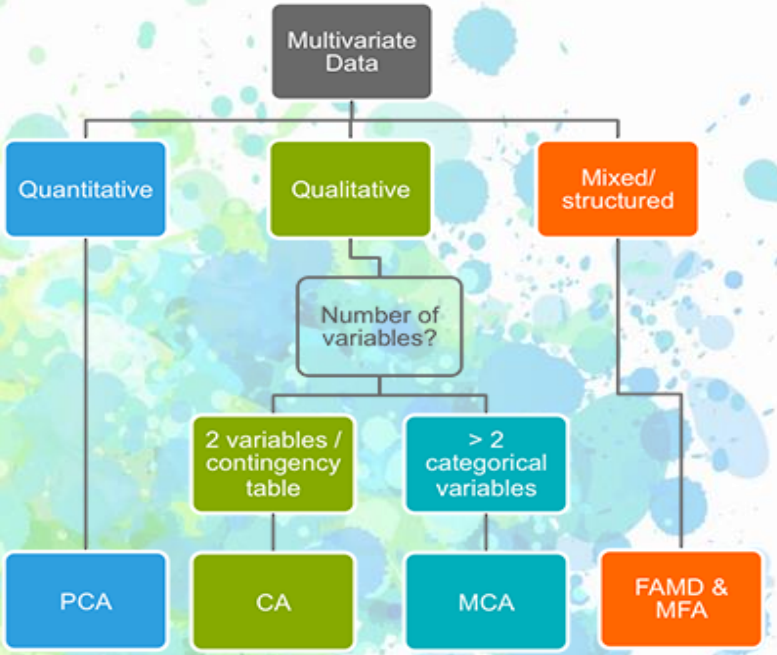
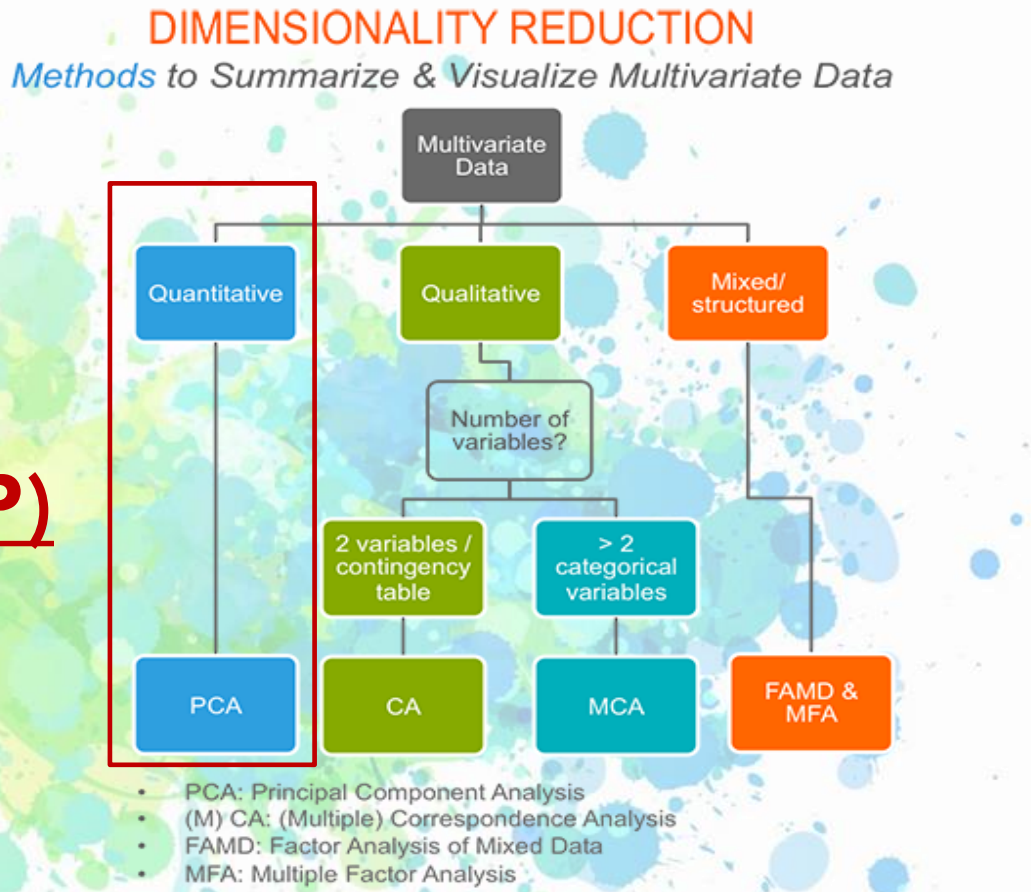
DIMENSIONALITY REDUCTION

Methods to Summarize & Visualize Multivariate Data



- PCA: Principal Component Analysis
- (M) CA: (Multiple) Correspondence Analysis
- FAMD: Factor Analysis of Mixed Data
- MFA: Multiple Factor Analysis

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)



L’ ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

L'analisi dei componenti principali, o ACP, è uno dei primi metodi nati con l’obiettivo di analizzare i dati tenendo conto del loro carattere multidimensionale

L'ACP è un metodo di **riduzione della dimensionalità** che viene spesso utilizzato per ridurre la dimensionalità di **grandi insiemi di dati**, **trasformando un grande insieme di variabili in uno più piccolo che contiene ancora la maggior parte delle informazioni nell'insieme di grandi dimensioni**

L'ACP consente di **sintetizzare l’informazione raccolta con un numero elevato di variabili cardinali fra loro correlate, attraverso un numero di solito molto più piccolo di nuove variabili, dette componenti principali**, ognuna delle quali esprime una combinazione lineare delle variabili originarie.

$A_{(n \times p)}$ ⇨ $B_{(n \times c)}$

N individui
P variabili
C componenti principali

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

Spesso, nella vita di tutti i giorni, ci capita di dover sintetizzare delle informazioni

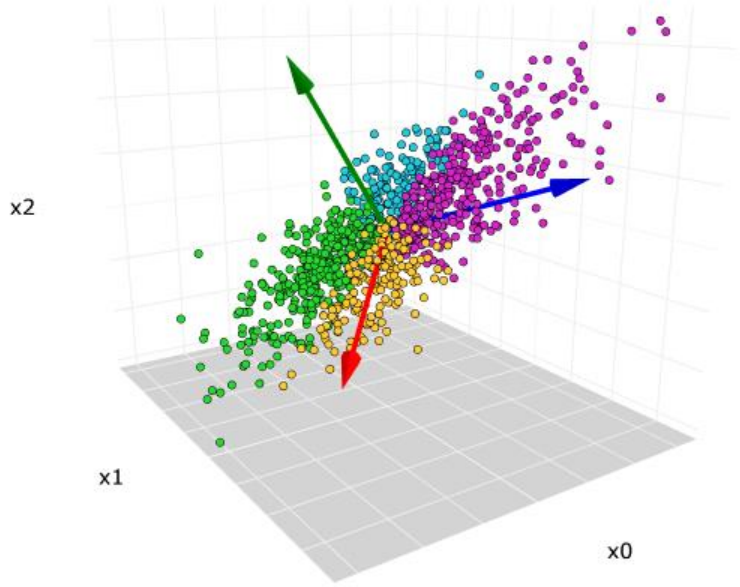
Ad esempio, quando presentiamo le caratteristiche di una persona possiamo effettuarne una descrizione minuziosa (la sua professione, il suo reddito, i suoi consumi, le sue attività nel tempo libero, etc.), oppure possiamo estrapolare da tutte queste caratteristiche quella o quelle dimensioni che le riassumono al meglio

In altri termini, anziché dire che Claudio è un avvocato, che guadagna cinque milioni al mese, che vive in una villa in un quartiere residenziale, che possiede una Ferrari, etc., possiamo dire che **Claudio è un benestante**

L’ ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

Il concetto di benestante equivale a una componente principale che riassume l’insieme delle caratteristiche di Claudio, cercando di rendere minima la perdita di informazioni.

Obiettivo dell’ACP è esattamente questo: **ridurre un insieme di informazioni alle sue componenti principali minimizzando la perdita di informazioni**

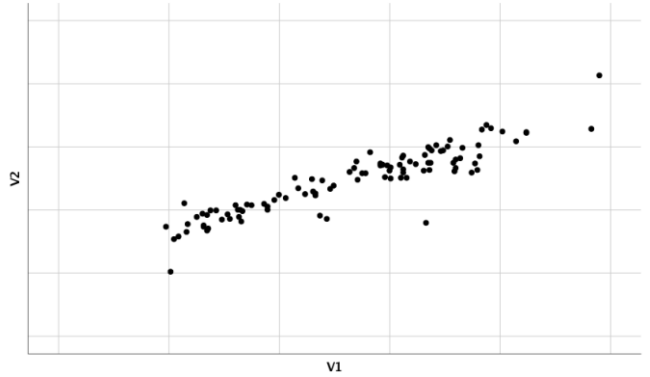


La riduzione del numero di variabili di un set di dati va naturalmente a scapito dell'accuratezza

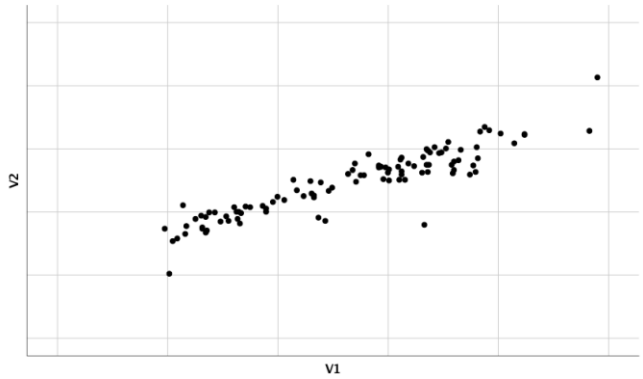
L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

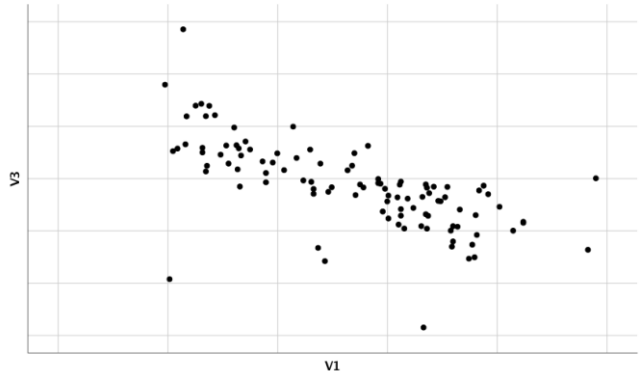
FACCIAMO UN PASSO INDIETRO….

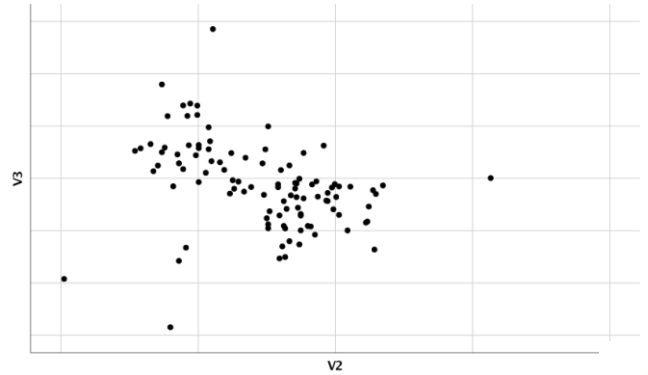
Quando si hanno solo **due variabili**, V1 and V2, l'interpretazione del grafico è, in genere, molto semplice.



Se le variabili sono **tre**, V1, V2 e V3, possiamo studiare i tre grafici piani che risultano dalle possibili combinazioni delle tre variabili osservate, ognuno dei quali rappresenta solo una parte dell'informazione contenuta nei dati.







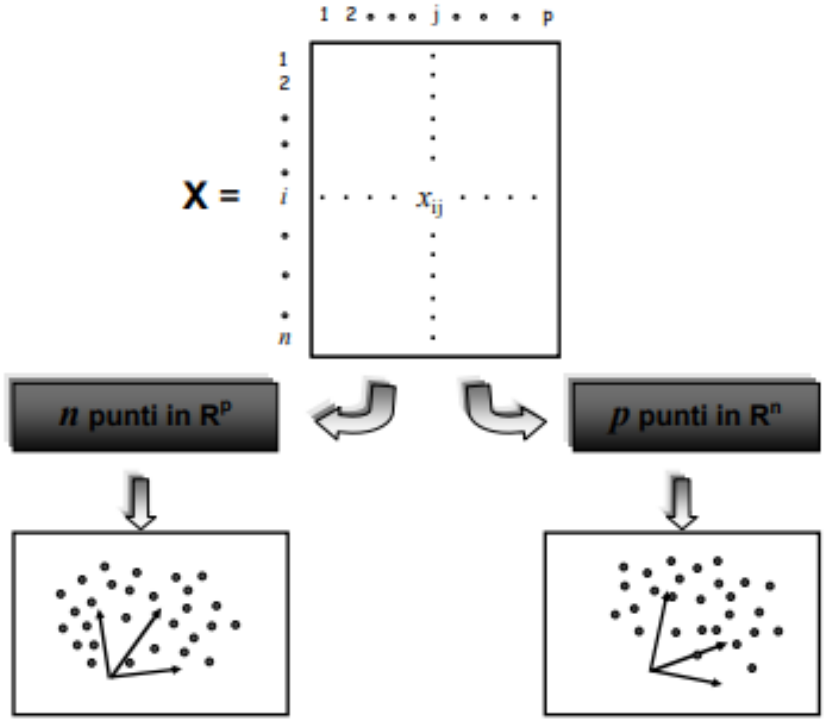
…..e se abbiamo **p** variabili?

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

DA UN PIANO AD UNO SPAZIO

Quando gli individui sono descritti da un **numero elevato di variabili**, diventa necessario disporre di uno strumento specifico che consenta di "esplorare" lo spazio in cui questi si trovano.

Analizzare i **punti-individuo** considerando tutte le variabili in modo simultaneo, significa identificare le loro **similarità**, per poter definire possibili, differenti tipologie.



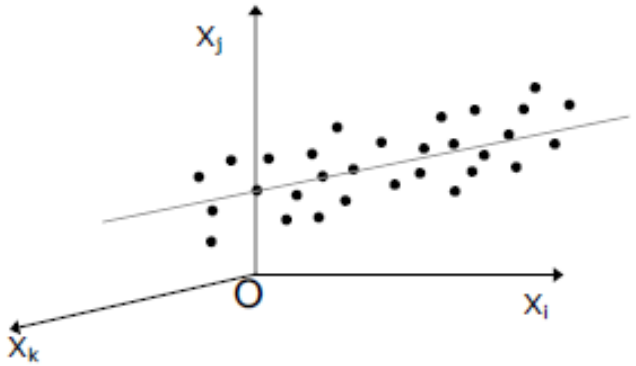
Allo stesso tempo, quando le variabili sono numerose è necessario individuare uno strumento in grado di sintetizzare e rappresentare, su grafici piani appropriati, le principali **relazioni** tra le variabili osservate.

