

Bike sharing a Parigi: un case study

Corso di Statistica Iterazione 18/19



Alessandro Casa



casa@stat.unipd.it



[alessandrocasa.github.io](https://github.com/alessandrocasa)

2 Maggio 2019

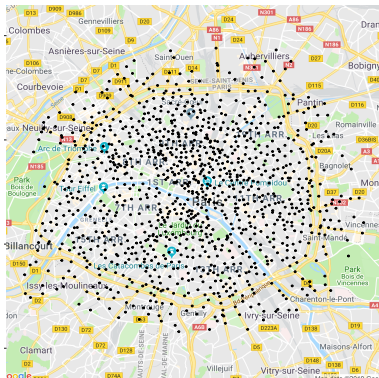


STATISTICA

Searching for a trail of evidence in a maze

- Dataset reali, discuterne, proporre analisi e trovare qualche soluzione (cerchiamo di convincerci che aver dovuto passare Statistica CP e Data Mining sia servito a qualcosa..)
- **Entry requirements:** partecipazione
(che sennò mi imbarazzo, è la mia prima lezione 🎂)

Biciclette e baguette



 **Velib Data**

bici disponibili

totale di postazioni

per 1189 docking stations a Parigi.
Misurazioni orarie per un'intera settimana (e chiamiamolo toy example...)

 **Let's play a game:**

il sindaco di Parigi vi chiede una consulenza. A che cosa potrebbe essere interessato?

Diamo un'occhiata

0.54	0.09	0.71	0.65						
0.71	1	0.09	0.11						
1	0.1	0.71	0.33	0.11	0.09				
0.92	0.65	0.12	0.29	0	0.54				
		0.09	0	0.11	0.09	0.92	0.1		
		0.09	0.65	1	0.29	0.29	0.29		
				0.87	0.11	0.87	0.02		
				0.87	0.12	0.78	1		

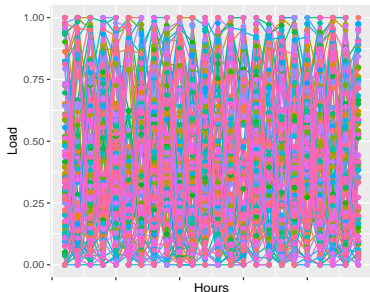
🚲 Struttura

un array con # righe = # stazioni,
colonne che indicano giorni della
settimana e strati che indicano le
ore del giorno

🚲 Matrice dove ogni cella è una
serie temporale con $T = 24$.

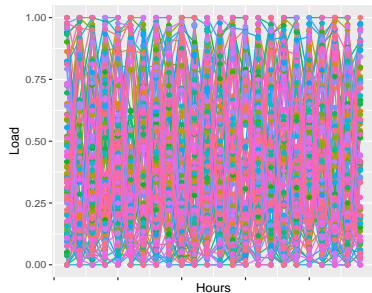
🚲 Prima cosa da fare?

Diamo un'occhiata - 2

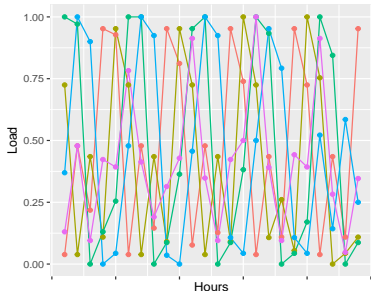


🚲 A volte dare un'occhiata ai dati
può non esser così semplice...

Diamo un'occhiata - 2



🚲 A volte dare un'occhiata ai dati può non esser così semplice...



🚲 5 docking stations, un solo giorno della settimana

- 🚲 Natura dei dati? Come e per quale obiettivo possiamo usarli?

🚲 Natura dei dati? Come e per quale obiettivo possiamo usarli?

🚲 **Focus:** ci concentriamo sulle osservazioni per il singolo giorno della settimana.

Unità statistica è una curva (sequenza di osservazioni) e non una singola osservazione.


