Statistica sociale - 12

Prof. Antonio Mussino

a. a. 2022-2023

SAPIENZA
UNIVERSITÀ DI ROMA

La Statistica multivariata

Statistica multivariata - 1

- Anche nella Statistica multivariata sono presenti gli approcci esplorativi, descrittivi e sintetici, nonché quelli inferenziali che abbiamo visto nell'analisi univariata e bivariata della matrice dei dati.
- Se l'obiettivo è quello di estrarre il massimo dell'informazione da grandi masse di dati e di sintetizzare in insiemi più piccoli (indicatori di sintesi) un grande numero di unità statistiche e/o di variabili (indicatori), allora è necessario utilizzare strategie e strumenti esplorativi che sono caratteristici dell'***Analyse des données***.

Statistica multivariata - 2

- Le strategie e gli strumenti (modelli) inferenziali saranno sviluppati in altri corsi: qui ne proporremo uno solo, a titolo esemplificativo, nell'ultima parte del corso.

L'*Analyse des données* - 1

- L'Analyse des données (Add) è un'etichetta attribuita ad alcune tecniche di analisi che cercano di esplorare, descrivere e sintetizzare il contenuto informativo presente in una matrice di dati.
- Il punto di partenza del padre dell'Add, J.P. Benzecrì, è indubbiamente esasperato

  <<Statistique n'est pas probabilitè>>
- ma fa capire lo spirito col quale si affrontano i dati presenti nella matrice, ovvero osservandoli <<per se stessi>>, indipendentemente da ipotesi su modelli teorici di riferimento da accettare o rifiutare: <<il modello deve seguire i dati e non viceversa>>.

L'*Analyse des données* - 2

- I metodi di Add che proporremo si possono distinguere a seconda dell'obiettivo di sintesi che ci poniamo:
  - se vogliamo operare sulle colonne della matrice (variabili), allora parleremo di ***metodi fattoriali***, che mirano a individuare variabili di sintesi;
  - se, invece, vogliamo sintetizzare le righe (unità), per individuare tipologie omogenee di comportamento, allora parleremo di ***metodi di classificazione automatica***.

I metodi fattoriali

- Iniziamo dal caso più semplice, in cui si analizza una matrice composta da tutte variabili quantitative: l'***Analisi in Componenti Principali* (ACP)**.
- È il caso cui possiamo ricondurre la tematica affrontata della ricerca di indici di sintesi per batterie di molti indici/indicatori.

L'ACP - 1

- Per descrivere e comprendere la strategia di questa tecnica, abbiamo bisogno della formalizzazione* degli ***spazi vettoriali***.
- Il modello matematico degli spazi vettoriali è caratterizzato dalla possibilità di rappresentare le relazioni fra le unità statistiche e/o le variabili, presenti nella matrice dei dati, su spazi "paralleli", ossia aventi caratteristiche geometriche simili.

* Per formalizzazione dei dati si intende una struttura concettuale, matematicamente tradotta, entro la quale prendono corpo specifiche tecniche di analisi (R.Coppi, 1981).

La matrice dei dati

Nome	Sport	corsa	salto	fless	spola	tapp	later	rank
Andrea	Calcio	9,8	177	6,2	17,9	38,6	Dx	6°
Carlo	Volley	10,2	169	10,2	18,2	38,4	Dx	22°
Enrico	Volley	9,5	178	11,9	17,6	38,1	Sn	3°
Gianni	Calcio	9,6	179	9,6	17,2	37,4	Dx	5°
Mario	Volley	9,2	182	6,4	16,8	36,2	Dx	10°
Mauro	Volley	9,1	186	10,1	16,4	37,4	Dx	7°
Nicola	Calcio	8,9	190	8,4	16,5	39,2	Sn	1°
Sandro	Atletica	9,3	180	10,4	17,0	39,6	Dx	12°
Silvano	Atletica	10,1	174	8,2	18,6	39,2	Dx	20°
Ugo	Atletica	9,7	177	8,4	17,9	38,1	Dx	14°

L'ACP - 2

- Riprendiamo in considerazione la matrice dei dati, così come l'abbiamo definita precedentemente.

M = (n,p)	X_{11}	X_{12}	X_{1j}	X_{1p}	
	X_{21}	X_{22}	X_{2j}	X_{2p}	
	X_{i1}	X_{i2}	X_{ij}	X_{ip}	$\rightarrow \mathbf{R}_p$
	X_{n1}	X_{n2}	X_{nj}	X_{np}	
	$\rightarrow \mathbf{R}_n$				

L'ACP - 3

- Le ***n*** righe (unità) sono rappresentabili come vettori su p dimensioni, definite dalle ***p*** colonne (variabili), in quello che è chiamato "spazio delle unità" $\mathbf{R}_p$.
- Le ***p*** colonne (variabili) sono rappresentabili come vettori su n dimensioni, definite dalle ***n*** righe (unità), in quello che è chiamato "spazio delle variabili" $\mathbf{R}_n$.