L’ ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

DA UN PIANO AD UNO SPAZIO

Obiettivo dell'ACP è individuare dei nuovi "**fattori**", ottenuti come **combinazioni lineari delle variabili di partenza**, in grado di **sintetizzare al meglio l'informazione contenuta nella matrice dei dati di partenza** e di **visualizzare le relazioni tra le variabili e/o tra le unità, su grafici piani appropriati**

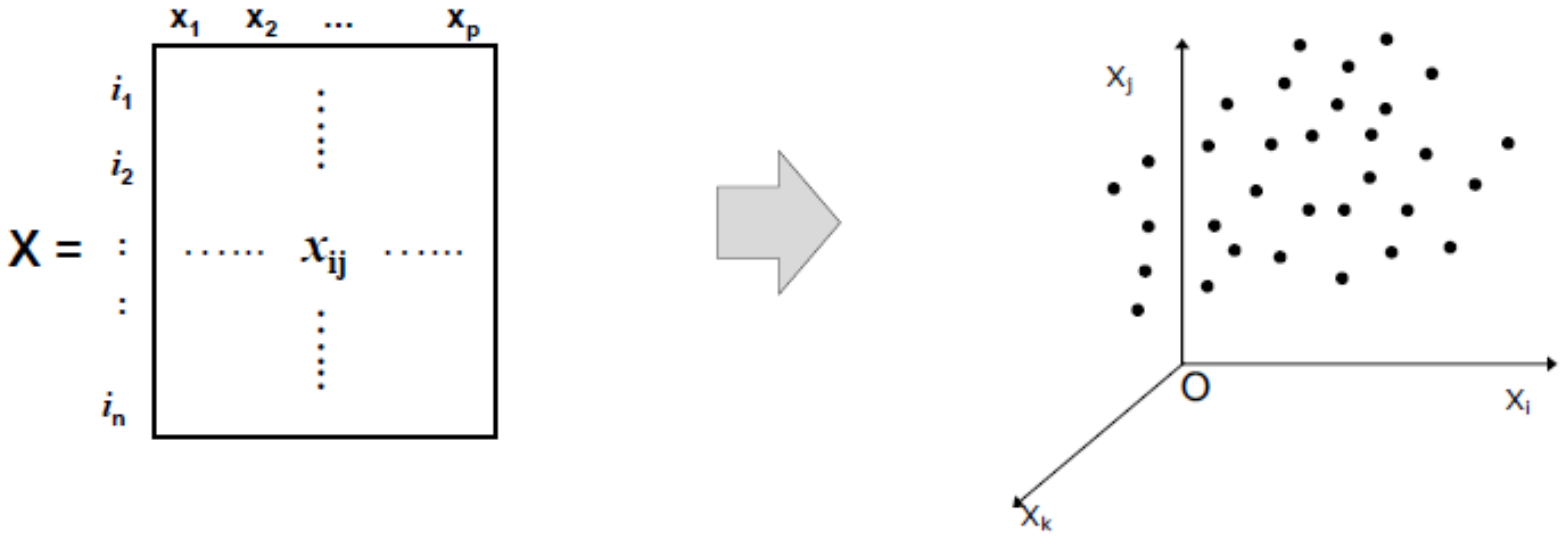
Obiettivo è determinare il miglior asse, poi il miglior piano, e così via, che consenta di ottenere la rappresentazione delle distanze degli n punti il più vicino possibile a quella nello spazio originario.



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

I PASSI PRINCIPALI

1. Consideriamo ***n* individui**, sui quali sono state osservate ***p* variabili quantitative**.



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

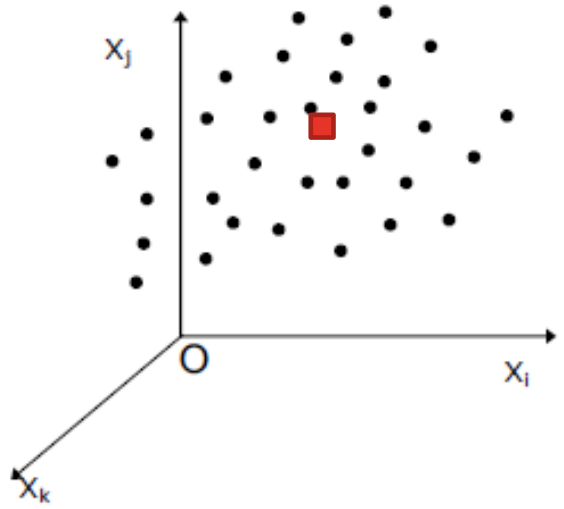
I PASSI PRINCIPALI

1 Consideriamo ***n* individui**, sui quali sono state osservate ***p* variabili quantitative**.

2 Da un punto di vista geometrico, **gli individui possono essere considerati come n punti di uno spazio a *p* dimensioni (R^P)**

X=

Sigla	Città	Abitanti	Reddito	Consumi	NumImp	Salecine	Palestre	Librerie	NO2
AG	Agrigento	54.600	9.864	8.676	9,7	1,9	2,5	3,0	46,5
AL	Alessandria	90.500	20.555	14.151	11,1	3,7	14,1	6,7	52,0
AN	Ancona	101.200	22.311	14.667	10,0	4,1	19,4	9,5	44,0
AO	Aosta	35.900	24.583	16.475	12,2	1,7	5,8	15,0	33,0
AP	Ascoli Piceno	52.400	18.179	13.015	12,1	3,0	16,5	6,5	37,0
AQ	L'Aquila	66.900	16.165	11.259	9,3	3,0	12,2	11,5	23,0
AR	Arezzo	90.600	19.780	12.602	11,2	5,0	10,3	7,5	32,0
AT	Asti	72.400	18.541	12.395	12,9	4,3	10,9	6,7	52,0
AV	Avellino	54.300	13.170	9.451	9,4	0,9	5,4	7,3	63,0
BA	Bari	341.300	14.771	10.794	9,5	2,5	7,3	6,9	35,0
BG	Bergamo	115.700	20.607	12.963	8,6	2,2	7,7	5,8	47,0
:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:
VE	Venezia	308.700	21.588	14.358	9,8	3,2	15,2	9,9	40,0
VI	Vicenza	107.100	24.067	14.254	10,3	3,6	12,9	6,3	49,0
VR	Verona	252.700	24.377	16.733	11,0	2,6	11,8	8,6	46,0
VT	Viterbo	58.400	15.855	11.724	13,1	3,4	8,6	9,6	30,0
VV	Vibo Valentia	31.500	10.071	5.113	7,2	2,3	2,8	2,8	36,0
MEDIA		*172.203*	*18.309*	*12.406*	*10,0*	*3,2*	*10,9*	*8,4*	*44,0*



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

I PASSI PRINCIPALI

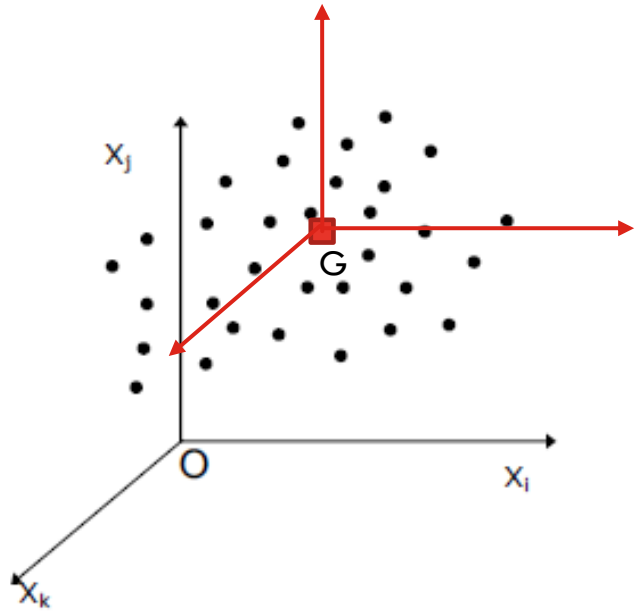
1 Consideriamo ***n* individui**, sui quali sono state osservate ***p* variabili quantitative**.

2 Da un punto di vista geometrico, **gli individui possono essere considerati come n punti di uno spazio a *p* dimensioni (R^P)**

3 Le **variabili** vengono **centrate**: $x_{ij}^* = x_{ij} - M(x_j)$

$X^*=$

Sigla	Città	Abitanti	Reddito	Consumi	Numimp	Salecine	Palestre	Librerie	NO2
AG	Agrigento	-117.603	-8.445	-3.730	-0,3	-1,2	-8,4	-5,5	2,5
AL	Alessandria	-81.703	2.246	1.744	1,1	0,5	3,2	-1,7	8,0
AN	Ancona	-71.003	4.002	2.261	0,0	0,9	8,5	1,1	0,0
AO	Aosta	-136.303	6.274	4.068	2,2	-1,5	-5,1	6,6	-11,0
AP	Ascoli Piceno	-119.803	-130	608	2,1	-0,2	5,6	-1,9	-7,0
AQ	L'Aquila	-105.303	-2.144	-1.148	-0,7	-0,2	1,2	3,1	-21,0
AR	Arezzo	-81.603	1.471	195	1,2	1,8	-0,6	-0,9	-12,0
AT	Asti	-99.803	232	-12	2,9	1,1	0,0	-1,8	8,0
AV	Avellino	-117.903	-5.139	-2.955	-0,6	-2,2	-5,5	-1,2	19,0
BA	Bari	169.097	-3.538	-1.613	-0,5	-0,7	-3,7	-1,5	-9,0
BG	Bergamo	-56.503	2.297	557	-1,4	-1,0	-3,2	-2,7	3,0
:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:
VE	Venezia	136.497	3.279	1.951	-0,2	0,0	4,3	1,5	-4,0
VI	Vicenza	-65.103	5.758	1.848	0,3	0,4	2,0	-2,1	5,0
VR	Verona	80.497	6.068	4.327	1,0	-0,6	0,8	0,2	2,0
VT	Viterbo	-113.803	-2.454	-683	3,1	0,3	-2,4	1,2	-14,0
VV	Vibo Valentia	-140.703	-8.238	-7.294	-2,8	-0,9	-8,1	-5,6	-8,0
MEDIA		*0,0*	*0,0*	*0,0*	*0,0*	*0,0*	*0,0*	*0,0*	*0,0*



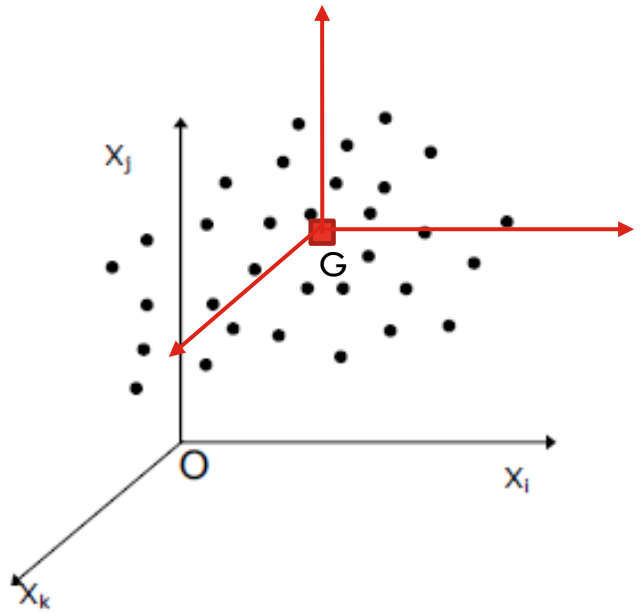
La centratura comporta una **traslazione** del sistema di riferimento nel **baricentro della nuvola di punti**

L’OPERAZIONE DI CENTRATURA

La centratura è una trasformazione che avviene sulle **colonne** della matrice

Questo vincolo comporta che in $\mathbf{R^n}$ i p punti variabile siano in realtà in uno spazio di dimensioni n-1 mentre in $\mathbf{R^p}$ le dimensioni non mutano poiché non vi sono vincoli sulle righe della matrice

Da un punto di vista geometrico, l’operazione di centratura, che in $\mathbf{R^p}$ comporta una traslazione del sistema di riferimento nel baricentro della nube degli individui (che rimangono comunque in uno spazio a p dimensioni), in $\mathbf{R^n}$ si traduce in una proiezione dei p punti-variabile in uno spazio di dimensioni n-1



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

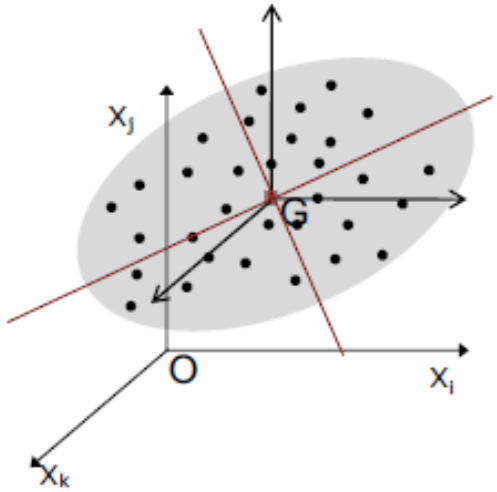
I PASSI PRINCIPALI

1. Consideriamo ***n* individui**, sui quali sono state osservate ***p* variabili quantitative**.

2. Da un punto di vista geometrico, **gli individui possono essere considerati come n punti di uno spazio a *p* dimensioni (R^P)**

3. Le **variabili** vengono **centrate**: $x_{ij}^* = x_{ij} - M(x_j)$

4. L'obiettivo è individuare l'asse che consenta la **migliore rappresentazione unidimensionale** della nuvola di punti, conservando al meglio la variabilità originaria e **minimizzando la perdita di informazione**



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

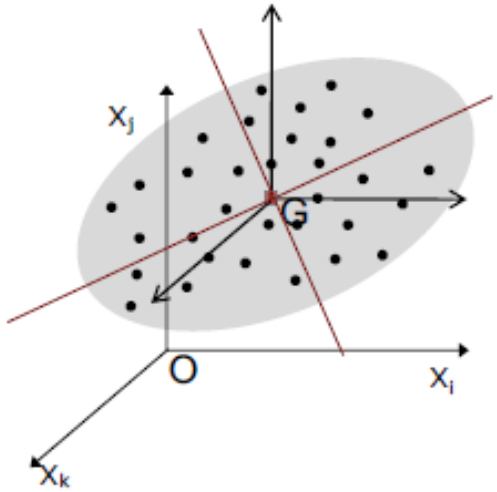
I PASSI PRINCIPALI

1. Consideriamo ***n* individui**, sui quali sono state osservate ***p* variabili quantitative**.

2. Da un punto di vista geometrico, **gli individui possono essere considerati come n punti di uno spazio a *p* dimensioni (R^P)**