🚲 **Possibile scopo:** cercare dei pattern di utilizzo tra le diverse docking station che tengano conto della natura dell'unità statistica e della struttura temporale

Big brother (Jeff Bezos) is watching you



Amazon Fine Food Reviews¹

Più di 10 anni di recensioni Amazon nella categoria *fine food*

 Numero di recensioni = 568464
Numero di prodotti recensiti = 74258
Numero di utenti diversi = 256059

Informazioni a disposizione

id prodotto, id utente, valutazione prodotto (rating da 1 a 5), summary recensione, recensione completa (testi)

¹Free download: <https://snap.stanford.edu/data/web-FineFoods.html>
<https://www.kaggle.com/snap/amazon-fine-food-reviews>

a Esempio di riga del dataset

PROD ID: B0009XLVGo

USER ID: A2725IB4YY9JEB

SCORE: ★★★★★

SUMMARY: *My cats LOVE this "diet" food better than their regular food*

TEXT: *One of my boys needed to lose some weight and the other didn't.*

*I put this food on the floor for the chubby guy, and the protein rich,
no by-product food up higher where only my skinny boy can jump.*

The higher food sits going stale. They both really go for this food.


And my chubby boy has been losing about an ounce a week.


a Tipologia di dati?


Possibili analisi?

Idee su informazioni che Amazon potrebbe estrarre?

Maggiori difficoltà nell'analizzarli?

 **Possibile scopo:** fornire un possibile sistema di raccomandazione personalizzato per ogni utente

 **Focus:** tralasciando informazioni testuali ci concentriamo sugli score forniti dagli utenti (da 1-pessimo a 5-ottimo)

 Qual è il maggior problema da affrontare in dati di questo tipo?

- Possibile analisi valida in entrambi gli esempi presentati

CLUSTERING



- Flessibile per diverse tipologie di dati
 - Utile analisi esplorativa per aver qualche informazione iniziale
 - Riassume anche grandi matrici di dati in un numero limitato di gruppi
- Output:
 - gruppi di docking stations con utilizzo simile nell'arco della settimana
 - gruppi di utenti Amazon con preferenze d'acquisto/gusti simili

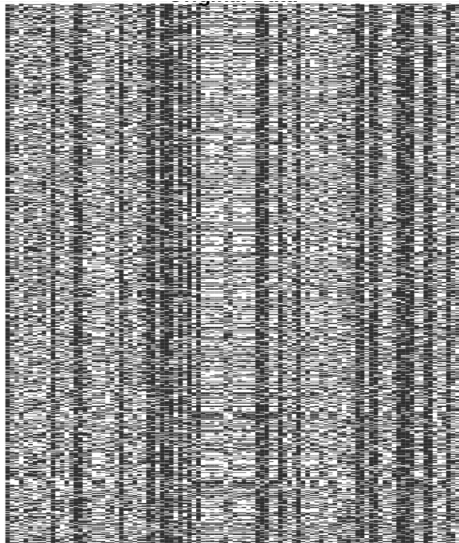
- Sia \mathbf{X} una matrice $n \times p$ di dati. Si assume che i dati provengano da un modello di mistura quindi la densità di una singola osservazione $x_i = \{x_1, \dots, x_p\}$ è data da

