.56 6.01 1.25 6.12
King        6.22 8.44 1.01 2.14 7.13 4.22 1.23
Magris      8.14 7.06 2.10 6.55 4.12 0.88 2.26
Mishima     7.99 7.12 5.98 5.45 5.01 1.08 4.16
Lucarelli   6.04 6.42 3.01 5.02 3.01 6.89 3.01
;
proc princomp cov out=b;
proc print data=b;
proc corr noprob;
proc gplot;
plot PRIN2*PRIN1=scrittore;

run;
```

Di seguito sono elencate alcune delle tabelle restituite dal programma SAS, la prima delle quali riporta il numero di osservazioni e di variabili utilizzate nella procedura, mentre la seconda riporta la matrice di covarianza delle 7 variabili.

Tabella 5.2.1
Osservazioni e variabili

Osservazioni	16
Variabili	7

Tabella 5.2.2
Matrice di covarianza

	x1	x2	x3	x4	x5	x6	x7
x1	1.118213333	0.234826667	0.857746667	-0.499860000	0.315453333	-1.192970000	1.809113333
x2	0.234826667	1.756105000	-1.427770000	1.512988333	1.605441667	1.323772500	-0.736453333
x3	0.857746667	-1.427770000	5.770633333	-3.057990000	-0.470576667	-2.633621667	5.293526667
x4	-0.499860000	1.512988333	-3.057990000	7.813118333	1.356885000	3.598234167	-4.450606667
x5	0.315453333	1.605441667	-0.470576667	1.356885000	3.084385000	0.510634167	-0.349126667
x6	-1.192970000	1.323772500	-2.633621667	3.598234167	0.510634167	5.085252917	-3.087440000
x7	1.809113333	-0.736453333	5.293526667	-4.450606667	-0.349126667	-3.087440000	7.834680000

I valori sulla diagonale principale mettono in evidenza come le varianze più elevate si abbiano per X_4 e X_7 , mentre valori un po' più bassi si hanno per X_3 e X_6 . Complessivamente, la varianza totale è pari a

Varianza totale	32.462387917
------------------------	--------------

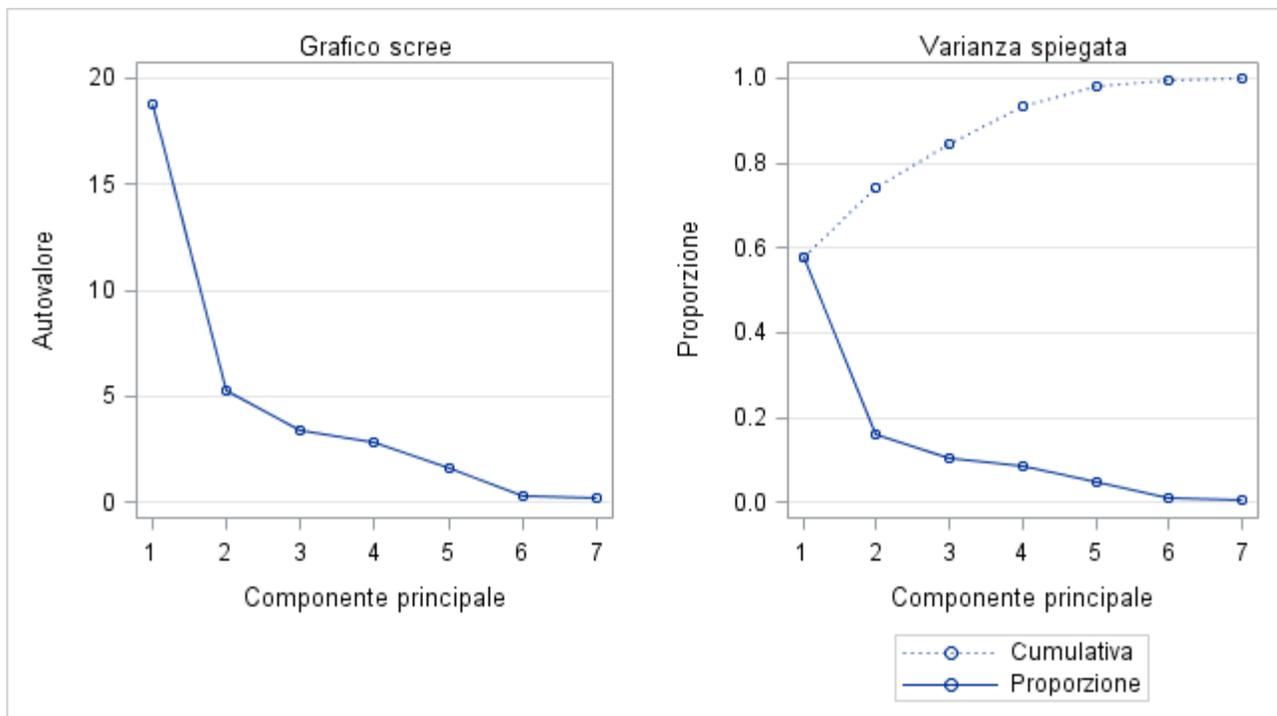
Nella successiva tabella 5.2.3 sono riportati gli autovalori calcolati sulla matrice di covarianza, la differenza fra ciascun autovalore e il successivo, la proporzione di varianza spiegata, pari al rapporto fra l'autovalore e la varianza totale, e la proporzione cumulata.

Tabella 5.2.3
Autovalori della matrice di covarianza

	Autovalore	Differenza	Proporzione	Cumulativa
1	18.8103483	13.5718820	0.5795	0.5795
2	5.2384663	1.8354318	0.1614	0.7408
3	3.4030345	0.5804841	0.1048	0.8457