3. Le **variabili** vengono **centrate**: $x_{ij}^* = x_{ij} - M(x_j)$

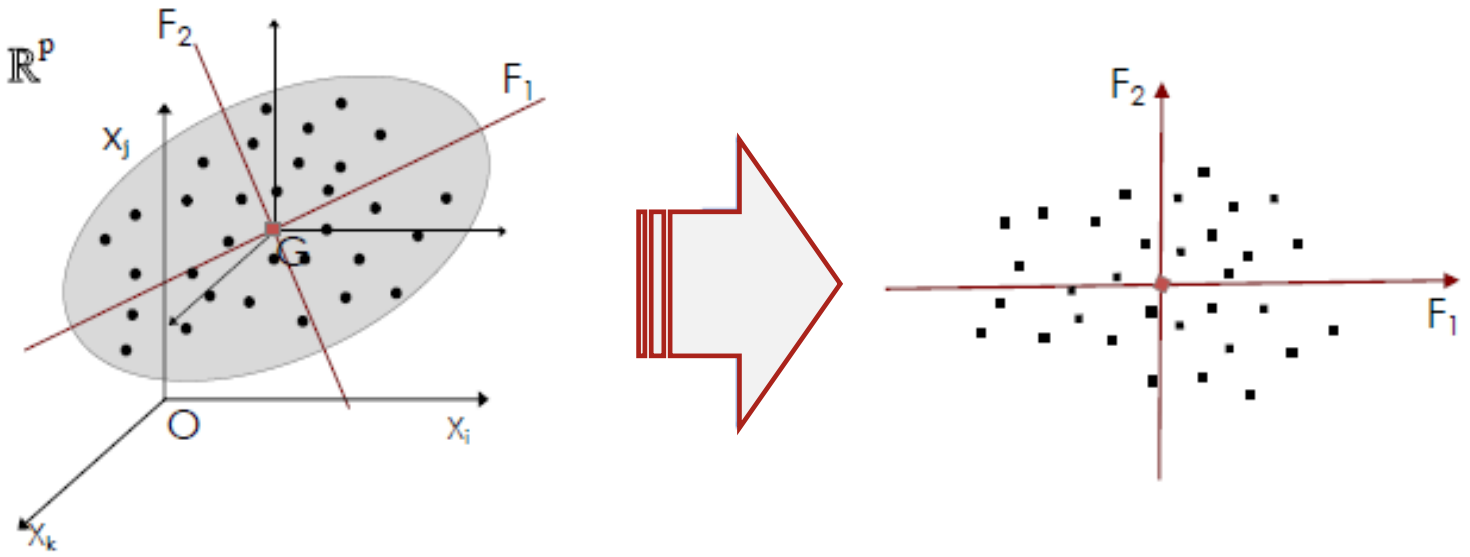
4. L'obiettivo è individuare l'asse che consenta la **migliore rappresentazione unidimensionale** della nuvola di punti, conservando al meglio la variabilità originaria e **minimizzando la perdita di informazione**



5. Quando il numero di variabili osservate è elevato, **un singolo asse potrebbe non essere sufficiente a rappresentare l'informazione** contenuta nella matrice dei dati. È allora possibile determinare **altri assi**, tra loro **ortogonali**, ognuno dei quali rappresenterà **una parte dell'informazione**, non spiegata dagli altri.

L’ ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

DEFINIZIONE DEL PIANO FATTORIALE



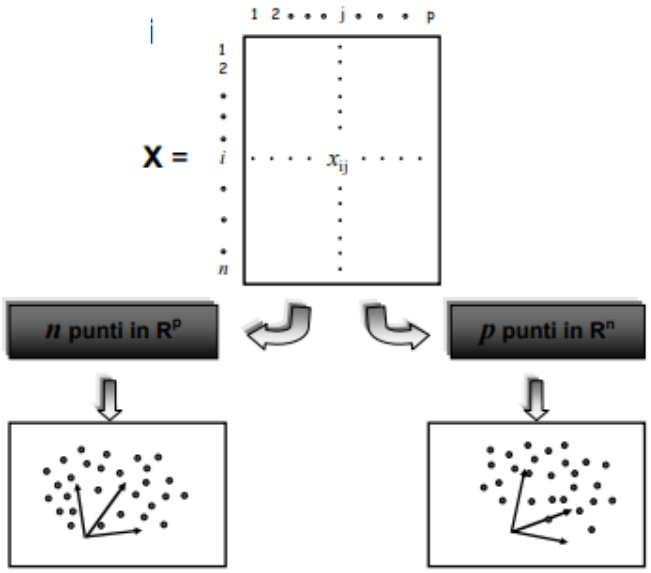
Spazio originario

Piano fattoriale

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

DOPPIA ANALISI

Data la matrice X, di dimensioni nxp, da un punto di vista geometrico, è possibile considerare i **punti individuo** come n punti di uno spazio a p dimensioni (**R^p**, *Spazio delle variabili*);



D'altra parte, è possibile immaginare le **p variabili** come elementi di uno spazio a n dimensioni, in cui ogni asse rappresenta un individuo (**R^n**, *Spazio degli individui*);

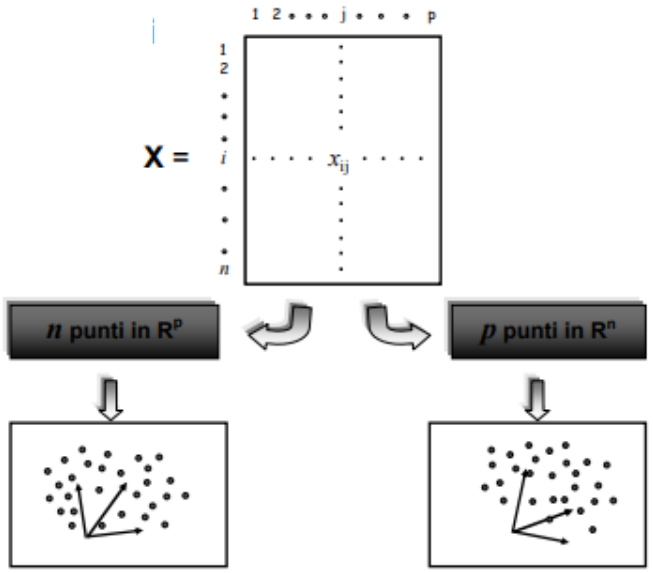
Mentre in **R^p** ci si interessa alla **distanza tra punti-unità**, in **R^n** si guarda piuttosto all'**angolo formato da coppie di vettori variabile** (coseno) analizzando la **sfera delle correlazioni**

Il **coseno al quadrato** è una misura della **qualità della rappresentazione** dei punti-unità sul sottospazio generato dai fattori scelti: quanto più risulta prossimo ad 1 tanto più il punto avrà conservato, in proiezione, la distanza dall'origine che aveva nello spazio iniziale, e risulterà quindi ben rappresentato

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

ANALISI PUNTI-INDIVIDUO

Analizzare i **punti-individuo** considerando tutte le variabili in modo simultaneo, significa identificare le loro **similarità**, per poter definire possibili, differenti tipologie.



La **similarità** viene espressa attraverso la **distanza** tra i punti

Quando le variabili sono espresse in unità di misura diverse, o anche quando siano espresse nella stessa unità di misura ma con campi di variazione molto diversi, la distanza tra due individui può essere **fortemente condizionata** da alcune di queste variabili

Si può allora scegliere di fare giocare, a ciascuna variabile, un **ruolo identico** nella definizione delle prossimità tra punti individuo. Questo può avvenire mediante la **standardizzazione delle variabili**

LA STANDARDIZZAZIONE

Lo scopo di questo passaggio è standardizzare l'intervallo delle variabili iniziali continue in modo che **ciascuna di esse contribuisca in egual modo all'analisi**

Il motivo per cui è fondamentale eseguire la standardizzazione prima dell’ACP, è che questo tipo di analisi è piuttosto sensibile per quanto riguarda le varianze delle variabili iniziali.

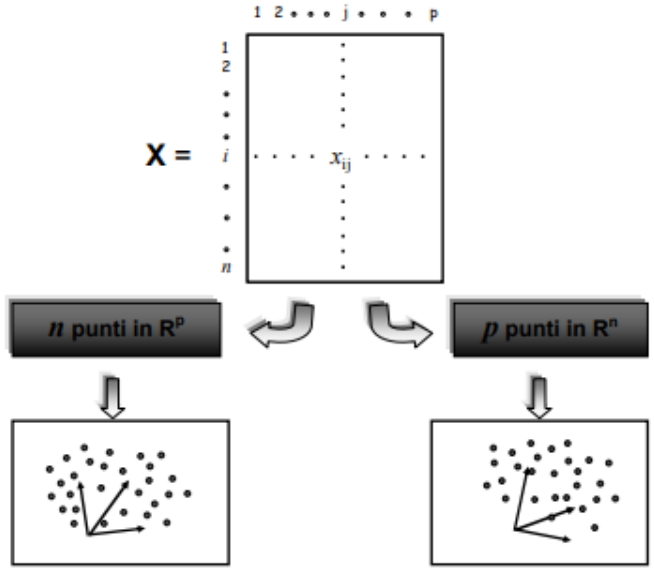
Se ci sono grandi differenze tra gli intervalli delle variabili iniziali, quelle variabili con intervalli più grandi domineranno su quelle con intervalli piccoli (ad esempio, una variabile compresa tra 0 e 100 dominerà su una variabile compresa tra 0 e 1), che porterà a risultati distorti. Quindi, trasformare i dati in scale comparabili può prevenire questo problema.

Matematicamente, questo può essere fatto sottraendo la media e dividendo per la deviazione standard per ogni valore di ciascuna variabile.

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}_{j}}{\sigma_{j}}$$

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

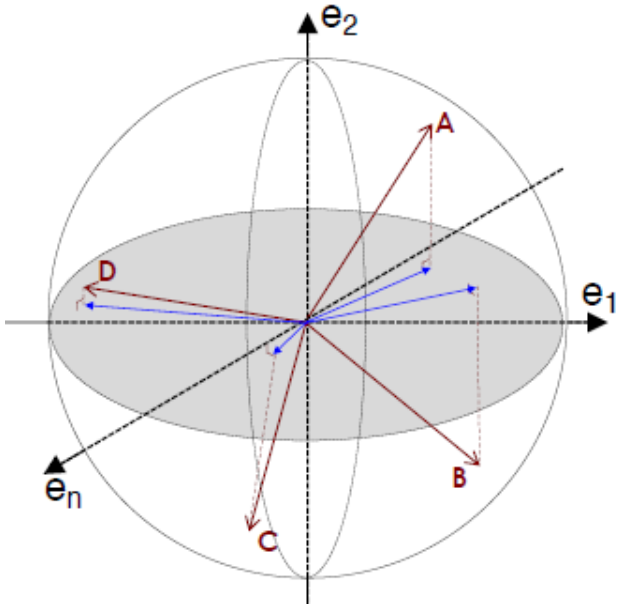
ANALISI PUNTI-VARIABILE



Nello spazio $\mathbf{R^n}$ ogni asse rappresenta un individuo mentre le variabili, tutte a distanza unitaria dall'origine e, dunque, tutte su un'**ipersfera di raggio unitario**, sono rappresentate da **vettori**, i cui **angoli** definiscono la **correlazione** tra le variabili corrispondenti.

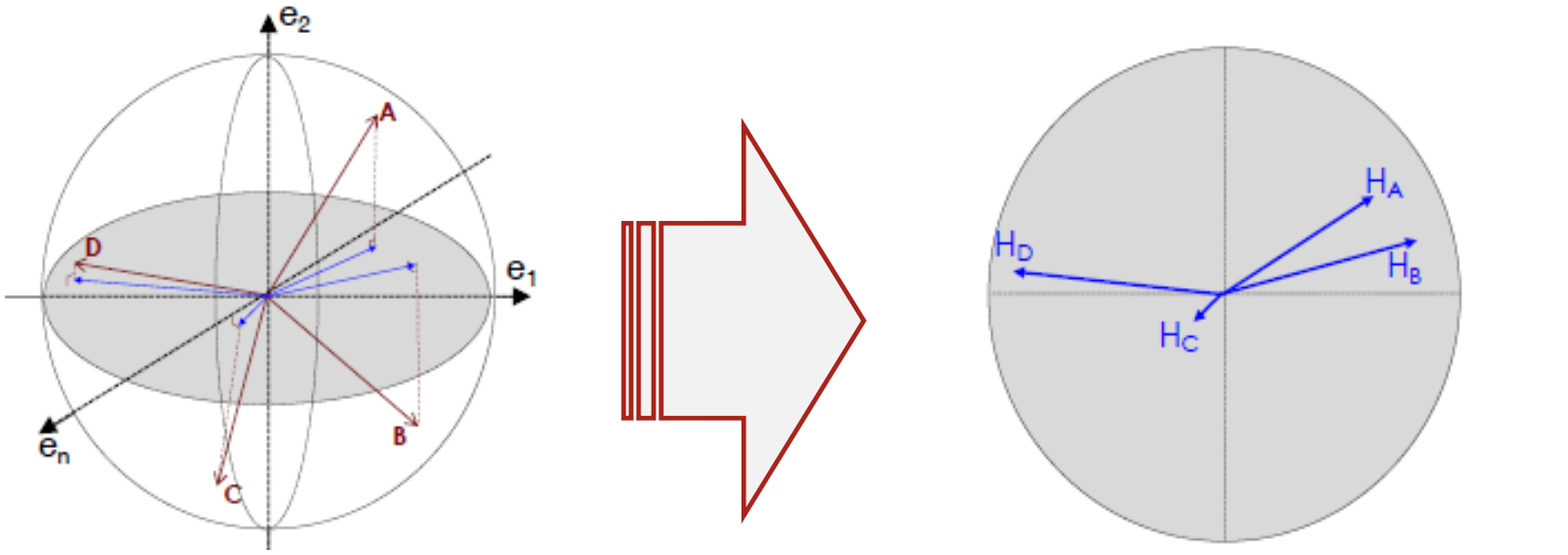
La proiezione delle variabili sul primo piano fattoriale comporterà, dunque, l'osservazione di vettori proiettati tutti all'interno di un cerchio di raggio unitario (poiché, ricordiamo, l'operazione di proiezione "schiaccia" le distanze originarie).