$$f(x_i|\Theta) = \sum_{k=1}^K \pi_k f_k(x_i|\theta_k)$$

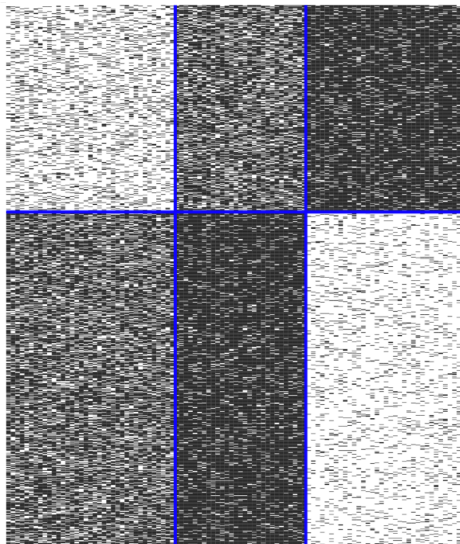
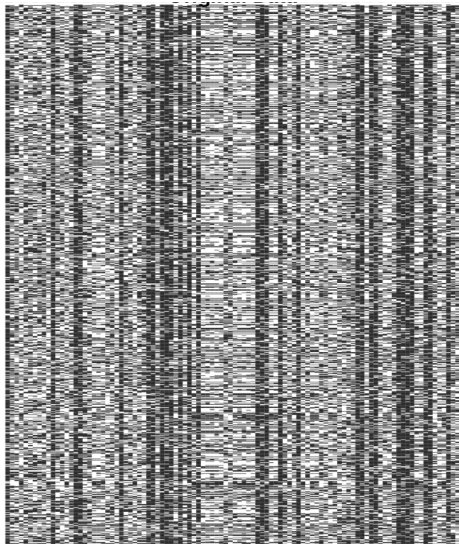
- $f_k(\cdot|\theta_k)$ può essere scelta per accomodare diversi tipologie di dati (ad es. funzioni o categoriali)
- Il modello viene stimato utilizzando l'algoritmo EM (v. lezione Menardi). Le osservazioni vengono allocate ad uno dei K gruppi valutando la probabilità a posteriori di appartenenza

- La riduzione della dimensionalità del problema operata dal clustering può non essere informativa in caso di matrici di grandi dimensioni
- Cibo dietetico per gatti \neq Nocciolata Rigoni.
- Nel caso in cui si voglia
 1. Maggiore riduzione della dimensionalità e della complessità del problema
 2. Gruppi coesi non solo per caratteristiche delle unità statistiche ma anche delle variabili rilevate
- **CO-CLUSTERING:** lo scopo è quello di ottenere gruppi sia di righe che di colonne. Studio congiuntamente similarità delle unità e delle variabili.

Per convincervi



Per convincervi



Un po' di notazione

- Diversi approcci al co-clustering, la maggior parte euristici/basati su distanza (come per il clustering)
- **Focus:** approccio basato su modello statistico
- But first:
 - $\mathbf{X} = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ matrice $n \times p$ di dati osservati
 - $K = \text{n.ro di cluster-riga}$
 $L = \text{n.ro di cluster-colonna}$
 - Struttura latente
 - $\mathbf{z} = (z_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ ➤ appartenenza cluster-riga
 - $\mathbf{w} = (w_{jl})_{1 \leq j \leq p, 1 \leq l \leq L}$ ➤ appartenenza cluster-colonna

$$z_{ik} = \begin{cases} 1 & \text{se } x_{i.} \in \text{cl}_k \\ 0 & \text{altrimenti} \end{cases} \qquad w_{jl} = \begin{cases} 1 & \text{se } x_{.j} \in \text{cl}_l \\ 0 & \text{altrimenti} \end{cases}$$

Latent Block Model

- Modello di co-clustering più utilizzando è il **Latent Block Model**
- **Assunzioni:**
 - \mathbf{z} e \mathbf{w} sono variabili casuali multinomiali indipendenti
 - Condizionatamente a \mathbf{z} e \mathbf{w} , le variabili x_{ij} sono indipendenti
- Il modello è definito come:

$$p(\mathbf{X}; \theta) = \sum_{\mathbf{z} \in \mathcal{Z}} \sum_{\mathbf{w} \in \mathcal{W}} p(\mathbf{z}; \theta) p(\mathbf{w}; \theta) p(\mathbf{X} | \mathbf{z}, \mathbf{w}; \theta)$$

- \mathcal{Z} e \mathcal{W} : possibili partizioni delle righe (colonne) in K (L) gruppi
- $p(\mathbf{z}; \theta) = \prod_{ik} \pi_k^{z_{ik}}$ e $p(\mathbf{w}; \theta) = \prod_{jl} \rho_l^{w_{jl}}$
- $p(\mathbf{X} | \mathbf{z}, \mathbf{w}; \theta) = \prod_{ijkl} p(x_{ij}; \theta_{kl})^{z_{ik} w_{jl}}$
dove $p(\cdot; \theta_{kl})$ va scelto tra i modelli probabilistici adeguati a descrivere il tipo di dato osservato