L'ACP - 4

- Entrambi questi spazi di riferimento sono "spazi vettoriali" ed è quindi possibile effettuare le operazioni di somma di vettori, di moltiplicazione di un vettore per uno scalare ed è introdotto un criterio (prodotto scalare) per misurare la distanza fra variabili o fra unità (nel loro spazio di pertinenza), cioè una "metrica": in questo caso la tradizionale metrica euclidea.

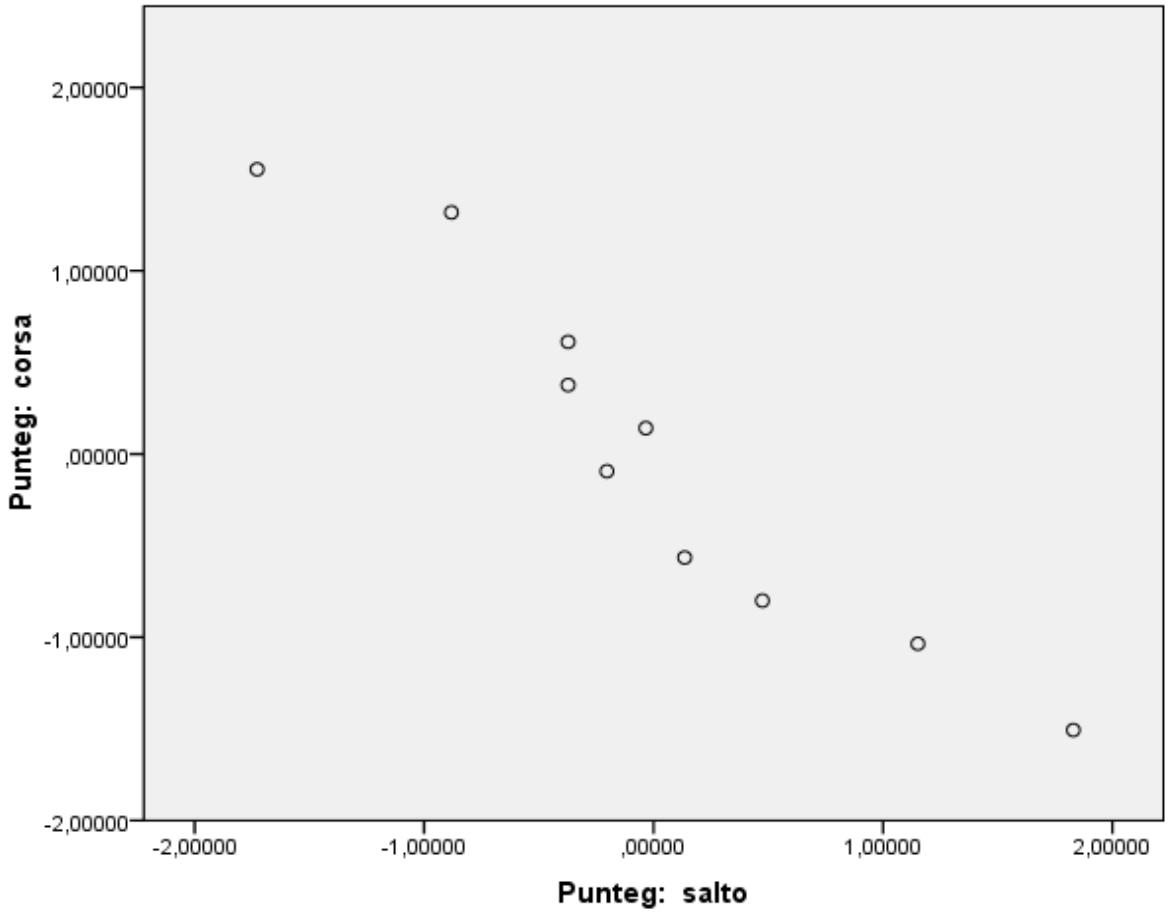
L'ACP - 5

- Riprendiamo l'esempio proposto della matrice (10 unità e 5 variabili) sugli allievi dei CAS, per meglio chiarire il significato delle relazioni che proporremo.
- Abbiamo visto:
  - la possibilità di rappresentare le unità (tramite *scatter diagram*) su spazi a 1 dimensione (la retta) e a 2 dimensioni (il piano);
  - come, su ciascuno dei molteplici spazi così individuabili (5 rette e 10 piani), le unità si dispongano su "nuvole" che prefigurano diversi tipi di relazione fra le variabili stesse.

L'ACP - 6

- Analogamente è possibile (anche se non percepibile) la rappresentazione su uno spazio a 5 dimensioni (il nostro ***spazio delle unità***).
- In questo approccio non tratteremo della strategia simmetrica che si può applicare allo ***spazio delle variabili***, lasciando ad altri corsi la sua trattazione.
- Ricordiamo sempre che le variabili vanno sempre centrate o, se si vuole eliminare l'effetto della diversa unità di misura o della diversa variabilità, standardizzate.