Le **variabili meglio rappresentate** saranno quelle che **conserveranno la distanza più vicina a quella nello spazio originario**, e dunque quelle, in proiezione, **più distanti dall'origine**



L’ ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

DEFINIZIONE DEL CERCHIO DELLE CORRELAZIONI



Spazio originario

Cerchio delle Correlazioni

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

LE COMPONENTI PRINCIPALI

Le componenti principali sono dei vettori-colonna ottenuti attraverso combinazioni lineari dei vettori-colonna che rappresentano le variabili originarie.

Il vettore-colonna che rappresenta una componente principale si chiama **autovettore**.

Ad ogni autovettore è associato un **autovalore** λ_j che è uno scalare (cioè un numero singolo).

L'autovalore esprime l'ammontare della varianza totale riprodotta da una componente.

Gli elementi dell'autovettore (ossia i pesi componenziali elevati al quadrato di ciascuna variabile) **esprimono analiticamente i contributi forniti dalle singole variabili alla componente** in termini di quantità della loro varianza ceduta alla componente stessa.

Se si sommano gli elementi dell'autovettore elevati al quadrato si ottiene esattamente la cifra espressa dall'autovalore.

LE COMPONENTI PRINCIPALI

$$C^1 = u_1^1 x_1 + u_2^1 x_2 + u_3^1 x_3 + \ldots + u_v^1 x_v$$

C^1 è la prima componente principale

I coefficienti u_v sono i *component loadings* (pesi componenziali) delle variabili sulla componente stessa

Questi pesi si possono interpretare come correlazioni fra ciascuna variabile e la componente principale
Essi hanno la proprietà di massimizzare la quota della varianza totale delle v variabili estratta dalla prima componente

Per il calcolo della seconda componente **C^2** bisogna imporre il vincolo di **ortogonalità** (assenza di correlazione lineare) **rispetto alla prima componente**

In forma generale, la generica componente **C^p** si ottiene in questo modo:

$$C^p = u_1^p x_1 + u_2^p x_2 + u_3^p x_3 + \ldots + u_v^p x_v$$

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

LE PROPRIETA' DELLE COMPONENTI PRINCIPALI

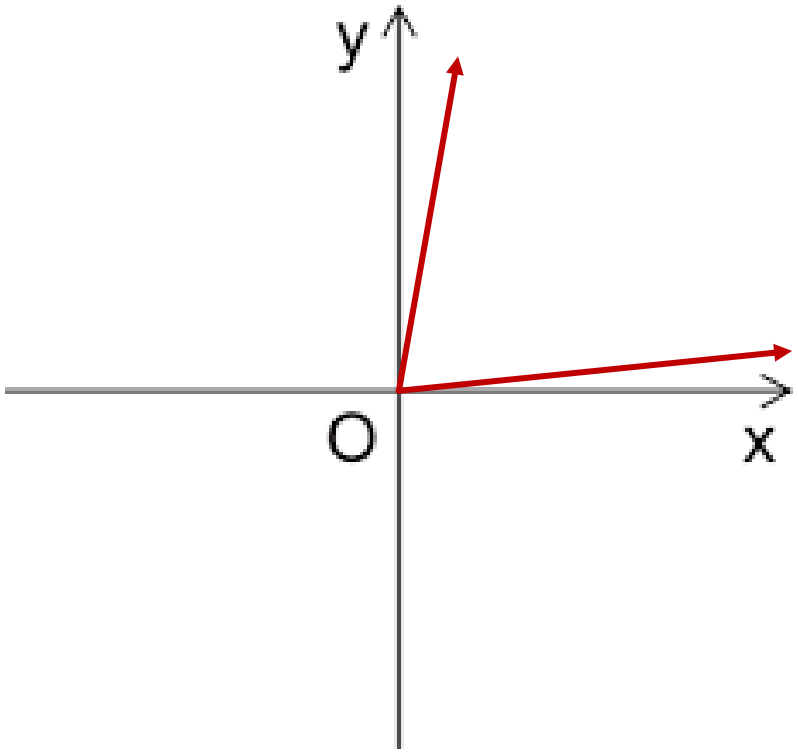
$$C^{p} = u_1^{p} x_1 + u_2^{p} x_2 + u_3^{p} x_3 + \ldots + u_v^{p} x_v$$

1. Due componenti principali qualsiasi sono linearmente indipendenti, in quanto i relativi autovettori sono ortogonali per effetto del procedimento di estrazione
2. L'autovalore λ_j rappresenta la varianza della j-esima componente
3. La somma degli autovalori è uguale alla traccia (ossia alla somma degli elementi diagonali) della matrice di correlazione

L’INTERPRETAZIONE DELLE COMPONENTI PRINCIPALI

Per interpretare le componenti si può ricorrere alla rappresentazione grafica della variabili su piani cartesiani definiti da coppie di componenti.
Le coordinate di ciascuna variabile sono costruite dai suoi pesi sulle due componenti considerate

Ogni asse rappresenta una componente.
Sull’asse delle ascisse si pone la prima componente estratta e sull’asse delle ordinate la seconda componente estratta

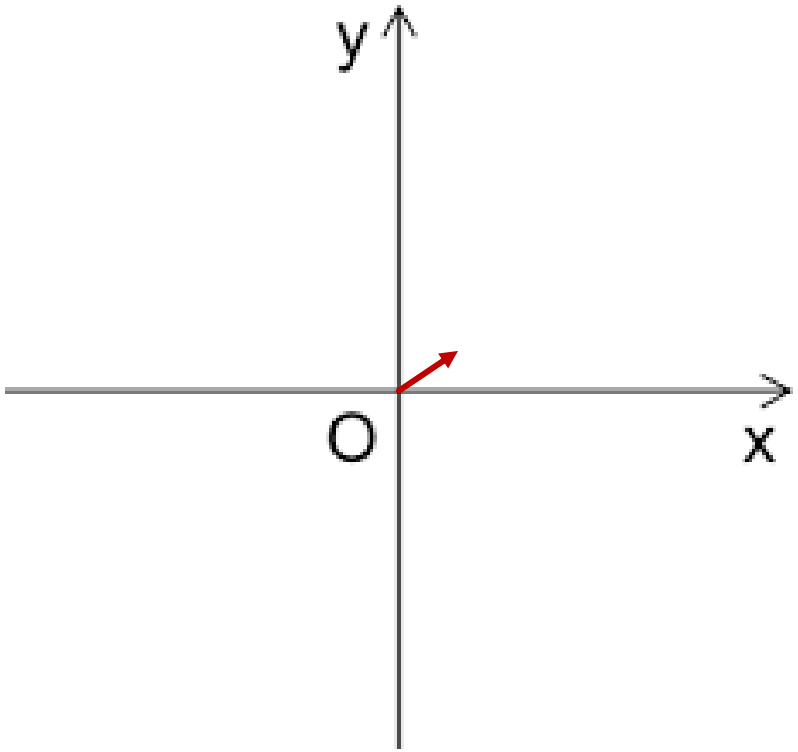


Dato che la posizione di una variabile nel diagramma dipende dai suoi pesi sulle componenti, più questi pesi sono elevati più la variabile si allontanerà dal baricentro. Si allontanerà lungo l’ascissa se ha un alto peso sulla prima componente, lungo l’ordinata se ha un alto peso sulla seconda componente.

L’INTERPRETAZIONE DELLE COMPONENTI PRINCIPALI

Per interpretare le componenti si può ricorrere alla rappresentazione grafica della variabili su piani cartesiani definiti da coppie di componenti.
Le coordinate di ciascuna variabile sono costruite dai suoi pesi sulle due componenti considerate

Ogni asse rappresenta una componente.
Sull’asse delle ascisse si pone la prima componente estratta e sull’asse delle ordinate la seconda componente estratta

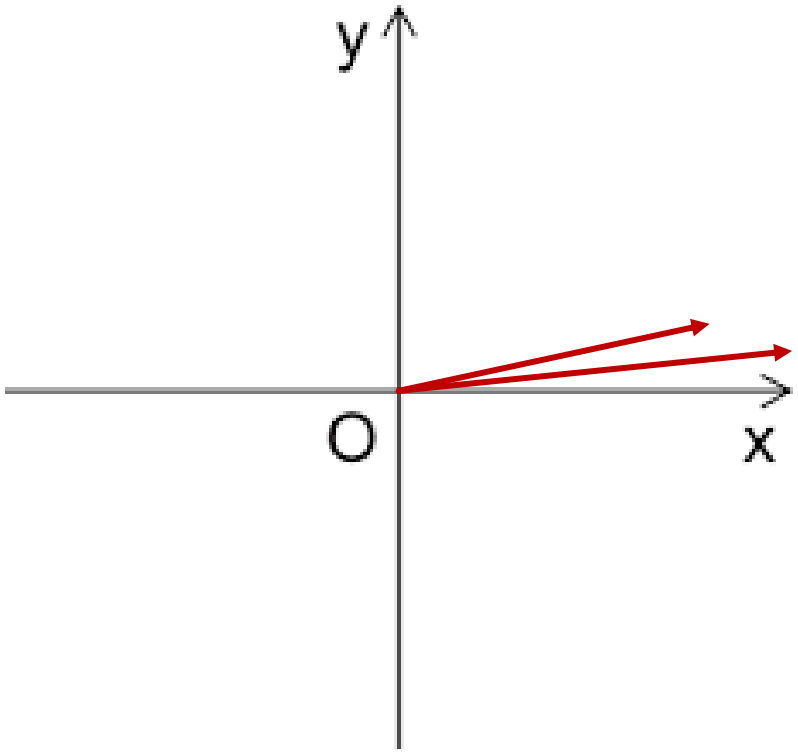


Se invece la variabile presenta pesi modesti (tendenti a zero) su entrambe le componenti considerate la sua posizione sarà prossima al baricentro.

L’INTERPRETAZIONE DELLE COMPONENTI PRINCIPALI

Per interpretare le componenti si può ricorrere alla rappresentazione grafica della variabili su piani cartesiani definiti da coppie di componenti.
Le coordinate di ciascuna variabile sono costruite dai suoi pesi sulle due componenti considerate

Ogni asse rappresenta una componente.
Sull’asse delle ascisse si pone la prima componente estratta e sull’asse delle ordinate la seconda componente estratta

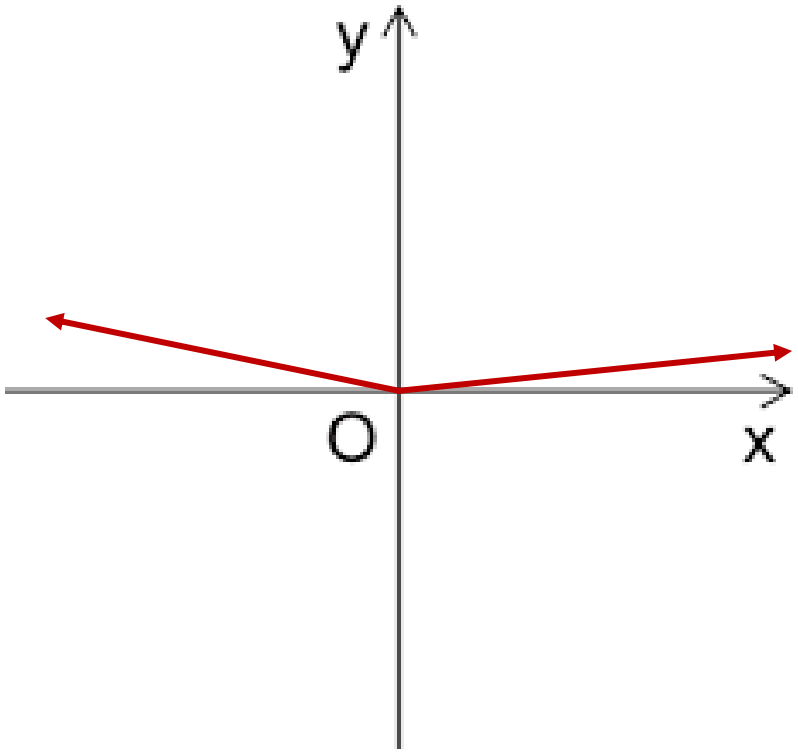


Se due variabili si collocano presso un’estremità del semiasse significa che hanno un’alta correlazione positiva tra loro.

L’INTERPRETAZIONE DELLE COMPONENTI PRINCIPALI

Per interpretare le componenti si può ricorrere alla rappresentazione grafica della variabili su piani cartesiani definiti da coppie di componenti.
Le coordinate di ciascuna variabile sono costruite dai suoi pesi sulle due componenti considerate

Ogni asse rappresenta una componente.
Sull’asse delle ascisse si pone la prima componente estratta e sull’asse delle ordinate la seconda componente estratta

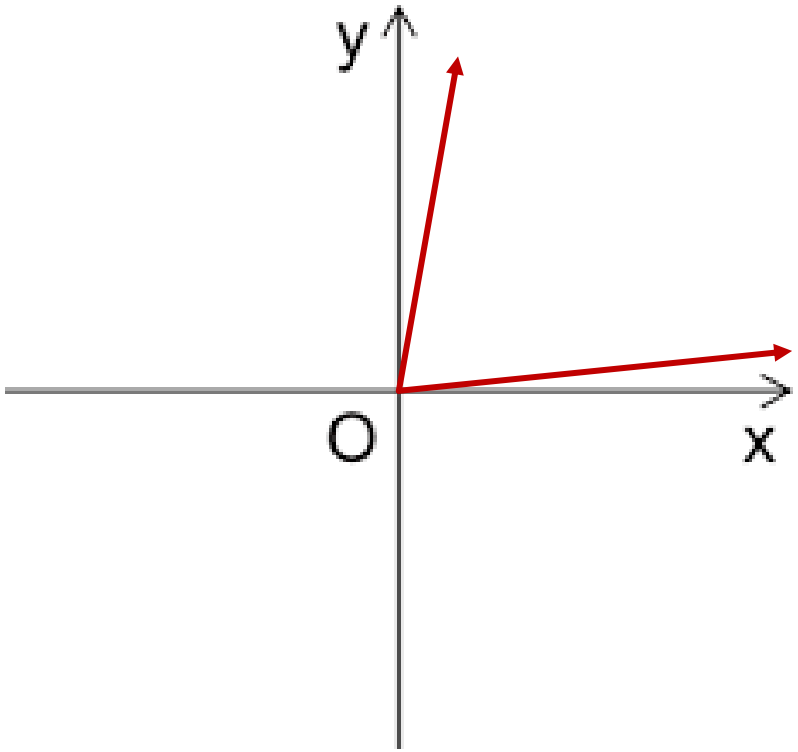


Se invece due variabili si distanziano dal baricentro, ma una sul semiasse positivo e l’altra sul semiasse negativo die una stessa componente vuol dire che fra le due variabili esiste un’alta correlazione negativa e che entrambe presentano pesi elevati su quella componenti (peso positivo per la variabile che si trova nel semiasse positivo e negativo per l’altra)