4	2.8225504	1.2104415	0.0869	0.9326
5	1.6121089	1.2651758	0.0497	0.9823
6	0.3469331	0.1179867	0.0107	0.9929
7	0.2289464		0.0071	1.0000

I due grafici successivi, forniti dal SAS, possono essere utilizzati per determinare il numero di componenti da utilizzare, sulla base del primo e del secondo criterio precedentemente elencati.

Figura 5.2.1
Grafici Screen Plot e della varianza spiegata



Come si vede sia dalla tabella 5.2.3 sia dal secondo grafico della figura 5.2.1, la prima CP spiega poco meno del 58% della variabilità complessiva, le prime due componenti arrivano al 74% circa, mentre per superare il 90% della variabilità spiegata occorrono le prime quattro CP.

In questo caso il criterio dello Screen Plot sembra suggerire l'utilizzo delle prime 2 o 3 CP, il secondo criterio suggerisce l'uso di 3 o 4 CP, mentre il terzo criterio suggerisce l'uso delle prime cinque CP, per cui la scelta varia a seconda del criterio ritenuto migliore nel caso in esame.

Per quanto riguarda l'interpretazione delle nuove variabili, le prime informazioni in merito sono contenute nella tabella 5.2.4, che riporta gli autovettori. I coefficienti delle combinazioni lineari coincidono infatti con le componenti degli autovettori corrispondenti agli autovalori.

Tabella 5.2.4
Autovettori

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
x1	0.117556	0.206176	-.168291	-.160390	0.413739	0.470786	0.704845
x2	-.146450	0.305104	-.381864	0.218073	0.165512	0.569192	-.583708
x3	0.458885	0.331067	0.430452	-.107551	-.565674	0.402445	-.031830
x4	-.521398	0.526986	0.386456	-.488322	0.184194	-.132315	-.105781
x5	-.096402	0.486961	-.602465	0.003088	-.489211	-.315045	0.228083
x6	-.381413	0.183016	0.350059	0.782928	-.067703	0.046199	0.280701
x7	0.572305	0.454376	0.075601	0.252404	0.451221	-.416563	-.139505

In linea con quanto rilevato a proposito della matrice di covarianza, si nota come la prima CP è legata fortemente alle variabili che presentano la varianza più elevata, con un coefficiente pari a -0.52 circa per X_4 e pari a 0.57 per X_7 . Coefficienti leggermente più bassi, in valore assoluto, si hanno per X_3 e X_6 .