Plot ZCORSA vs. ZSALTO



L’ACP - 7

- Si è visto come la relazione, evidenziata da un alto coefficiente di correlazione lineare di Pearson, fra le prove di corsa e di salto sia riassumibile sinteticamente su di una retta, che rappresenta una variabile latente che, secondo la teoria dell’allenamento, possiamo definire la “capacità di forza veloce (o rapida)”.

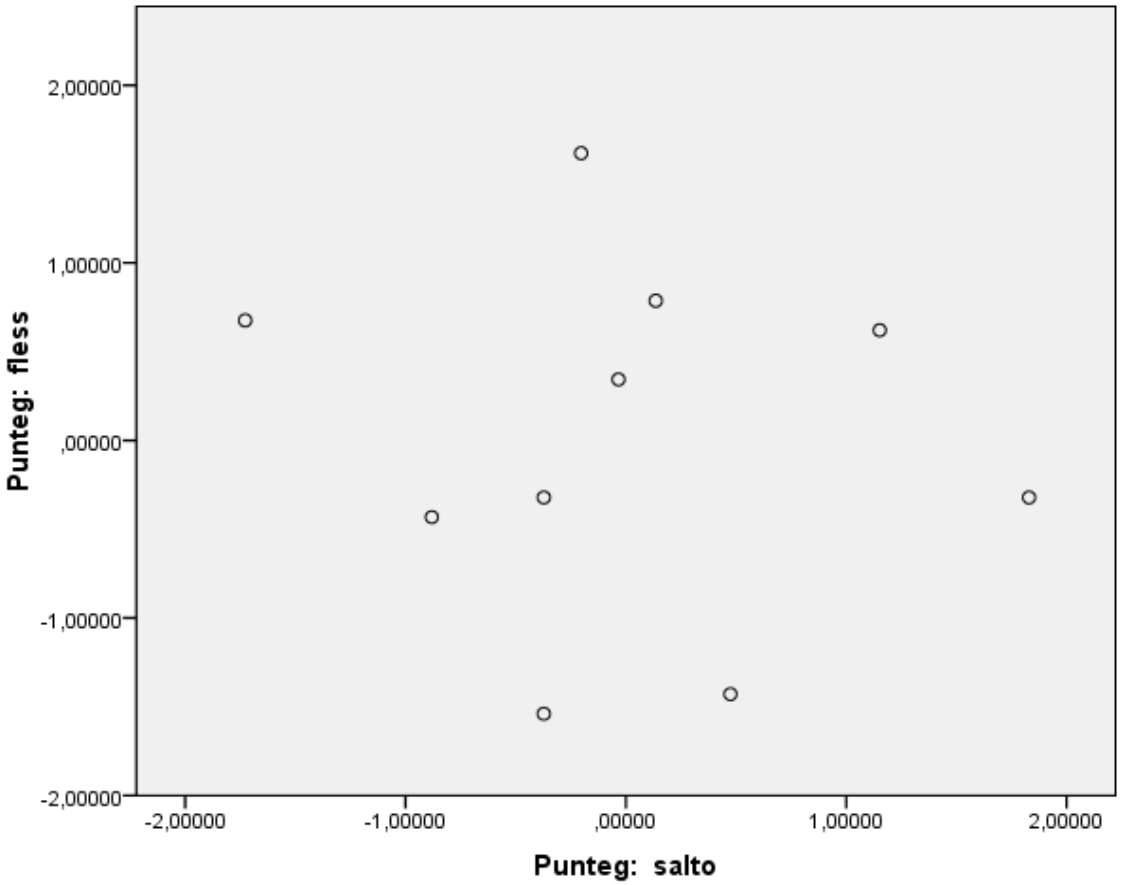
L'ACP - 8

- La variabilità dei risultati degli allievi può essere rappresentata non più su due dimensioni X_1 e X_2, bensì su una sola retta F_1: infatti la reale variabilità dei risultati fra i ragazzi non è quella misurata nelle prove di corsa e di salto, bensi quella nella dimensione latente, ovvero non direttamente misurabile, della "forza veloce", la cui conseguenza sono i risultati (variabili) nella corsa e nel salto.

L’ACP - 9

- La “forza veloce” ha, infatti, influenzato i risultati nelle prove e i punti che rappresentano i ragazzi nel piano non cadono esattamente sulla retta perché in ogni prova sono presenti situazioni o condizioni di disturbo o di incentivazione, che fanno sì che gli allievi migliorino o peggiorino le loro *performance* teoriche.