L’INTERPRETAZIONE DELLE COMPONENTI PRINCIPALI

Per interpretare le componenti si può ricorrere alla rappresentazione grafica della variabili su piani cartesiani definiti da coppie di componenti.
Le coordinate di ciascuna variabile sono costruite dai suoi pesi sulle due componenti considerate

Ogni asse rappresenta una componente.
Sull’asse delle ascisse si pone la prima componente estratta e sull’asse delle ordinate la seconda componente estratta



Se due variabili sono una vicina ad un asse e l’altra al secondo asse significa che tra le due variabili ci sono basse correlazioni e che presentano pesi alti su una componente e bassi sull’altra

I CRITERI PER LA SCELTA DEL NUMERO DI COMPONENTI

a. Criterio della variabilità spiegata

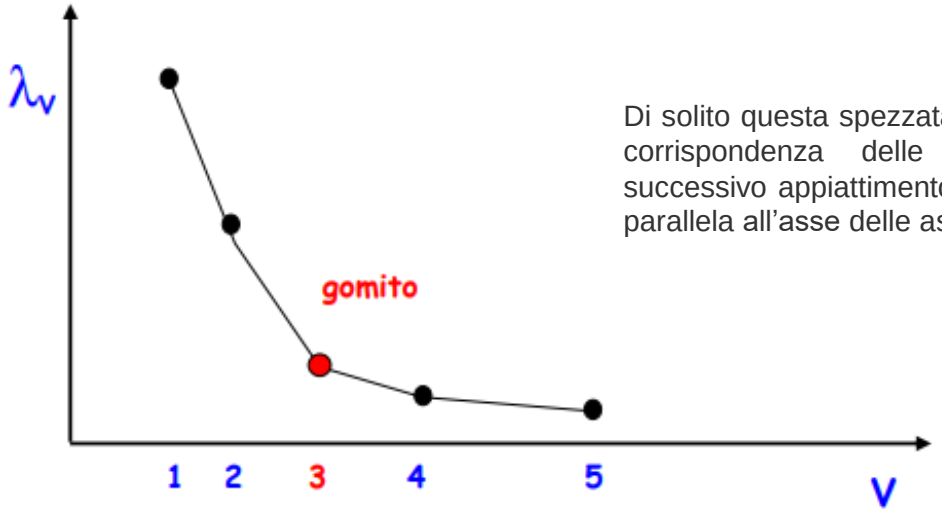
Si considera un numero di componenti tale che esse tengano conto di una percentuale sufficientemente elevata (ad esempio, almeno l'80%) della varianza totale

b. Criterio dell'eigenvalue one (o "Autovalore uno")

Se le variabili sono standardizzate, ciascuna avrà varianza unitaria. In questo caso, si conserveranno solo quei fattori la cui varianza (espressa dal corrispondente autovalore) è maggiore di 1, ossia di quella espressa da una qualsiasi delle variabili originarie.

c. Criterio dello scree test

La percentuale di variabilità spiegata è decrescente e tende a stabilizzarsi su valori poco significativi. Si considereranno, quindi, solo i fattori corrispondenti agli autovalori che precedono la "regolarizzazione" dell'istogramma (considerando che un istogramma "regolare" è indice di uno spazio in cui la nube dei punti è sferica).



Di solito questa spezzata presenta forte inclinazione in corrispondenza delle prime componenti e un successivo appiattimento che la porta ad essere quasi parallela all'asse delle ascisse

Osservazione 1

Non sempre l'andamento del grafico fornisce una risposta univoca, poiché la diminuzione degli autovalori può essere graduale, senza salti evidenti.

Osservazione 2

Alcuni autori suggeriscono di escludere tra le CP scelte quelle sul gomito (Harman, 1976). Altri suggeriscono di includere tra le CP scelte quelle sul gomito (Cattell, 1966).

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

I DATI: PROFILI DELLE CITTÀ METROPOLITANE

PERIODO DI RIFERIMENTO: **ANNI 2020-2022**
DATA DI PUBBLICAZIONE: **02 FEBBRAIO 2023**

Analisi multitematica sulle Città metropolitane – "enti territoriali di area vasta" che hanno sostituito le province in 10 aree urbane di regioni a statuto ordinario – che dispongono di propri organi di governo e di territori coincidenti con quelli delle ex province: Roma, Torino, Milano, Venezia, Genova, Bologna, Firenze, Bari, Napoli e Reggio Calabria. A queste si aggiungono quattro città metropolitane delle regioni a statuto speciale: Palermo, Catania, Messina e Cagliari.

L'analisi è articolata in un set di indicatori chiave che consente di identificare le principali caratteristiche, le diversità o i fattori comuni di questi territori. Sono affrontati alcuni aspetti socio demografici e alcuni elementi di contesto economico fra cui la dinamica della popolazione, l'invecchiamento, la mortalità, le scelte insediative, il mercato del lavoro, il livello di istruzione, il pendolarismo e le caratteristiche del tessuto produttivo

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

I DATI: PROFILI DELLE CITTÀ METROPOLITANE

Città	Invecchiamento	PopStraniera	Occupazione	Pendolarismo	Densità	Reddito
Bari	179,8	3,4	59,1	14,9	74,9	11103,88
Bologna	199,8	11,9	77,2	24,8	93,5	18062,56
Cagliari	226,7	3,6	62,2	23,5	83,0	13265,93
Catania	147,6	3,1	50,5	17,4	63,6	9077,1
Firenze	214,6	12,7	75,9	23,5	99,6	15934,83
Genova	268,7	9,1	72,7	10,8	87,8	16911,15
Messina	202,1	4,4	53,8	11,9	69,4	10444,2
Milano	175,6	14,4	75,8	26,3	106,9	19747,71
Napoli	130,3	4,1	50,6	16,6	66,5	8991,32
Palermo	156,8	2,8	48,9	8,9	56,2	9451,17
Reggio Calabria	168,3	5,5	53,6	11,3	58,6	9377,34
Roma	172,0	11,8	67,2	11,8	87,8	15534,95
Torino	215,4	9,5	72,9	25,6	85,3	15993,31
Venezia	215,3	10,4	72,9	24,6	84,3	14964,71

PERIODO DI RIFERIMENTO: **ANNI 2020-2022**
DATA DI PUBBLICAZIONE: **02 FEBBRAIO 2023**

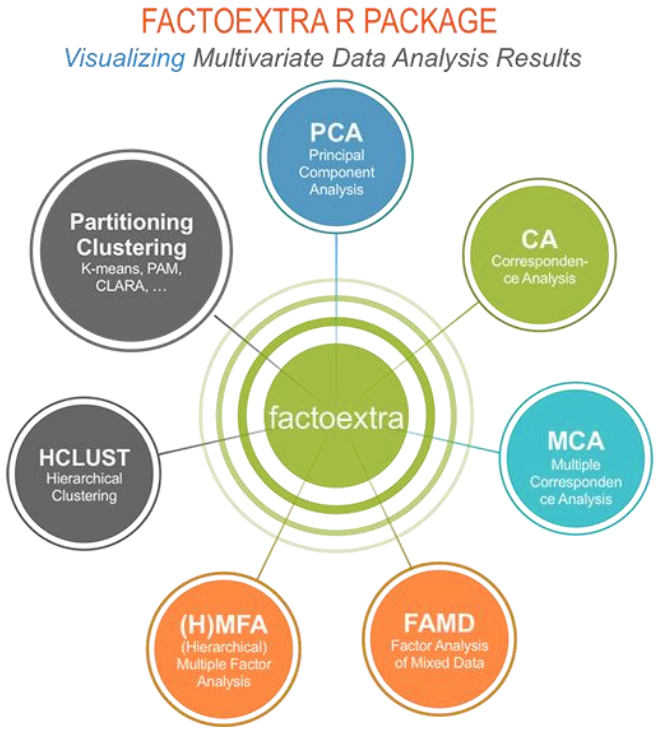
Invecchiamento	Indice di vecchiaia nelle città metropolitane - Anno 2021
PopStraniera	Popolazione straniera residente nelle città metropolitane - Anno 2021 (% sul totale popolazione dell'area)
Occupazione	Tasso di occupazione 25-64 anni totale nelle città metropolitane - Anno 2019 (valori % sulla popolazione 25-64 anni)
Pendolarismo	Persone che si spostano fuori comune per motivi di studio e di lavoro nelle città metropolitane - Anno 2019 (% sul totale popolazione dell'area)
Densità	Densità delle unità locali per 1000 abitanti nelle città metropolitane - Anno 2020 (valori per 1.000 abitanti)
Reddito	Reddito medio pro capite nelle città metropolitane - Anno 2020 (valori in euro)

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

I DATI: PROFILI DELLE CITTÀ METROPOLITANE

 # Libraries

```
library(summarytools)  # summarytools provides a coherent set of functions centered on data exploration and simple reporting
library(tidyverse) # The tidyverse is an opinionated collection of R packages designed for data science
library(FactoMineR) # Exploratory data analysis methods to summarize, visualize and describe datasets
library(factoextra) # Provides some easy-to-use functions to extract and visualize the output of multivariate data analyses
```



factoextra is an R package making easy to extract and visualize the output of exploratory **multivariate data analyses**

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

I DATI: PROFILI DELLE CITTÀ METROPOLITANE

 # Data Import

```
data<- read.csv(file="cittàMetropolitane.csv", header=T, sep=";", dec=",")
```

 # Specify the column with id. Labels

```
data <- as.data.frame(data)
data <- column_to_rownames(data, "Città")
```

	Invecchiamento	PopStraniera	Occupazione	Pendolarismo	Densità	Reddito
Bari	179.8	3.4	59.1	14.9	74.9	11103.88
Bologna	199.8	11.9	77.2	24.8	93.5	18062.56
Cagliari	226.7	3.6	62.2	23.5	83.0	13265.93
Catania	147.6	3.1	50.5	17.4	63.6	9077.10
Firenze	214.6	12.7	75.9	23.5	99.6	15934.83
Genova	268.7	9.1	72.7	10.8	87.8	16911.15
Messina	202.1	4.4	53.8	11.9	69.4	10444.20
Milano	175.6	14.4	75.8	26.3	106.9	19747.71
Napoli	130.3	4.1	50.6	16.6	66.5	8991.32
Palermo	156.8	2.8	48.9	8.9	56.2	9451.17
Reggio Calabria	168.3	5.5	53.6	11.3	58.6	9377.34
Roma	172.0	11.8	67.2	11.8	87.8	15534.95
Torino	215.4	9.5	72.9	25.6	85.3	15993.31
Venezia	215.3	10.4	72.9	24.6	84.3	14964.71

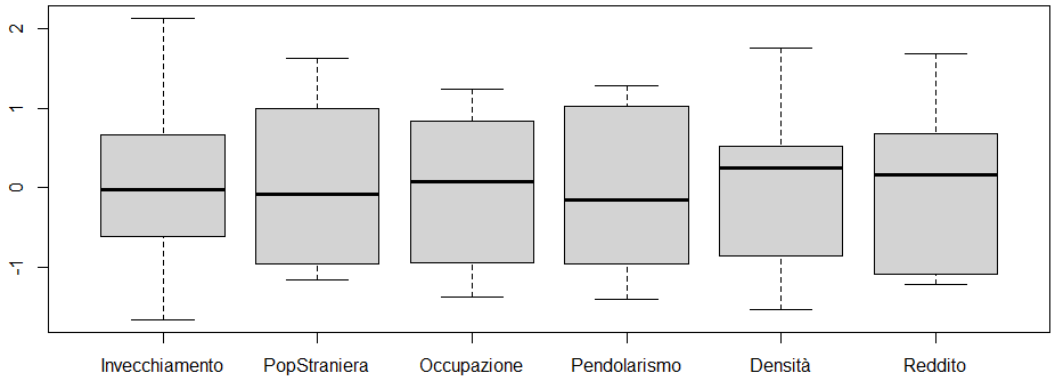
L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

STATISTICHE DESCRITTIVE

```
 Invecchiamento   PopStraniera     Occupazione     Pendolarismo      Densità          Reddito
Min.   :130.3   Min.   : 2.800   Min.   :48.90   Min.   : 8.90   Min.   : 56.20   Min.   : 8991
1st Qu.:169.2   1st Qu.: 3.725   1st Qu.:53.65   1st Qu.:11.82   1st Qu.: 67.22   1st Qu.: 9699
Median :189.8   Median : 7.300   Median :64.70   Median :17.00   Median : 83.65   Median :14115
Mean   :190.9   Mean   : 7.621   Mean   :63.81   Mean   :17.99   Mean   : 79.81   Mean   :13490
3rd Qu.:215.1   3rd Qu.:11.450   3rd Qu.:72.90   3rd Qu.:24.32   3rd Qu.: 87.80   3rd Qu.:15979
Max.   :268.7   Max.   :14.400   Max.   :77.20   Max.   :26.30   Max.   :106.90   Max.   :19748
```

```
> round(cor(data[2:7]), digits=3)
              Invecchiamento PopStraniera Occupazione Pendolarismo Densità Reddito
Invecchiamento          1.000
PopStraniera            0.351        1.000
Occupazione             0.642        0.899       1.000
Pendolarismo            0.230        0.530       0.656        1.000
Densità                 0.513        0.876       0.930        0.674   1.000
Reddito                 0.574        0.905       0.961        0.597   0.946   1.000
>
```

L'analisi della matrice di correlazione è un primo, importante passo per comprendere le relazioni tra le variabili



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

VARIABILITA' SPIEGATA

```
> summary(res.pca)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6
Standard deviation     2.1319 0.8963 0.70986 0.30071 0.20173 0.12984
Proportion of Variance 0.7575 0.1339 0.08398 0.01507 0.00678 0.00281
Cumulative Proportion  0.7575 0.8913 0.97534 0.99041 0.99719 1.00000
>
```