In definitiva, quindi, la prima CP è una combinazione lineare delle 7 variabili originarie, i cui pesi corrispondono agli elementi degli autovettori, per cui la prima CP è approssimativamente pari a

$$Y_1 = 0.1176X_1 - 0.1465X_2 + 0.4589X_3 - 0.5214X_4 - 0.0964X_5 - 0.3814X_6 + 0.5723X_7$$

Le affermazioni precedenti circa il peso associato alle diverse variabili nel determinare la prima CP sono confermate dalla matrice di covarianza fra variabili originarie e componenti che è riportata nella tabella 5.2.5.

Tabella 5.2.5
Matrice di correlazione fra variabili originarie e CP

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
x1	0.48215	0.44625	-0.29358	-0.25482	0.49678	0.26223	0.31893
x2	-0.47931	0.52696	-0.53158	0.27647	0.15858	0.25299	-0.21076
x3	0.82850	0.31543	0.33056	-0.07522	-0.29899	0.09868	-0.00634
x4	-0.80901	0.43151	0.25505	-0.29350	0.08367	-0.02788	-0.01811
x5	-0.23807	0.63462	-0.63282	0.00295	-0.35368	-0.10566	0.06214
x6	-0.73356	0.18575	0.28636	0.58329	-0.03812	0.01207	0.05956
x7	0.88678	0.37154	0.04983	0.15150	0.20468	-0.08766	-0.02385

Si può quindi concludere che la componente Y_1 tiene conto soprattutto del contenuto, indicando una certa preferenza per i seguenti temi: politica, fantasia, thriller e attualità.

L'interpretazione della seconda CP risulta però meno chiara, dato che risulta abbastanza legata con tutte le variabili considerate, tranne che con X_6 , che contribuisce in misura molto esigua rispetto alle restanti. Nella tabella 5.2.6 sono indicati i valori assunti da tutte le CP sui 16 scrittori esaminati e sono questi i valori che potrebbero essere considerati in analisi successive, se si volessero utilizzare la CP al posto delle variabili originarie.