Plot: ZFLESS vs. ZSALTO

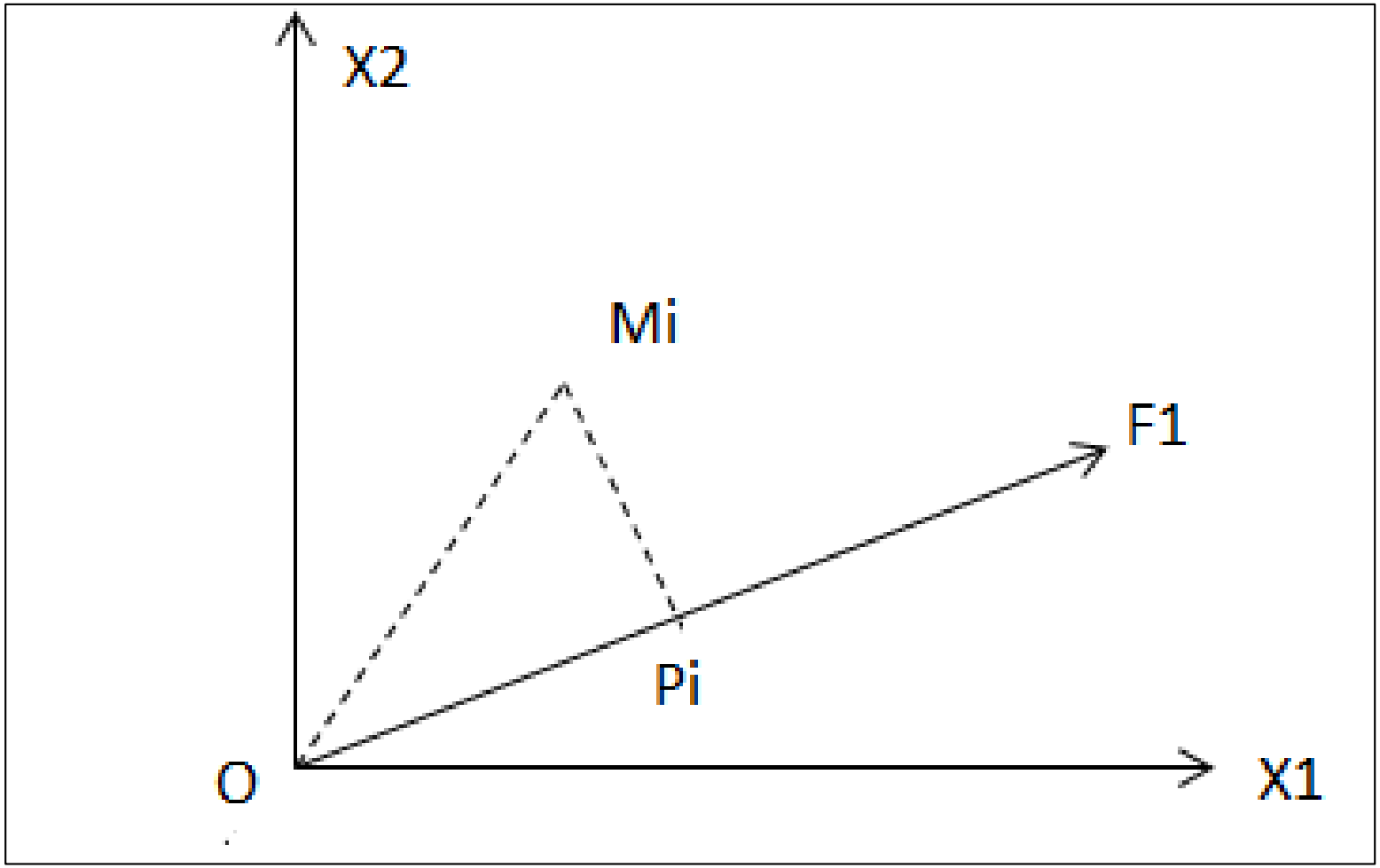


L'ACP - 10

- Invece nell'altro esempio non c'è alcuna relazione: infatti le due variabili misurano due capacità motorie e coordinative diverse, per cui non è possibile ridurre la ***dimensionalità*** del problema: la nuvola dei punti presenta una dispersione dovuta alla variabilità dei risultati sulle due dimensioni.

L'ACP - 11

- Ma come trovare la retta F_1 che rappresenta la dimensione "forza veloce"?
- Si consideri una generica retta passante per l'origine e si prenda un punto (M_i) sul piano: quello che rappresenta i risultati di un allievo (u_i) nelle prove misurate da X_1 e X_2.


X2
Mi
F1
Pi
O
X1

L'ACP – 11 bis

- La proiezione (perpendicolare) del punto M_i sulla retta individuerà il punto P_i.
- La distanza di M_i dall'origine (O) è l'ipotenusa di un triangolo rettangolo i cui cateti sono OP_i e M_iP_i:

$$O\,M_i^2 = O\,P_i^2 + M_iP_i^2$$

L'ACP - 12

- Più $M_iP_i^2$ è piccolo, meglio la retta rappresenterà M_i, fino a che $M_iP_i^2 = 0$ e la retta passerà per il punto.
- Ma più $M_iP_i^2$ diventa piccolo, più OP_i^2 cresce.
- La relazione che ci interessa è però quella complessiva, che tiene conto di tutti i punti M_i, con i=1,n :