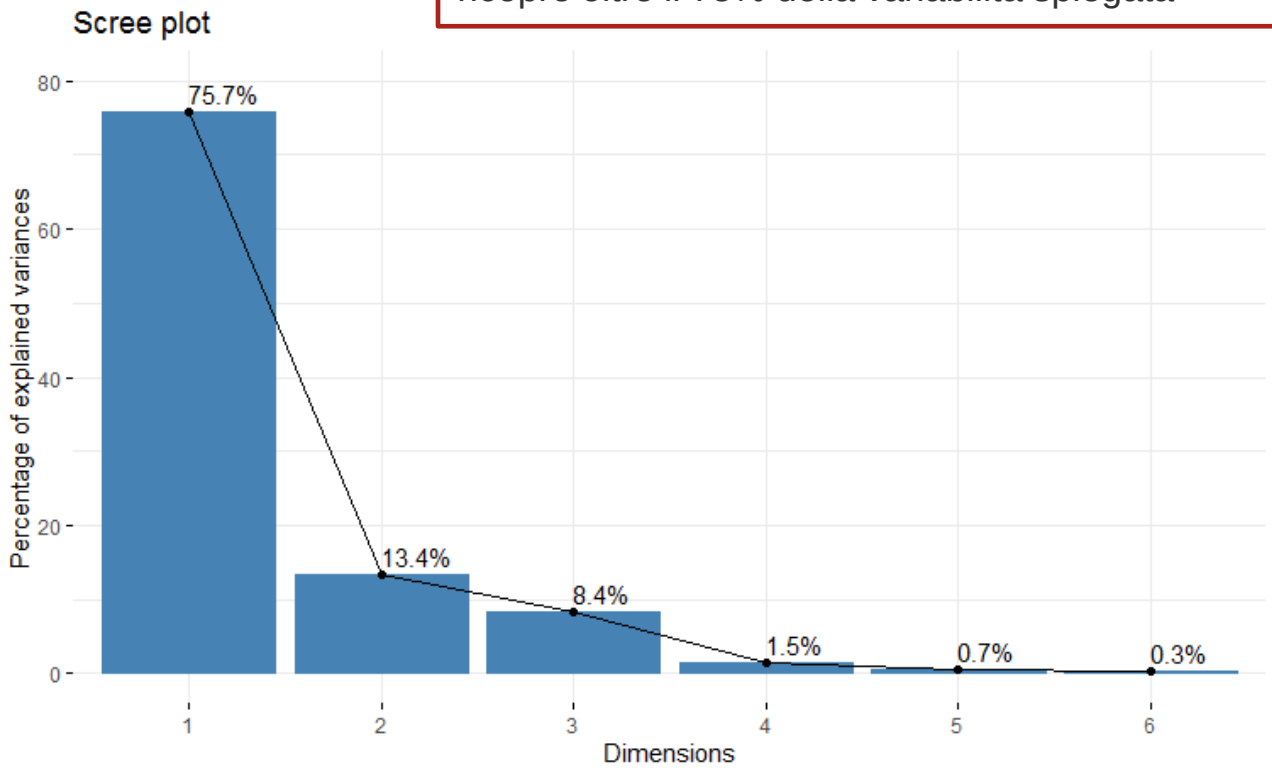
Dall'output possiamo vedere che le prime tre PC spiegano più del 97% della variabilità globale

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

SCREE PLOT E LA SCELTA DELLE COMPONENTI

```
> get_eig(res.pca)
       eigenvalue variance.percent cumulative.variance.percent
Dim.1 4.54483028       75.7471714                    75.74717
Dim.2 0.80329057       13.3881762                    89.13535
Dim.3 0.50389809        8.3983015                    97.53365
Dim.4 0.09042584        1.5070973                    99.04075
Dim.5 0.04069555        0.6782591                    99.71901
Dim.6 0.01685967        0.2809945                   100.00000
>
```

Sia lo scree plot che l'analisi degli autovalori che il criterio della varianza spiegata suggerirebbero di non usare più di una componente, dato che da sola ricopre oltre il 75% della variabilità spiegata



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

ANALISI PUNTI-VARIABILI

```
# Results for Variables
res.var <- get_pca_var(res.pca)

res.var$coord      # Coordinates
res.var$cor        #Correlations between variables and dimensions
res.var$contrib    # Contributions to the PCs
res.var$cos2       # Quality of representation
```

```
> res.var$coord
                Dim.1       Dim.2
Invecchiamento 0.6115076 -0.74686047
PopStraniera   0.9038704  0.16206291
Occupazione    0.9880316 -0.06056941
Pendolarismo   0.7052006  0.45700043
Densità        0.9652625  0.07717566
Reddito        0.9739922 -0.02741325
```

```
> res.var$cor
                Dim.1       Dim.2
Invecchiamento 0.6115076 -0.74686047
PopStraniera   0.9038704  0.16206291
Occupazione    0.9880316 -0.06056941
Pendolarismo   0.7052006  0.45700043
Densità        0.9652625  0.07717566
Reddito        0.9739922 -0.02741325
```

```
> res.var$contrib
                Dim.1       Dim.2
Invecchiamento  8.227844 69.43945149
PopStraniera   17.976067  3.26959970
Occupazione    21.479492  0.45670311
Pendolarismo   10.942278 25.99923428
Densità        20.500913  0.74146046
Reddito        20.873406  0.09355096
```

```
> res.var$cos2
                Dim.1       Dim.2
Invecchiamento 0.3739416 0.557800566
PopStraniera   0.8169817 0.026264386
Occupazione    0.9762064 0.003668653
Pendolarismo   0.4973080 0.208849397
Densità        0.9317317 0.005956082
Reddito        0.9486609 0.000751486
```

ANALISI PUNTI-VARIABILI

```
# Results for Variables
res.var <- get_pca_var(res.pca)

res.var$contrib        # Contributions to the PCs
```

```
                Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6
Invecchiamento   8.23 69.44 12.89  1.46  2.58  5.40
PopStraniera    17.98  3.27 24.72 27.41 12.25 14.38
Occupazione     21.48  0.46  0.15  8.12  1.20 68.60
Pendolarismo    10.94 26.00 57.76  2.60  0.00  2.70
Densità         20.50  0.74  0.65 55.02 22.62  0.46
Reddito         20.87  0.09  3.83  5.40 61.35  8.45
>
```

I **contributi** rappresentano la percentuale di varianza delle variabili rappresentata da ciascuna componente. Per tale ragione, possono essere interpretati in termini di contributo dato da ciascuna variabile alla definizione delle componenti

Una variabile con un contributo alto è importante ai fini dell'interpretazione semantica della componente

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

ANALISI PUNTI-VARIABILI

```
# Results for Variables
res.var <- get_pca_var(res.pca)

res.var$cos2        # Quality of representation
```

I **coseni quadri** (o contributi relativi) rappresentano la quota di variabilità della variabile "spiegata" dalla componente

Una variabile con un coseno quadrato alto è ben rappresentata dalla componente

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Invecchiamento	0.37	0.56	0.06	0.00	0.00	0.00
PopStraniera	0.82	0.03	0.12	0.02	0.00	0.00
Occupazione	0.98	0.00	0.00	0.01	0.00	0.01
Pendolarismo	0.50	0.21	0.29	0.00	0.00	0.00
Densità	0.93	0.01	0.00	0.05	0.01	0.00
Reddito	0.95	0.00	0.02	0.00	0.02	0.00

>

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

ANALISI PUNTI-VARIABILI

Quando le variabili sono numerose, può essere utile tenere a mente la distinzione fra contributi e coseni quadrati.

```
> get_eig(res.pca)
       eigenvalue variance.percent cumulative.variance.percent
Dim.1 4.54483028        75.7471714                    75.74717
Dim.2 0.80329057        13.3881762                    89.13535
Dim.3 0.50389809         8.3983015                    97.53365
Dim.4 0.09042584         1.5070973                    99.04075
Dim.5 0.04069555         0.6782591                    99.71901
Dim.6 0.01685967         0.2809945                   100.00000
>
```

Ad esempio, la varianza della variabile Occupazione sulla prima componente è pari al 21,48% del 4,54% di varianza spiegata dalla componente stessa

Anche se questa variabile contribuisce in misura maggiore al sesto asse (68,60%), è meglio rappresentata dal primo. Il sesto asse, infatti, riproduce solo lo 0,02% della varianza totale

res.var$contrib # Contributions to the PCs

```
                Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6
Invecchiamento   8.23 69.44 12.89  1.46  2.58  5.40
PopStraniera    17.98  3.27 24.72 27.41 12.25 14.38
Occupazione     21.48  0.46  0.15  8.12  1.20 68.60
Pendolarismo    10.94 26.00 57.76  2.60  0.00  2.70
Densità         20.50  0.74  0.65 55.02 22.62  0.46
Reddito         20.87  0.09  3.83  5.40 61.35  8.45
>
```

res.var$cos2 # Quality of representation

```
                Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6
Invecchiamento   0.37  0.56  0.06  0.00  0.00  0.00
PopStraniera     0.82  0.03  0.12  0.02  0.00  0.00
Occupazione      0.98  0.00  0.00  0.01  0.00  0.01
Pendolarismo     0.50  0.21  0.29  0.00  0.00  0.00
Densità          0.93  0.01  0.00  0.05  0.01  0.00
Reddito          0.95  0.00  0.02  0.00  0.02  0.00
>
```

```
21,48 * 4,54 =  0,98
```

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

ANALISI PUNTI-INDIVIDUI

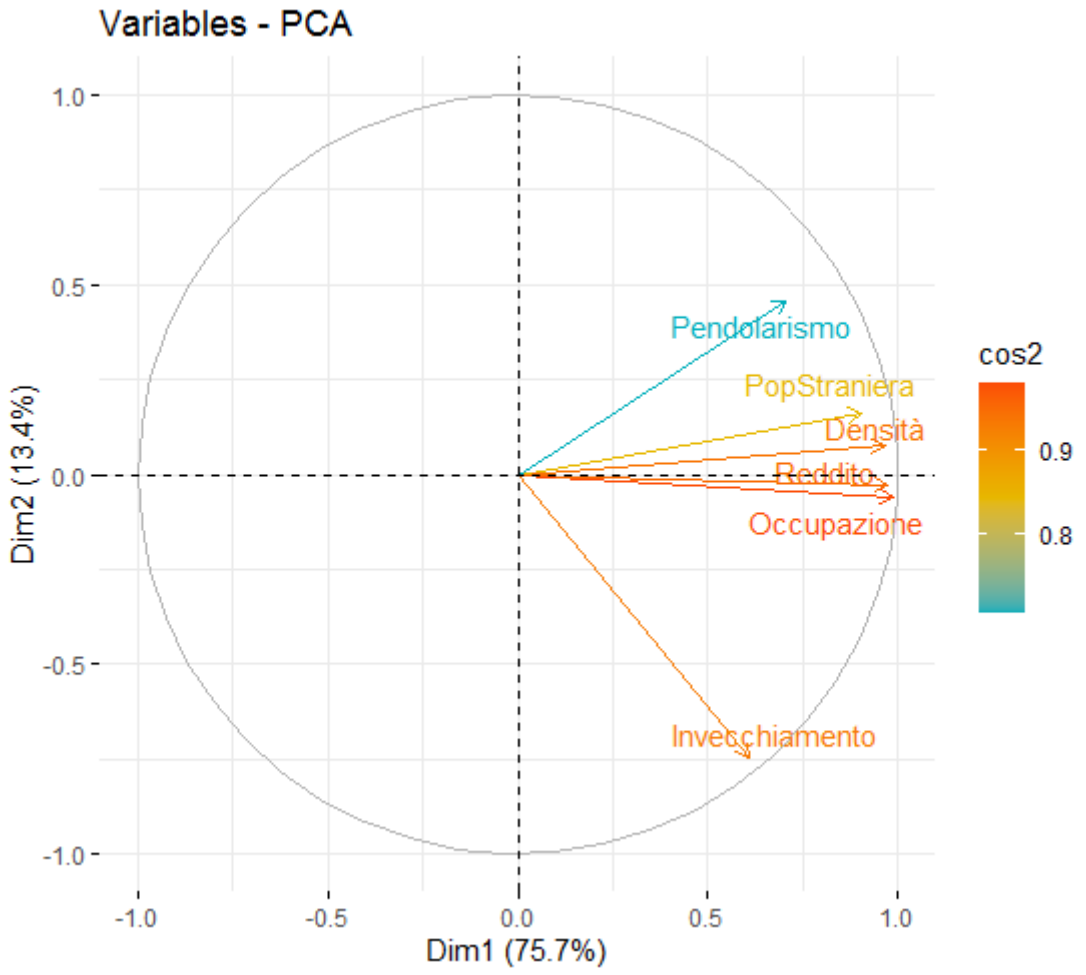
```
# Results for individuals
res.ind <- get_pca_ind(res.pca)

res.ind$coord      # Coordinates
res.ind$contrib    # Contributions to the PCs
res.ind$cos2       # Cos2 for the individuals (Quality of representation)
```

```
> round(res.ind$contrib, digits =3)
                Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6
Bari            2.719  0.202  0.560 11.489  0.000 48.872
Bologna         8.994  2.001  0.004  1.593 14.242  4.741
Cagliari        0.036  2.492 31.059 14.654  0.078  3.714
Catania         9.217  5.376  2.350  0.026  0.002  0.640
Firenze         8.940  0.145  0.000  0.330 43.856  1.859
Genova          3.224 47.936  3.661  0.802  1.601  0.002
Messina         4.336  6.370  0.012  0.630  7.418 16.074
Milano         14.858 15.604  3.523 11.121  2.150 14.557
Napoli          9.190 12.175  0.002  0.817  2.170  0.334
Palermo        14.660  0.201  2.747  0.231 13.284  0.628
Reggio Calabria 8.342  0.134  2.179 20.038  0.328  0.786
Roma            0.550  0.157 35.137  1.067  0.456  0.435
Torino          4.184  0.049  7.307  8.306  6.364  0.002
Venezia         3.607  0.014  4.316 21.752  0.908  0.214
>
```

```
> round(res.ind$cos2, digits =3)
                Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6
Bari            0.843 0.011 0.019 0.071 0.000 0.056
Bologna         0.944 0.037 0.000 0.003 0.013 0.002
Cagliari        0.009 0.104 0.815 0.069 0.000 0.003
Catania         0.884 0.091 0.025 0.000 0.000 0.000
Firenze         0.954 0.003 0.000 0.001 0.042 0.001
Genova          0.266 0.698 0.033 0.001 0.001 0.000
Messina         0.774 0.201 0.000 0.002 0.012 0.011
Milano          0.812 0.151 0.021 0.012 0.001 0.003
Napoli          0.808 0.189 0.000 0.001 0.002 0.000
Palermo         0.969 0.002 0.020 0.000 0.008 0.000
Reggio Calabria 0.926 0.003 0.027 0.044 0.000 0.000
Roma            0.122 0.006 0.866 0.005 0.001 0.000
Torino          0.801 0.002 0.155 0.032 0.011 0.000
Venezia         0.796 0.001 0.106 0.096 0.002 0.000
>
```