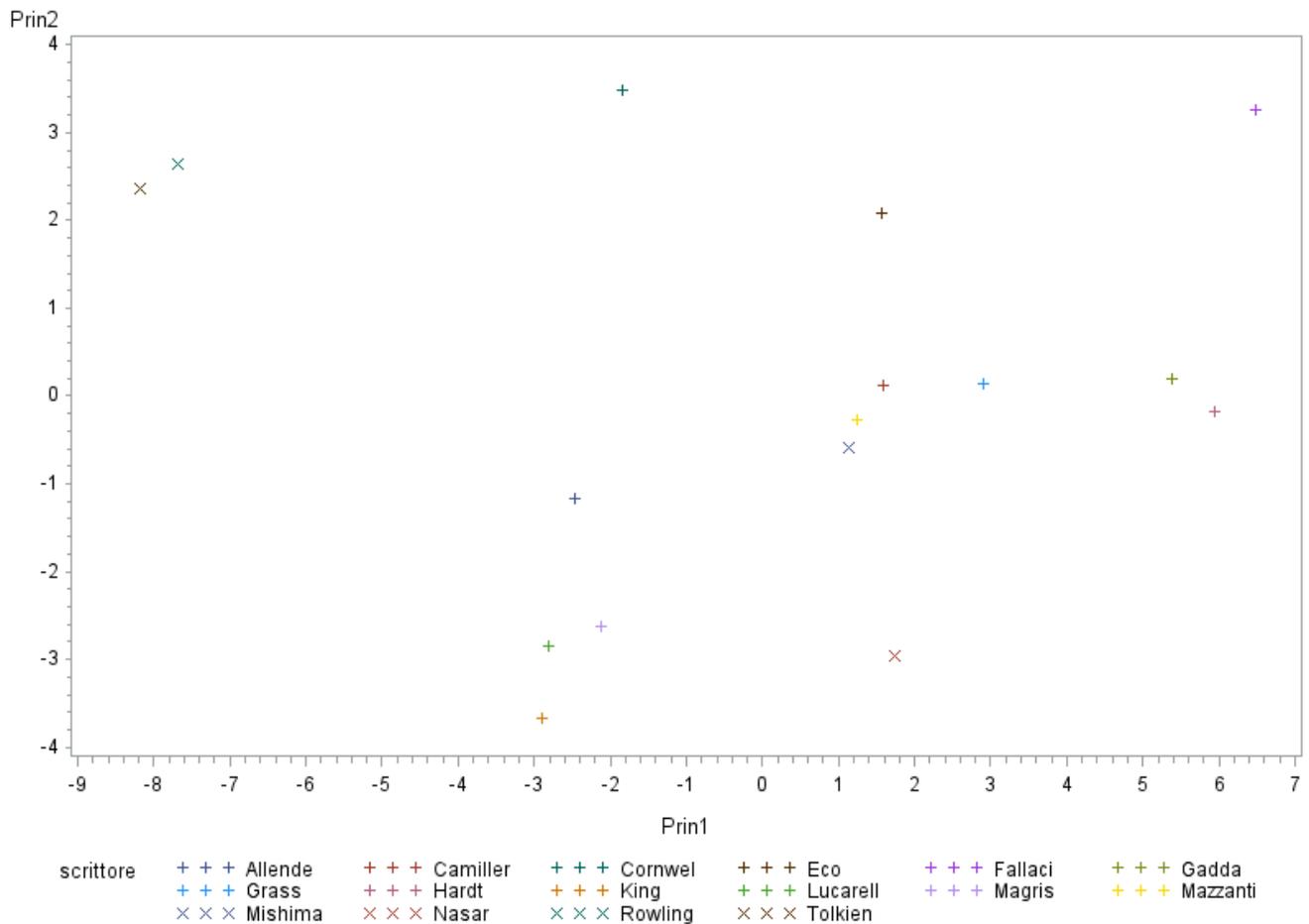
Tabella 5.2.6
Valori assunti dalle CP

Oss	scrittore	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
1	Fallaci	6.48661	3.25930	-1.35271	0.86108	0.52656	0.15530	-0.85130
2	Tolkien	-8.18547	2.36065	-1.58861	-0.04951	0.45478	-0.00654	0.45424
3	Rowling	-7.68814	2.63810	-0.05456	0.20981	0.66402	0.86944	0.03588
4	Hardt	5.94977	-0.17930	1.17941	-0.46690	0.80854	0.03240	0.69002
5	Cornwel	-1.83147	3.47416	3.80844	1.19446	-0.47752	-0.51117	-0.23447
6	Camiller	1.58515	0.11932	-0.53370	0.20476	-3.18147	0.36259	0.25917
7	Mazzanti	1.24945	-0.26167	-1.84153	0.39668	1.73428	-1.17164	0.19839
8	Allende	-2.46853	-1.17020	-0.03267	-1.02531	-1.36823	-0.50497	-0.80822
9	Nasar	1.73024	-2.95328	0.39434	2.40693	1.38780	0.66717	-0.68862
10	Eco	1.56659	2.09006	-2.15392	-0.16620	0.26859	-0.28923	0.02244

Oss	scrittore	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
11	Gadda	5.39328	0.20173	-0.36449	0.98869	-0.03799	0.71264	0.76166
12	Grass	2.90403	0.13602	0.99564	-1.94304	-0.87951	-0.78502	0.13471
13	King	-2.89927	-3.66484	-2.97656	1.64665	-1.33428	-0.17127	0.03871
14	Magris	-2.11708	-2.61417	0.12285	-3.59726	1.59081	0.16723	-0.07143
15	Mishima	1.13581	-0.58863	1.04769	-2.80136	-0.45037	0.77518	-0.22524
16	Lucarell	-2.81095	-2.84725	3.35038	2.14052	0.29400	-0.30210	0.28407

Infine, nel grafico successivo, è stato riportato lo scatter dei valori assunti dalle prime due CP sui 16 autori considerati, con l'indicazione del nome di ciascuno scrittore.