$$\Sigma OM_i^2 = \Sigma OP_i^2 + \Sigma M_iP_i^2$$

- Quindi la ***migliore retta interpolante*** la nuvola dei punti (criterio dei "minimi quadrati") è quella in cui risulta minimo il valore

$$\Sigma M_iP_i^2$$

L'ACP - 13

- La ΣOM_i^2 è la distanza al quadrato di tutti i punti dal centro di gravità della nuvola ed è definita ***inerzia***.
- La migliore retta interpolante, quindi, è quella in cui è preservato il massimo di tale inerzia.
- Una volta calcolata la F_1, per la cui individuazione è sufficiente il vettore unitario **1**, in quanto parte dall'origine, si possono proiettare i punti della nuvola su di essa e calcolare le ***coordinate*** di ogni unità su di essa.
- I ragazzi possono così essere ordinati secondo la loro "capacità di forza veloce".

L'ACP - 14

- Ma, a meno che le due variabili siano perfettamente correlate (r=+1 o r=-1), la quantità spiegata dalla F_1 rappresenta solo una parte dell'inerzia/variabilità totale.
- Ne rimane fuori una parte «non spiegata».
- Se tale quantità è proporzionalmente piccola, allora ci è sufficiente una dimensione per spiegare la dispersione della nuvola.
- Altrimenti il guadagno in ***sintesi*** (da 2 a 1 dimensione) viene perso per l'***errore di rappresentazione*** commesso.

L'ACP - 15

- Si tratta, pertanto, di scegliere la ***prospettiva ottimale*** dalla quale osservare la nuvola per sintetizzare le relazioni espresse nella matrice dei dati.
- Se questa prospettiva ci può trarre in inganno (come nel secondo grafico), allora la riduzione di dimensioni non risulta corretta!
- È evidente che queste considerazioni crescono di importanza all'aumentare delle dimensioni di riferimento! Pensiamo a una batteria di 30 prove: sarà necessario sintetizzare i risultati in un numero ridotto di dimensioni per ***ordinare*** e ***classificare*** gli allievi.

L'ACP - 16

- Possiamo così proporre la definizione classica di ACP (Hotelling, 1933).
- L'obiettivo dell'ACP è quello di ***rappresentare e interpretare gli elementi di partenza di una matrice di dati tramite la loro proiezione su sottospazi vettoriali di dimensioni ridotte, sotto la condizione di "minima perdita dell'informazione" rispetto alla struttura originaria***.

L'ACP - 17

- Se nel caso della nostra matrice volessimo utilizzare tutte e 5 le dimensioni previste, la rappresentazione diverrebbe illeggibile: si presenta quindi la necessità di cercare uno spazio di dimensioni ridotte.
- La soluzione su di una retta non può essere adeguata, perché abbiamo visto che almeno due delle 5 variabili sono incorrelate.

L'ACP - 18

- Possiamo così proporre una soluzione su due dimensioni, un piano, con il vincolo che la seconda retta F_2 sia ortogonale alla prima F_1, ovvero le due dimensioni siano incorrelate.
- Anche per individuare F_2 useremo la stessa strategia usata per individuare F_1, in modo che sia massima anche la proiezione delle distanze dei punti della nuvola dal centro di gravità sul piano (F_1,F_2).
- Reiterando il processo, avremo che le variabili X_1, X_2, X_3, X_4, X_5 individueranno le nuove variabili F_1, F_2, F_3, F_4, F_5.

L'ACP - 19

- Le F_1, F_2, F_3, F_4, F_5 sono definite le ***componenti principali***!
- Complessivamente i due insiemi spiegano ciascuno il 100% della inerzia/variabilità, ma mentre le X_i hanno tutte la varianza unitaria, e quindi uguale, le F_i hanno la capacità di spiegare tale informazione con peso gerarchicamente decrescente.
- ***Ovvero la prima spiega il max dell'inerzia, la seconda il max della residua (dopo aver tolto quella spiegata dalla prima), la terza il max della residua (dalle prime due) e così via.***

L'ACP - 20

- Se prendiamo un numero ridotto di componenti, ovvero quelle che spiegano *molta* inerzia, si rappresenteranno i dati con un guadagno nella ***sintesi*** e nella capacità di ***interpretare*** l'informazione sottostante, ma con una perdita di informazione rispetto a quella contenuta nella matrice di partenza.
- La scelta su quante siano le componenti da considerare può essere guidata da alcuni parametri numerici, ma è legata sostanzialmente alla valutazione soggettiva del ricercatore.
- Non essendo una strategia inferenziale non si può nemmeno utilizzare un test di ipotesi e un p-value.

L'ACP - 21

- Riassumendo, l'ACP è un metodo ***fattoriale***, in quanto la riduzione della complessità non è basata su una strategia soggettiva, ma sulla individuazione oggettiva di nuove variabili ***sintetiche***, ***incorrelate***, di ***varianza massima*** e di ***importanza decrescente***, ottenute come combinazioni lineari delle variabili di partenza.