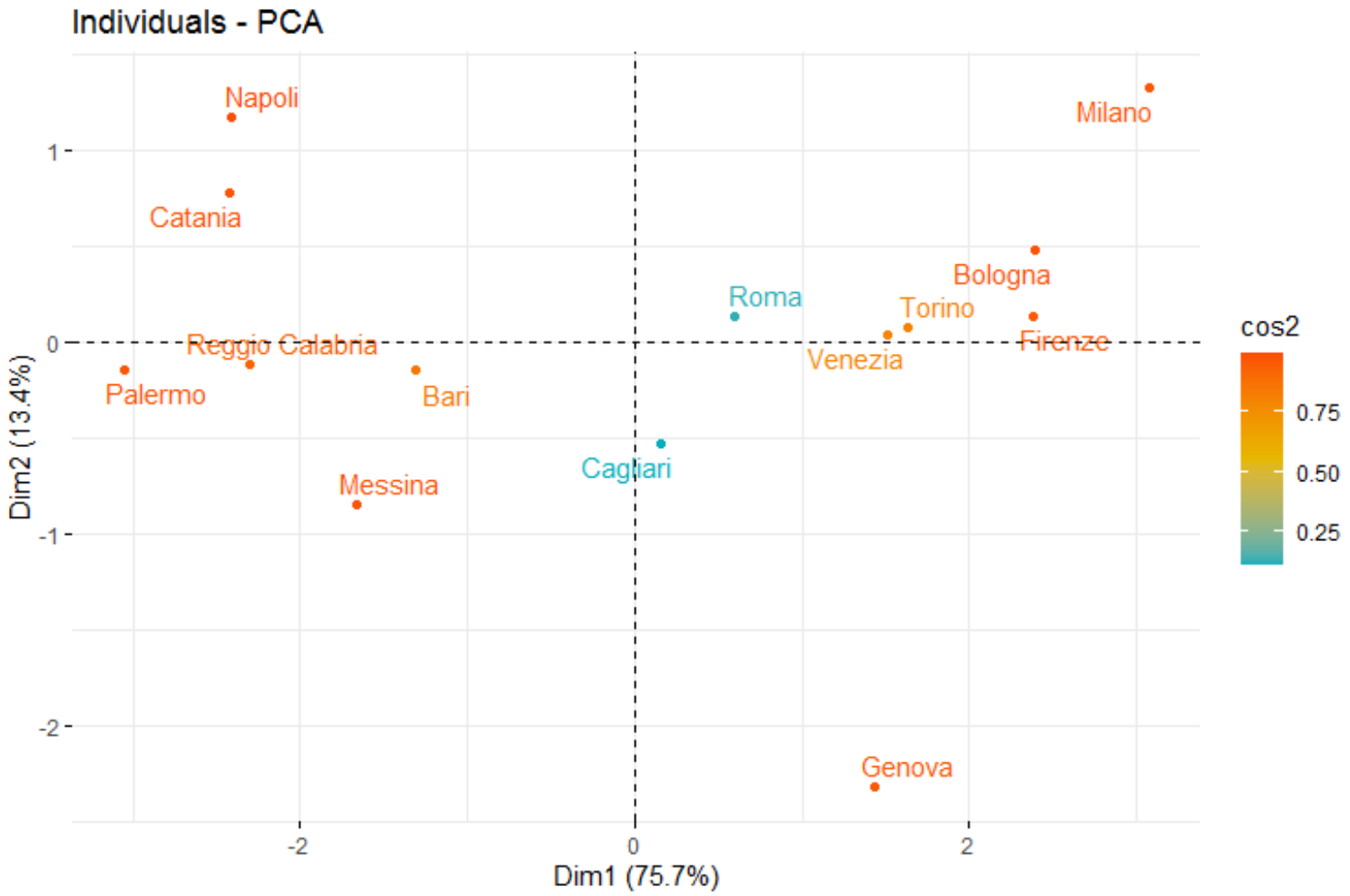
L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

IL CERCHIO DELLE CORRELAZIONI



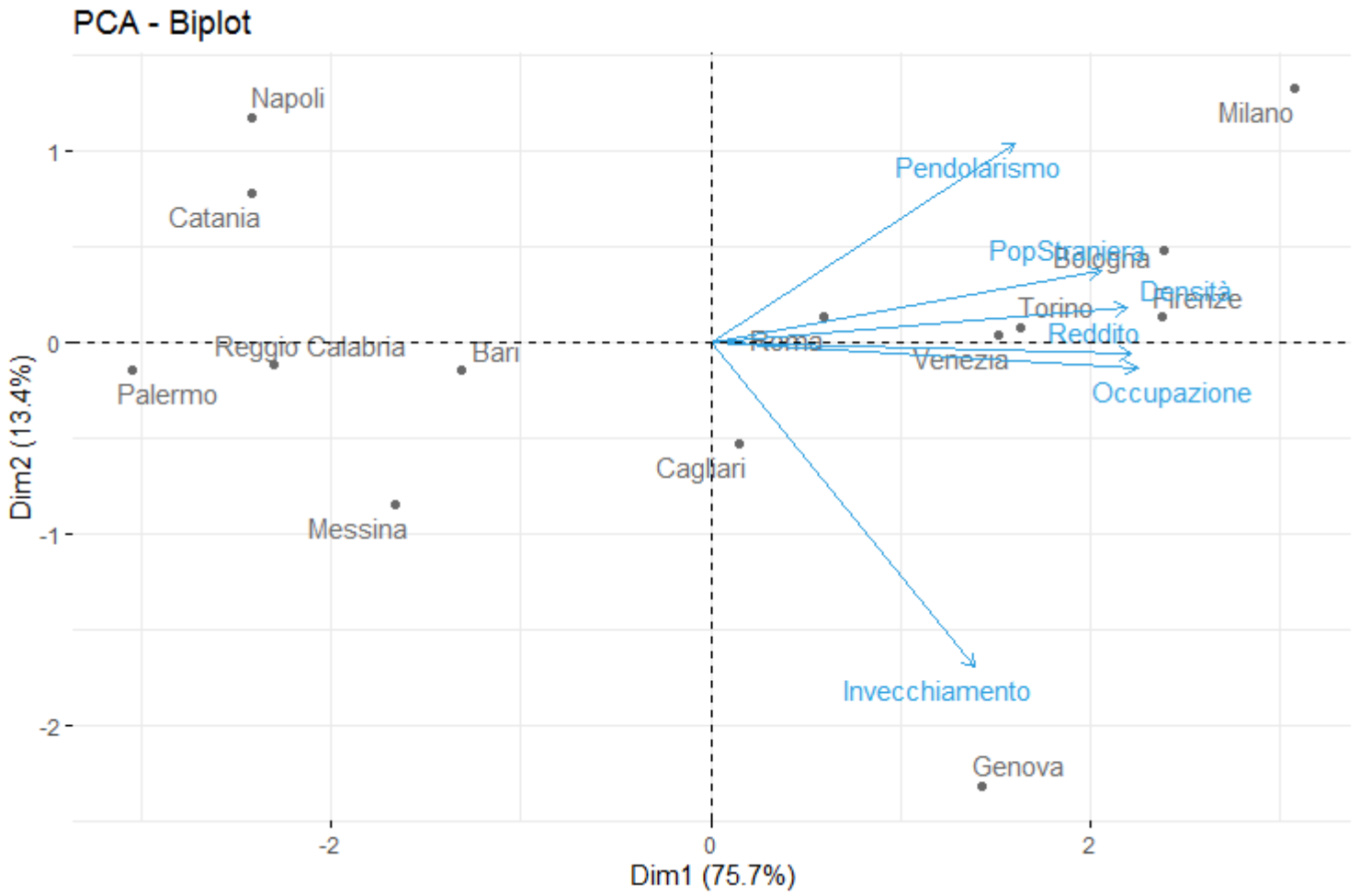
L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

IL GRAFICO DELLE OSSERVAZIONI



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

PROIEZIONE DEI PUNTI NELLO SPAZIO R^P



L'origine degli assi rappresenta la media delle variabili, o il valore rispetto al quale le variabili sono state normalizzate.

I punti individuo sono sempre centrati sull'origine che, per la centratura dei dati, coincide con il baricentro, e rappresenta **l'individuo medio**

Più un punto è distante dal baricentro più sarà caratterizzato da alcuni aspetti specifici


ORA TOCCA
A VOI!

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

I DATI: VIOLENT CRIME RATES BY US STATE

 # Data Import

Data(USArrests)

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Format

[,1]	Murder	numeric	Murder arrests (per 100,000)
[,2]	Assault	numeric	Assault arrests (per 100,000)
[,3]	UrbanPop	numeric	Percent urban population
[,4]	Rape	numeric	Rape arrests (per 100,000)

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

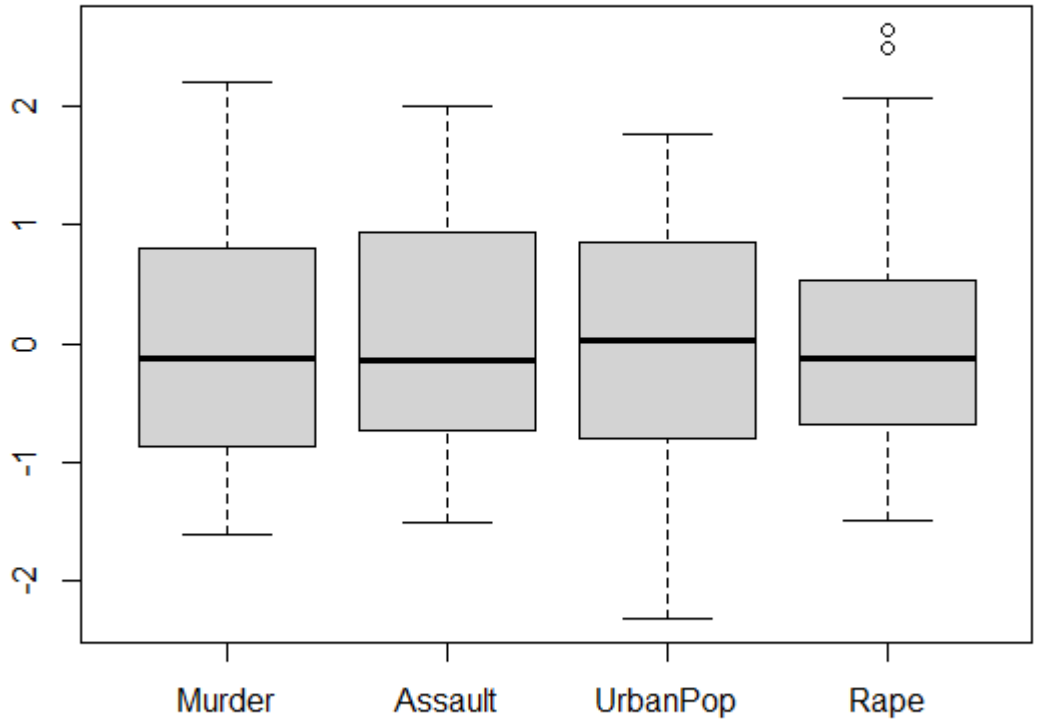
STATISTICHE DESCRITTIVE

 # Descriptive Statistics

```
summary(USArrests)
boxplot(scale(data))
cor(data)
```

```
> summary(USArrests)
     Murder          Assault         UrbanPop          Rape
 Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
 Median : 7.250   Median :159.0   Median :66.00   Median :20.10
 Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
 Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```
> round(cor(USArrests), digits=3)
         Murder Assault UrbanPop  Rape
Murder    1.000
Assault   0.802   1.000
UrbanPop  0.070   0.259    1.000
Rape      0.564   0.665    0.411 1.000
```



L'analisi della matrice di correlazione è un primo, importante passo per comprendere le relazioni tra le variabili

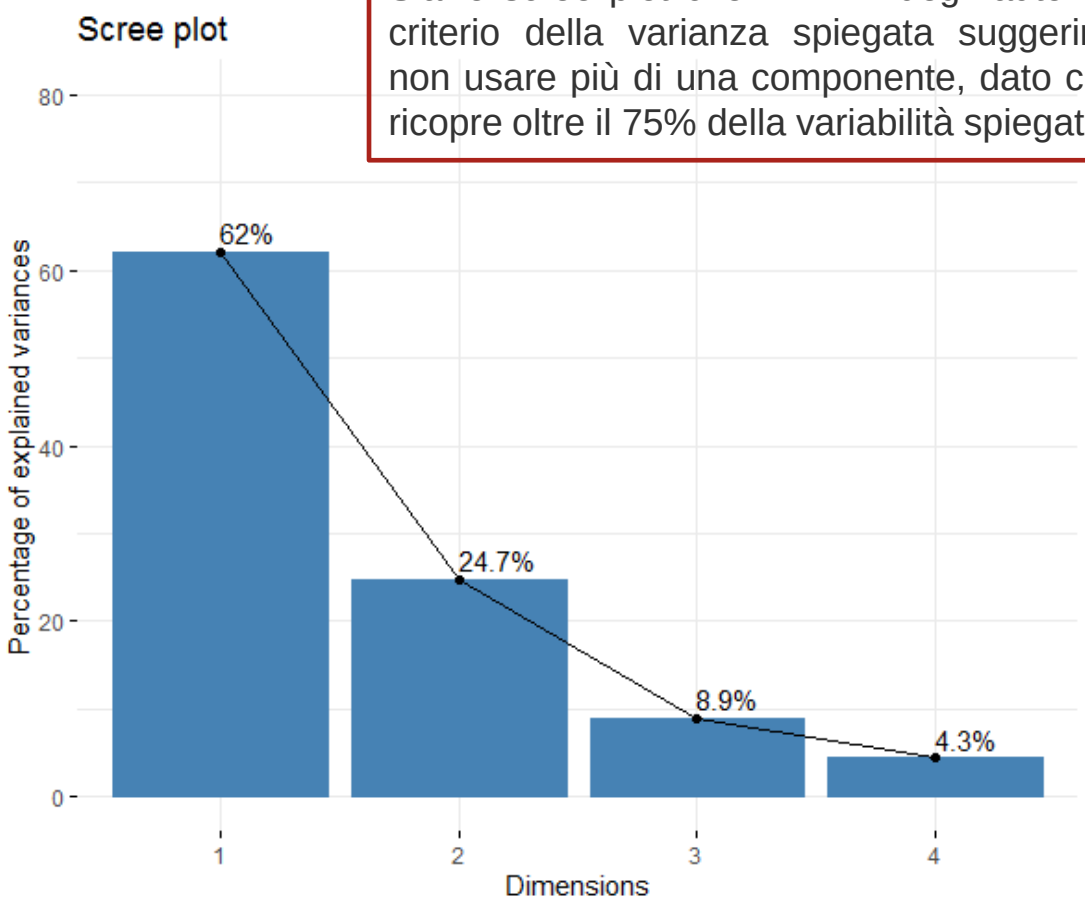
L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