Figura 5.2.2
Diagramma di dispersione delle prime due CP



5.4 Proc princomp sulla matrice di correlazione

Quando l'analisi in CP viene effettuata sulla matrice di correlazione, la prima CP è sempre la combinazione che ha la massima varianza fra tutte le possibili combinazioni lineari delle variabili X_i , ma è anche la combinazione che massimizza la somma dei coefficienti di determinazione R^2 fra la componente stessa e le variabili originarie, per cui è la combinazione che permette di ricostruire al meglio il valore delle X_i .

In questo caso l'autovalore λ_j corrisponde quindi sia alla varianza della j -esima componente sia alla somma dei coefficienti di determinazione lineare della j -esima CP contro le variabili originarie. Per effettuare l'analisi in CP basata sulla matrice di correlazione sui dati precedentemente utilizzati l'unica differenza nel listato SAS è l'omissione dell'opzione `cov`, dato che per default il SAS effettua l'analisi in CP sulla matrice di correlazione.

Sulla base del listato seguente

```
proc princomp out=b;  
proc print data=b;  
proc corr noprob;  
proc gplot;  
plot PRIN2*PRIN1=scrittore;  
run;
```

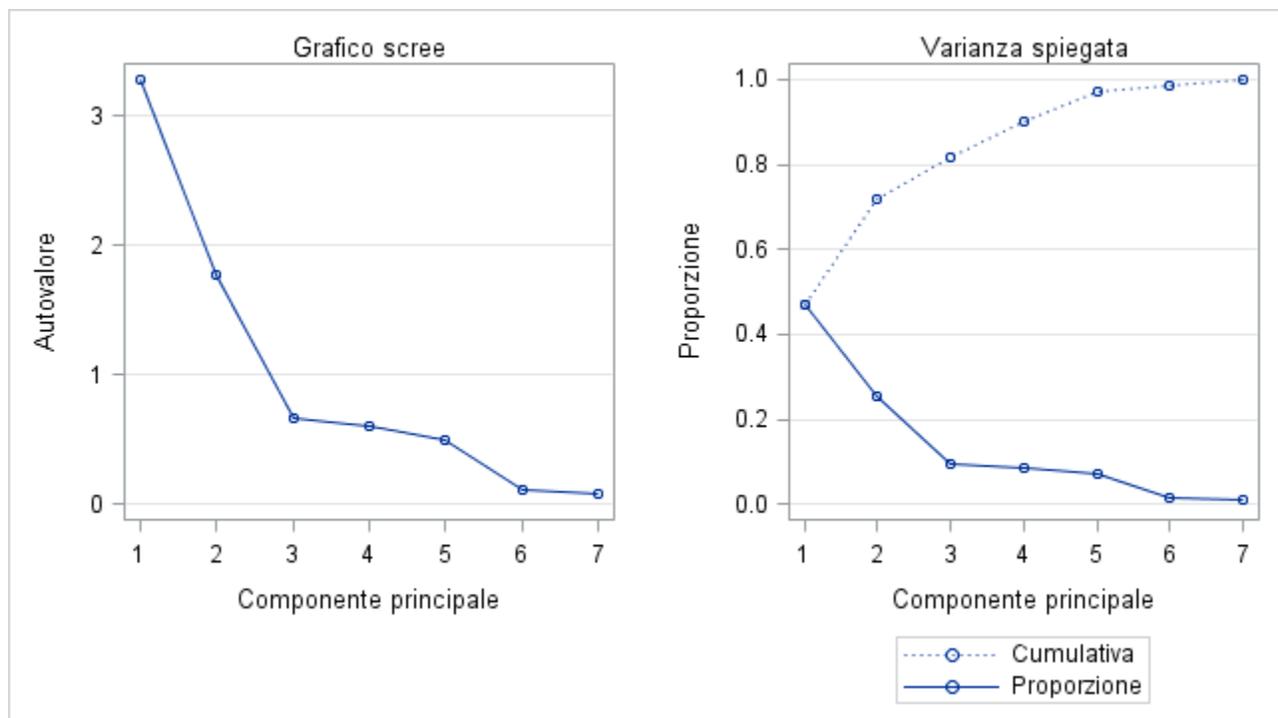
il SAS restituisce le informazioni riportate nelle tabelle successive.