L'ACP - 22

- L'ACP può anche essere vista come un metodo per eliminare la ***ridondanza*** dell'informazione, sostituendo le "p" variabili di partenza con un numero ridotto di componenti "k", di varianza massima e di importanza decrescente.
- Queste componenti, abbiamo visto, possono anche rappresentare caratteristiche delle unità non direttamente rilevabili (variabili "latenti").

L'ACP - 23

- In sintesi il modello delle Componenti principali può essere così formulato:

$$F_i = a_{1i}X_1 + a_{2i}X_2 + \ldots\ldots a_{ji}X_j + \ldots\ldots a_{pi}X_p$$

- Ciò evidenzia chiaramente la differenza col modello dell'Analisi Fattoriale:

$$X_i = b_{1i}F_1 + b_{2i}F_2 + \ldots\ldots b_{ji}F_j + \ldots\ldots a_{ki}F_k + \varepsilon_i$$

- con F_j fattore latente che influenza, secondo la nostra opinione i risultati della variabile X_i;
- con k<<p !!!!
- ed ε_i che rappresenta quella parte della variabilità della X_i che non è spiegata dal modello fattoriale.

L’interpretazione dei risultati di un ACP:

il *malessere demografico* in Liguria

Esempio - 1

- Per "***malessere demografico***" si intende il risultato, in termini di struttura e dinamica di una popolazione, di
- una prolungata presenza di ***bassa fecondità*** ed ***elevata età media***,
- che porta a un sempre più forte ***invecchiamento*** della popolazione
- con rischio di conseguente malessere ***economico*** e malessere ***sociale***, ma anche di malessere ***psicologico***.

Esempio - 2

- Il caso di studio è relativo a una batteria di 6 indicatori calcolati per i 235 comuni della **Liguria** su dati di **Censimento** (1981): matrice 235x6.

- Gli indicatori sono:
  - **P60ω81**: percentuale di ultrasessantenni sul totale della popolazione;
  - **P54981**: % di bambini fra 0 e 5 anni sulle donne in età feconda (tra 15 e 49 anni);
  - **TER81**: % di occupati nel settore terziario (esclusa PA) sul totale degli occupati;

Esempio - 3

- **AGR81**: % di occupati nell'agricoltura sul totale degli occupati;
- **DMF81**: dimensione media della famiglia (numero dei componenti);
- **DIPE81**: rapporto di dipendenza (pop. 0-14 anni più pop. 60-ω anni su pop. 15-59) per 100.

Esempio - 4

- L'applicazione dell'ACP alla matrice ha portato tecnicamente all'individuazione di 6 componenti, delle quali, però, solo 3 spiegano in modo significativo la variabilità complessiva.
- La variabilità di ciascun indicatore, standardizzato, è **1**: quindi quella complessiva **6**.
- L'individuazione delle componenti passa per gli ***autovalori*** della ***matrice delle correlazioni***, la cui somma è sempre **6**, ma i valori sono risultati: **3,18**; **1,32**; **0,78**; 0,36; 0,24; 0,12.
- Così $\mathbf{F_1}$ spiega il 53%, $\mathbf{F_2}$ spiega il 22%; complessivamente il 75%; con $\mathbf{F_3}$ che spiega il 13%, si arriva all'88%.

Esempio - 5

- Tecnicamente la soluzione ottimale prevederebbe la soluzione con 2 componenti (autovalore > 1), ma la terza è – come vedremo – interessante.
- Infatti ora si devono "battezzare" le componenti, ovvero comprendere il loro significato per decidere il loro ruolo nell'analisi.
- Per far questo calcoliamo i punteggi delle unità sulle componenti e, grazie a questi, le correlazioni variabile-componente.
- Ecco la matrice con queste nuove correlazioni.

Esempio - 6

Indicatore	F_1	F_2	F_3
P60ω81	**-0,96**	0,13	0,14
DMF81	**0,83**	-0,37	-0,13
DIP81	**-0,91**	0,08	0,13
P54981	0,45	-0,37	**0,81**
TER81	0,37	**0,83**	-0,17
AGR81	**-0,62**	**-0,58**	0,30

L’interpretazione dell’ACP: un caso di studio

Indicatori BES: la Salute - 1

- Tra gli indicatori utilizzati per il Rapporto BES 2013, abbiamo selezionato quelli relativi all'area Salute: si tratta di una serie di indicatori (indici demografici e sanitari) calcolati con dati relativi agli anni 2005-2011, avendo come unità di analisi la regione.
- Sono 17 indici, già ampiamente descritti, che presentano vari aspetti della salute dei cittadini; il dato è aggregato a livello regionale: pertanto si lavora con correlazioni ecologiche che misurano relazioni fra situazioni *medie* regionali, non applicabili a livello dei singoli cittadini.