SCREE PLOT E LA SCELTA DELLE COMPONENTI

 # compute PCA

res.pca <- prcomp(USArrests, scale = TRUE)

```
>
> get_eig(res.pca) # tabella degli autovalori
       eigenvalue variance.percent cumulative.variance.percent
Dim.1  2.4802416        62.006039                    62.00604
Dim.2  0.9897652        24.744129                    86.75017
Dim.3  0.3565632         8.914080                    95.66425
Dim.4  0.1734301         4.335752                   100.00000
>
>
```



Sia lo scree plot che l'analisi degli autovalori che il criterio della varianza spiegata suggerirebbero di non usare più di una componente, dato che da sola ricopre oltre il 75% della variabilità spiegata

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

ANALISI PUNTI-VARIABILI

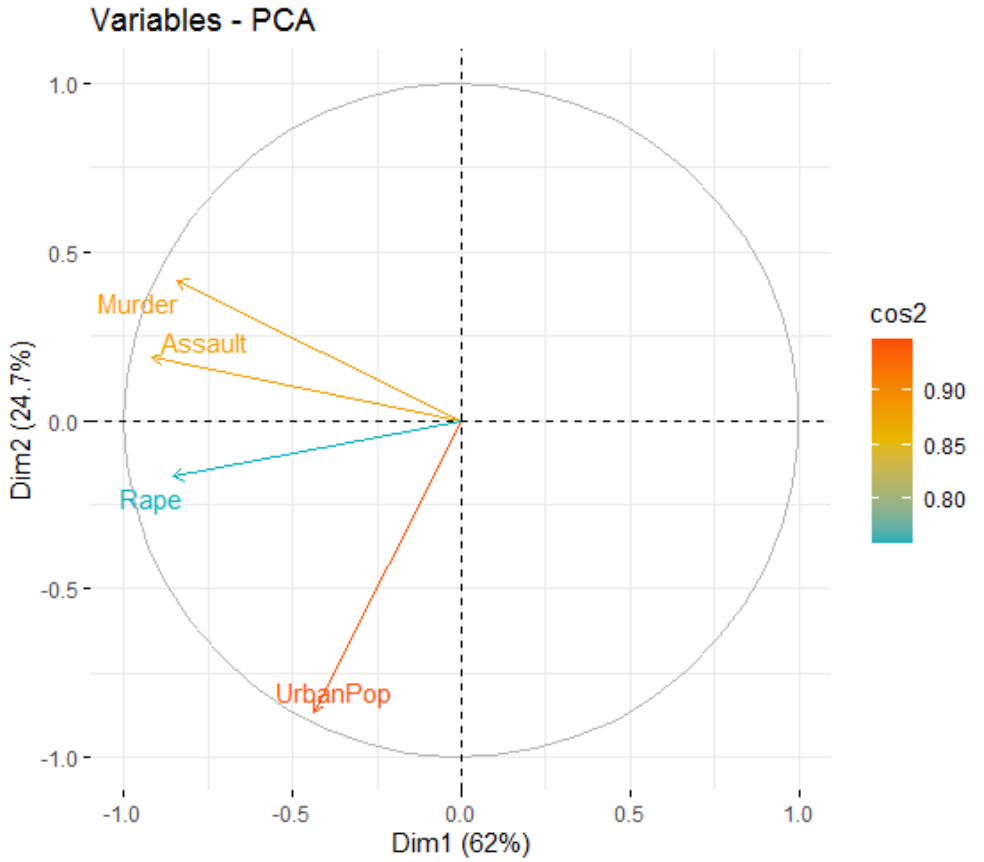
```
# Results for Variables
res.var <- get_pca_var(res.pca)

res.var$coord          # Coordinates
res.var$cor            #Correlations between variables and dimensions
res.var$contrib        # Contributions to the PCs
res.var$cos2           # Quality of representation
```

```
> res.var$contrib        # Contributions to the PCs
             Dim.1     Dim.2     Dim.3     Dim.4
Murder   28.718825 17.487524 11.643977 42.149674
Assault  34.010315  3.533859  7.190358 55.265468
UrbanPop  7.739016 76.179065 14.289594  1.792325
Rape     29.531844  2.799553 66.876071  0.792533
> res.var$cos2           # Cos2 for the variables (Quality of representation)
             Dim.1     Dim.2      Dim.3       Dim.4
Murder   0.7122962 0.1730854 0.04151814 0.073100217
Assault  0.8435380 0.0349769 0.02563817 0.095846950
UrbanPop 0.1919463 0.7539938 0.05095143 0.003108430
Rape     0.7324611 0.0277090 0.23845544 0.001374491
```

L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

IL CERCHIO DELLE CORRELAZIONI

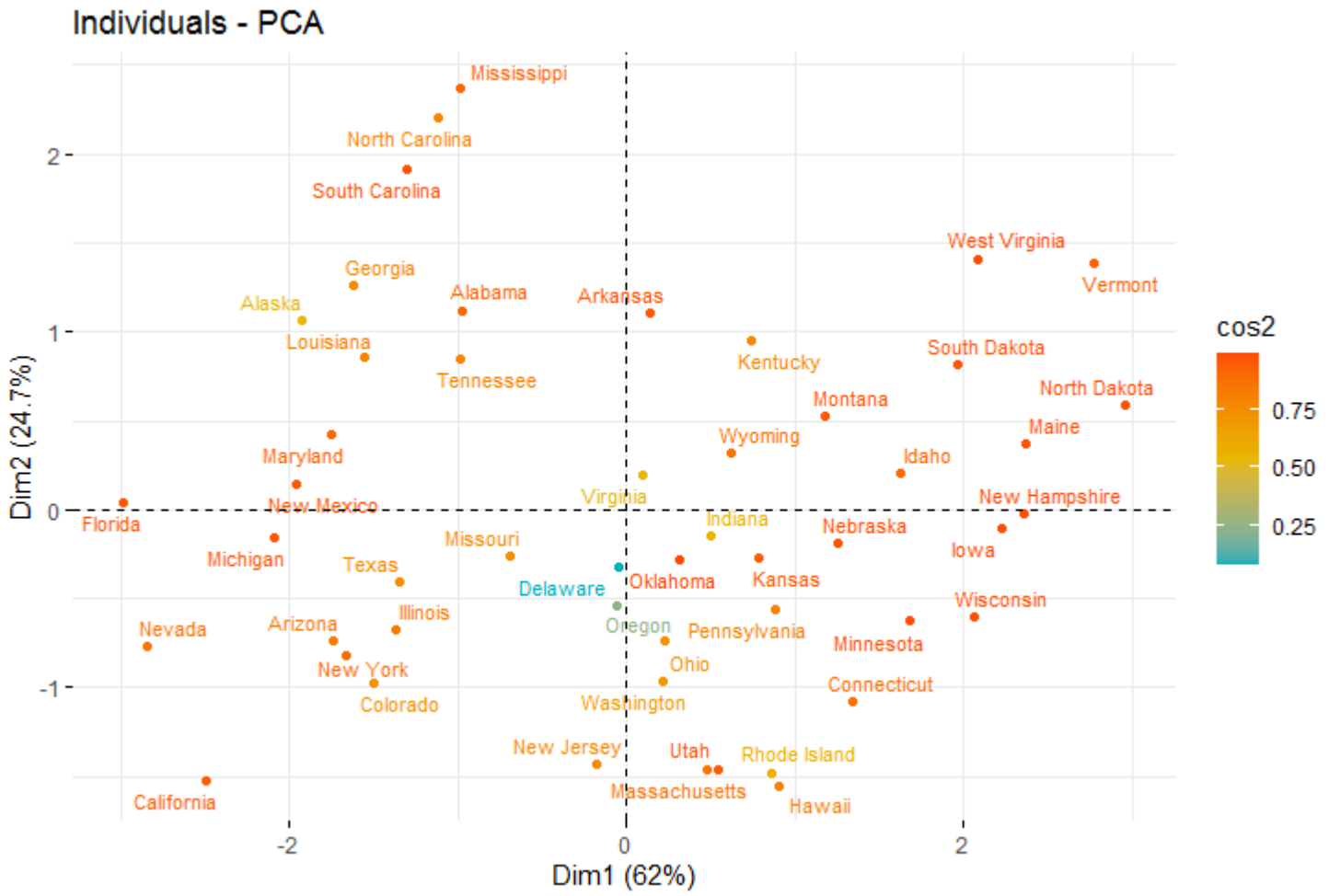
```
fviz_pca_var(res.pca,
             col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

IL GRAFICO DELLE OSSERVAZIONI

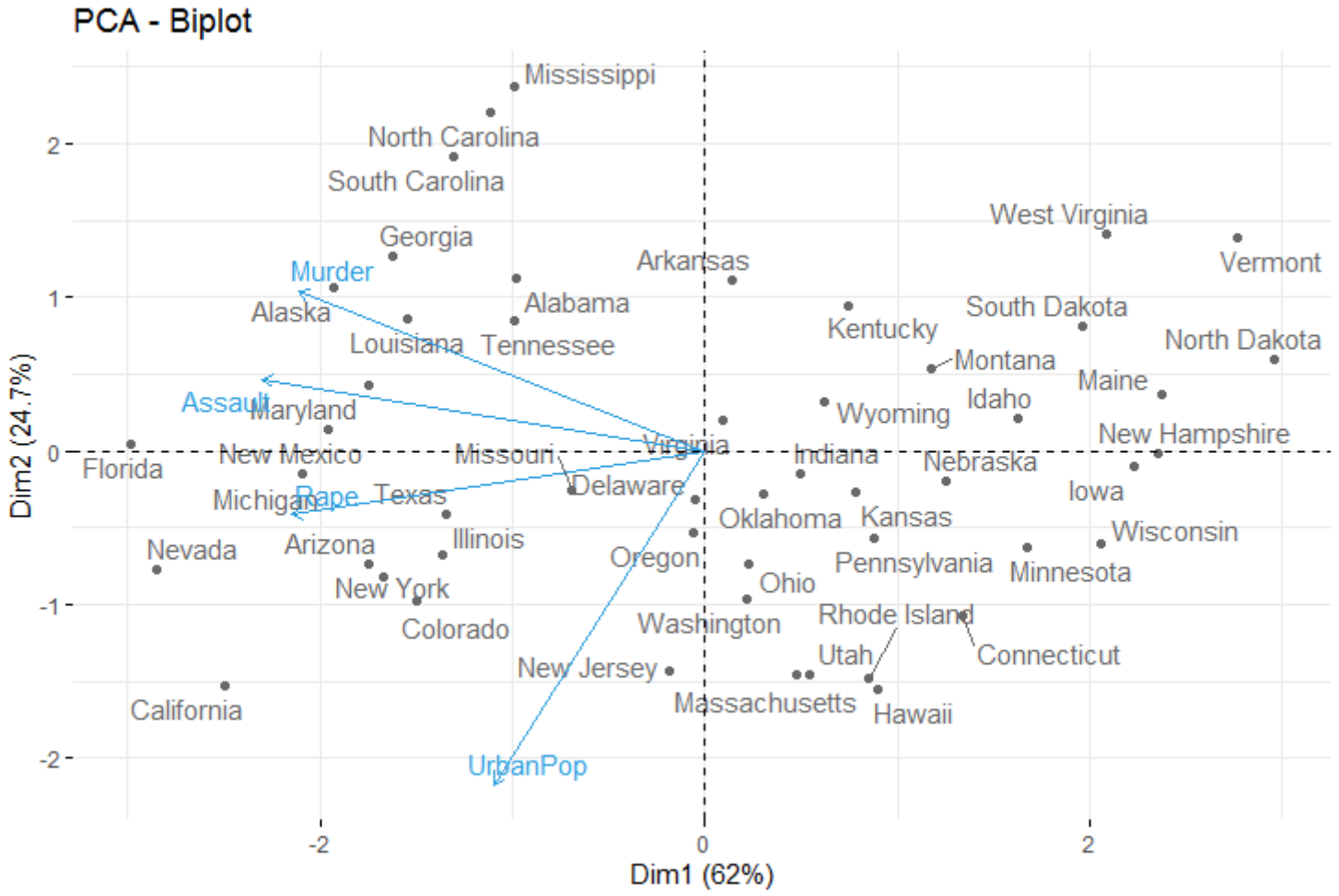
```
# Il grafico delle osservazioni

fviz_pca_ind(res.pca,
             col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```



L' ANALISI DELLE COMPONENTI PRINCIPALI (ACP)

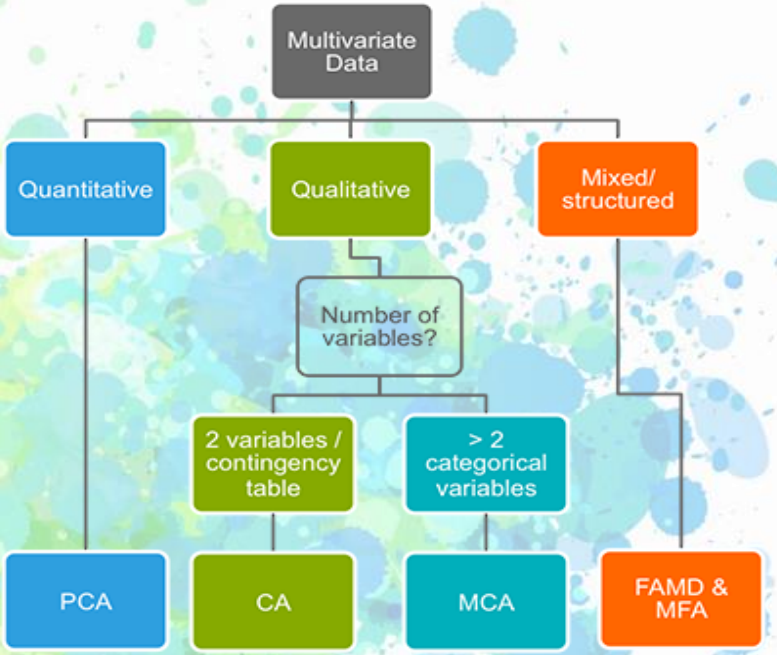
IL GRAFICO DELLE OSSERVAZIONI E DELLE VARIABILI

```
# Il biplot

fviz_pca_biplot(res.pca,,
            col.ind =  "#696969",
            col.ind =  "#2E9FDF",
            repel = TRUE)
```



DIMENSIONALITY REDUCTION

Methods to Summarize & Visualize Multivariate Data



- PCA: Principal Component Analysis
- (M) CA: (Multiple) Correspondence Analysis
- FAMD: Factor Analysis of Mixed Data
- MFA: Multiple Factor Analysis