Tabella 5.3.1
Matrice di correlazione

	x1	x2	x3	x4	x5	x6	x7
x1	1.0000	0.1676	0.3377	-.1691	0.1699	-.5003	0.6112
x2	0.1676	1.0000	-.4485	0.4085	0.6898	0.4430	-.1985
x3	0.3377	-.4485	1.0000	-.4554	-.1115	-.4862	0.7873
x4	-.1691	0.4085	-.4554	1.0000	0.2764	0.5708	-.5688
x5	0.1699	0.6898	-.1115	0.2764	1.0000	0.1289	-.0710
x6	-.5003	0.4430	-.4862	0.5708	0.1289	1.0000	-.4891
x7	0.6112	-.1985	0.7873	-.5688	-.0710	-.4891	1.0000

Tabella 5.3.2
Autovalori della matrice di correlazione

	Autovalore	Differenza	Proporzione	Cumulativa
1	3.27798430	1.51014486	0.4683	0.4683
2	1.76783943	1.10280529	0.2525	0.7208
3	0.66503414	0.06767418	0.0950	0.8158
4	0.59735996	0.10613423	0.0853	0.9012
5	0.49122574	0.37661284	0.0702	0.9713
6	0.11461290	0.02866938	0.0164	0.9877
7	0.08594353		0.0123	1.0000



Anche in questo caso il numero di CP da considerare varia a seconda del criterio scelto:

- 1) Dallo Screen Plot, il numero di CP dovrebbe essere pari a 3, in quanto in corrispondenza di questo valore si ha un brusco cambiamento di pendenza;
- 2) In base alla proporzione di varianza spiegata, i primi due autovalori ricoprono il 72% della varianza totale, mentre se si considera anche il terzo autovalore si arriva all'81.58%.

3) il numero di autovalori maggiori di 1 è pari a due, per cui secondo il terzo criterio andrebbero prese solo le prime due CP.

Sulla base degli autovalori (riportati nella tabella 5.3.3) e della matrice di correlazione fra CP e variabili originarie (riportata nella tabella 5.3.4) si vede che la prima CP è costituita essenzialmente dalle medesime variabili che erano state considerate nel caso precedente, ossia nel caso dell'analisi in CP effettuata sulla matrice di covarianza.

Tabella 5.3.3
Autovettori

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
x1	-0.279416	0.515065	-0.344102	0.507845	0.185852	0.382034	0.315960
x2	0.325223	0.543889	-0.003251	-0.136897	0.376952	-0.639533	0.169022
x3	-0.450791	0.077255	0.599412	0.059668	-0.328944	-0.237633	0.513119
x4	0.416740	0.098041	0.142498	0.699529	-0.445621	-0.189093	-0.269694
x5	0.189346	0.591075	0.108140	-0.462786	-0.475020	0.385473	-0.121158
x6	0.435284	-0.066540	0.600868	0.102460	0.448569	0.450353	0.174724
x7	-0.462545	0.263657	0.359510	0.076044	0.299542	-0.037614	-0.700262

Tabella 5.3.4
Matrice di correlazione fra variabili originarie e CP

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
x1	-0.50589	0.68483	-0.28061	0.39251	0.13026	0.12934	0.09263
x2	0.58882	0.72316	-0.00265	-0.10581	0.26420	-0.21651	0.04955
x3	-0.81617	0.10272	0.48882	0.04612	-0.23055	-0.08045	0.15043
x4	0.75452	0.13036	0.11621	0.54066	-0.31232	-0.06402	-0.07906
x5	0.34282	0.78589	0.08819	-0.35768	-0.33293	0.13050	-0.03552
x6	0.78809	-0.08847	0.49001	0.07919	0.31439	0.15246	0.05122
x7	-0.83745	0.35056	0.29318	0.05877	0.20994	-0.01273	-0.20529

Come si è ricordato in precedenza, alla determinazione della prima componente (Y_1) contribuisce essenzialmente il contenuto del libro (politica, fantasia, thriller e attualità) mentre alla determinazione

della seconda componente (Y_2) contribuisce in questo caso il modo in cui è stato scritto (giudizio, leggibilità e rilettura).