Indicatori BES: la Salute - 2

- Un primo passaggio potrebbe essere quello di calcolare i coefficienti di correlazione lineare di Pearson fra tutti gli indici: si ricaverebbe una matrice delle correlazioni di dimensione 17x17 (ogni indice è messo in relazione con tutti gli altri), di cui però i valori utili per l'analisi sono solo 136 (infatti $r_{xx} = 1$ e $r_{xy} = r_{yx}$).
- Leggere tale matrice è piuttosto complesso e poi molte delle correlazioni sono alte, quindi molti indicatori misurano aspetti della sanità nelle regioni che sono sovrapponibili.

Indicatori BES: la Salute - 3

- Per vedere quanti e quali sono questi aspetti, e poter poi classificare le varie regioni o costruirne graduatorie, cerchiamo di sintetizzare la variabilità presente nella matrice dei dati.

L'ACP - 1

- L'algoritmo delle Componenti Principali parte proprio dalla matrice delle correlazioni sopra citata, calcolandone gli *autovalori* e gli *autovettori*.
- Questi ultimi sono costituiti dai coefficienti angolari che individuano la direzione delle 17 rette F_1, F_2,……., F_{17}, che interpolano al meglio lo spazio a 17 dimensioni definito da X_1, X_2, ……, X_{17}.
- Gli autovalori, invece, misurano la variabilità riprodotta sulle nuove rette che, ricordiamo, non è uniforme come nelle X_i, (unitaria, dato che le variabili sono standardizzate), bensì decrescente dalla F_1, dove è massima, alla F_{17}, dove è minima.

L’ACP – 2

- Nell’output seguente si vede molto bene questa capacità di spiegare la variabilità che diminuisce dal primo autovalore, che sintetizza quella di più di 7 indici di partenza, a quella insignificante degli ultimi autovalori.

Tabella 1 – Output dell'ACP: autovalori e % spiegata.

Varianza totale spiegata						
Componente	Autovalori iniziali			Pesi dei fattori non ruotati		
	Totale	% di varianza	% cumulata	Totale	% di varianza	% cumulata
1	7,459	43,877	43,877	7,459	43,877	43,877
2	2,293	13,491	57,368	2,293	13,491	57,368
3	1,896	11,154	68,523	1,896	11,154	68,523
4	1,467	8,627	77,150	1,467	8,627	77,150
5	,883	5,195	82,345			
6	,767	4,511	86,856			
7	,653	3,839	90,695			
8	,470	2,763	93,458			
9	,393	2,315	95,773			
10	,241	1,418	97,191			
11	,163	,961	98,152			
12	,135	,791	98,944			
13	,082	,481	99,424			
14	,051	,299	99,723			
15	,029	,171	99,894			
16	,015	,091	99,985			
17	,003	,015	100,000			

Metodo di estrazione: Analisi componenti principali.

L'ACP – 3

- Un criterio per scegliere quanti autovalori, e quindi di quante componenti tener conto, è di considerare solo quelli superiori all'***unità***: infatti questi spiegano più di quanto non spieghi una qualsiasi delle variabili di partenza (che sono standardizzate). Gli indici sono standardizzati perché calcolati con unità di misura diverse.
- Nel caso le unità di misura fossero le stesse e si volesse considerare la ***covarianza***, allora si sceglierebbero gli autovalori maggiori di n (numero degli indici) volte la media degli autovalori stessi: è il caso, ad esempio, di un'analisi su variabili economiche tutte espresse in euro.

L'ACP – 4

- Gli autovalori scelti sono pertanto 4 e, complessivamente, spiegano più del 77% della variabilità presente nello spazio originario a 17 dimensioni: c'è quindi un risparmio di dimensioni (da 17 a 4), a fronte di una perdita di informazione (il 23%).
- Per vedere rispetto a quali indicatori questa perdita di informazione è maggiore, si può consultare la tabella 2.

Tabella 2 – Capacità delle componenti scelte di rappresentare la variabilità degli indici di partenza.		
Comunalità		
	Iniziale	Estrazione
Speranza di vita a 0 anni - m	1,000	,900
Speranza di vita a 0 anni - f	1,000	,837
Speranza di vita a 0 anni in buona salute - m	1,000	,881
Speranza di vita a 0 anni in buona salute – f	1,000	,875
Indice di stato fisico	1,000	,755
Indice di stato psicologico	1,000	,886
Tasso di moralità infantile	1,000	,825
Tasso standard di mortalità per accidenti	1,000	,591
Tasso standard di mortalità per tumori	1,000	,770
Tasso standard di mortalità per demenze sistema nervoso	1,000	,462
Speranza di vita senza limitazioni ai 65 anni - m	1,000	,779
Speranza di vita senza limitazioni ai 65 anni - f	1,000	,815
Eccesso peso	1,000	,673
Fumo	1,000	,567
Alcol	1,000	,764
Sedentarietà	1,000	,898
Alimentazione	1,000	,837
Metodo di estrazione: Analisi componenti principali.		

L'ACP – 5

- Nella seconda colonna c'è la variabilità uniforme e unitaria di tutti gli indici; nella terza, invece, quella che è in comune (***comunalità***) con gli altri, spiegata dalle componenti prescelte (le prime 4).
- Il risultato è confortante: solo 3 indici hanno una comunalità inferiore a .75, ovvero il tasso di mortalità per accidenti nei trasporti, quello per malattie del sistema nervoso e quello dei fumatori dai 14 anni in poi.

L’ACP – 6

- Effettivamente questi tre aspetti sembrano poco legati al funzionamento di un sistema sanitario, che è poi quello che vogliamo misurare con gli indicatori scelti; approfondendo l’analisi potremmo scoprire che ognuno di questi indici è legato a una singola componente specifica (***specificità***).

L'ACP – 7

- Abbiamo così individuato che l'insieme degli indici rilevati sull'area Salute può essere sintetizzato da sole 4 componenti, spiegando più del 77% della variabilità nelle regioni, e che la prima di queste componenti ne spiega, da sola, quasi il 44%.
- Ora dobbiamo cercare di vedere le relazioni fra componenti e indici di partenza, capire - se è possibile - il significato delle componenti e ***battezzarle***, ovvero assegnare loro un'etichetta per poterle riutilizzare come indicatori nel prosieguo delle analisi.

Tabella 3 – Correlazioni degli indici originari con le componenti scelte.

Matrice di componenti[a]				
	Componente			
	1	2	3	4
Speranza di vita a 0 anni - m	,566	-,401	,619	-,191
Speranza di vita a 0 anni - f	,613	-,492	,457	-,100
Speranza di vita a 0 anni in buona salute - m	,795	,387	,284	,134
Speranza di vita a 0 anni in buona salute – f	,785	,356	,132	,337
Indice di stato fisico	,766	,397	-,035	-,101
Indice di stato psicologico	,621	,234	-,107	,659
Tasso di moralità infantile	-,813	,080	,297	,264
Tasso standard di mortalità per accidenti	,366	-,632	-,217	,103
Tasso standard di mortalità per tumori	-,372	,610	-,443	-,251
Tasso standard di mortalità per demenze sistema nervoso	,446	-,044	-,499	-,108
Speranza di vita senza limitazioni ai 65 anni - m	,328	,531	,543	-,309
Speranza di vita senza limitazioni ai 65 anni - f	,875	,173	-,095	,104
Eccesso peso	-,768	-,071	,183	,211
Fumo	-,367	,532	,276	-,271
Alcol	,800	-,006	-,266	,232
Sedentarietà	-,923	,084	-,077	,180
Alimentazione	,567	-,118	-,351	-,615
Metodo estrazione: analisi componenti principali. a. 4 componenti estratti				

L'ACP – 8

- Questo passaggio si può ottenere calcolando una sorta di correlazione fra componente e indici originari, che è riportato nella cosiddetta **matrice delle componenti.**
- Come si può notare la prima componente cattura variabilità comune da quasi tutti gli indici: ha un andamento discordante con la mortalità infantile, con il tasso di sedentarietà e con il tasso di obesità; concordante con tutte le speranze di vita e gli indici di salute.

L'ACP – 9

- Potremmo definirla "livello del sistema sanitario" e assumerla come indicatore concordante con tale aspetto.
- A questo punto, utilizzando le proprietà degli spazi vettoriali, possiamo proiettare i punti unità su questa dimensione: ogni regione avrà una coordinata sulla nuova retta e le regioni con i punteggi più alti avranno un livello più alto del sistema sanitario.

L'ACP – 10

- La graduatoria della regioni sulla ***prima componente*** è la seguente:

Bolzano 2,17; Trento 1,45; Valle d'Aosta 1,37; Friuli-Venezia Giulia 0,78; Veneto 0,61; Lombardia 0,55; Piemonte 0,54; Toscana 0,49; Emilia-Romagna 0,30; Liguria 0,25;

Marche -0,06; Abruzzo -0,12; Umbria -0,16; Sardegna -0,34; Lazio -0,34; Molise -0,55; Basilicata -0,83; Puglia -1,04; Sicilia -1,57; Calabria -1,64; Campania -1,68.

L'ACP – 11

- Come si può vedere si tratta di una variabile standardizzata: valori positivi indicano una situazione migliore della media; negativi peggiore; la media dei valori regionali ha coordinata 0, ma quella effettiva italiana, essendo ponderata per il numero di abitanti con 14 anni e più, si va a collocare nel versante negativo a -0,18, dove le più popolose regioni meridionali la *attraggono*.

L'ACP – 12

- Un'anomalia che evidenziamo in questa analisi è la correlazione molto alta del tasso standardizzato di "popolazione con un comportamento a rischio nel consumo di alcol" con la prima componente.
- È uno dei casi, già segnalati, di correlazione spuria: non è il maggior consumo di alcol a migliorare il livello sanitario di una regione, ma probabilmente questa abitudine ha maggior incidenza nelle zone montane, che sono prevalenti nelle regioni settentrionali, e lì il sistema sanitario è migliore.