Misture sotto steroidi

- LBM è una modello di mistura con uno "strato aggiuntivo" (con qualche assunzione e difficoltà -v. stima- in più)
- **Modello di mistura**

$$\begin{aligned}p(\mathbf{X}; \theta) &= \sum_{z \in Z} p(z; \theta) p(\mathbf{X}|z; \theta) \\&= \sum_{z \in Z} \prod_{i=1}^n \prod_{k=1}^K [\pi_k f(x_i; \theta_k)]^{z_{ik}}\end{aligned}$$

- **Latent Block Model**

$$\begin{aligned}p(\mathbf{X}; \theta) &= \sum_{z \in Z} \sum_{w \in W} p(z; \theta) p(w; \theta) p(\mathbf{X}|z, w; \theta) \\&= \sum_{z \in Z} \sum_{w \in W} \prod_{i=1}^n \prod_{j=1}^p \prod_{k=1}^K \prod_{l=1}^L [\pi_k \rho_l f(x_{ij}; \theta_{kl})]^{z_{ik} w_{jl}}\end{aligned}$$

- **Stima:** massimizzazione diretta della verosimiglianza non è possibile e si deve ricorrere all'EM sfruttando la struttura latente del modello
- La doppia struttura latente complica l'E-step che in questo caso non ammette semplici aggiornamenti in forma chiusa.
- Problema intrattabile quindi si usano soluzioni approssimate:
 - Variational EM
 - Classification EM
 - Stochastic EM-Gibbs

Selezione del modello

- La scelta del numero di blocchi viene posta in termini di selezione del modello (come nel clustering)
- Usualmente si utilizzano criteri di informazione quali ICL e BIC con una preferenza per il primo nell'ambito di co-clustering (o approcci ibridi)
- Insiemi di dati molto grandi possono richiedere la stima di molti modelli
 - **Greedy Search Algorithm:** sub-ottimale ma più rapido

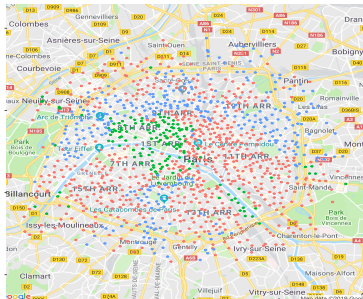
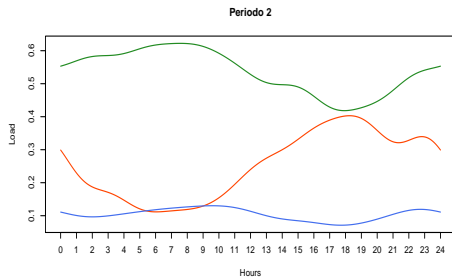
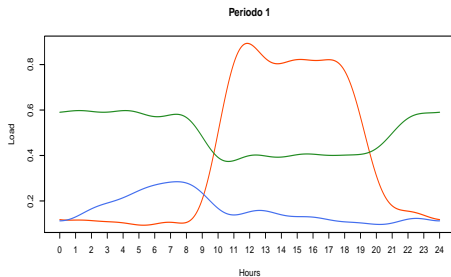
Tornando a Parigi

- Dati funzionali $\mathbf{X} = (x_{ij}(t))_{1 \leq i \leq n, 1 \leq j \leq p}$ con $t \in [1, \dots, 24]$:
 - Necessario pre-smoothing delle curve per lavorarci
 - Per $p(\mathbf{X}|\mathbf{z}, \mathbf{w}; \theta)$ viene assunto un modello probabilistico che descriva le curve in un adeguato sottospazio
- Risultati: $K = 3, L = 2$

Zona1	Zona2	Zona3
440	181	568
(0.37)	(0.15)	(0.48)

Periodo1	Periodo2
5	2
(0.71)	(0.29)

Tornando a Parigi



- ❓ Commenti
- ❓ Domande
- ❓ Interpretazioni
- ❓ Habitù di Parigi

Tornando ai gatti in dieta













- Maggiore problema da affrontare: **SPARSITÀ**
Utenti con più di 20 recensioni e prodotti recensiti più di 50 volte



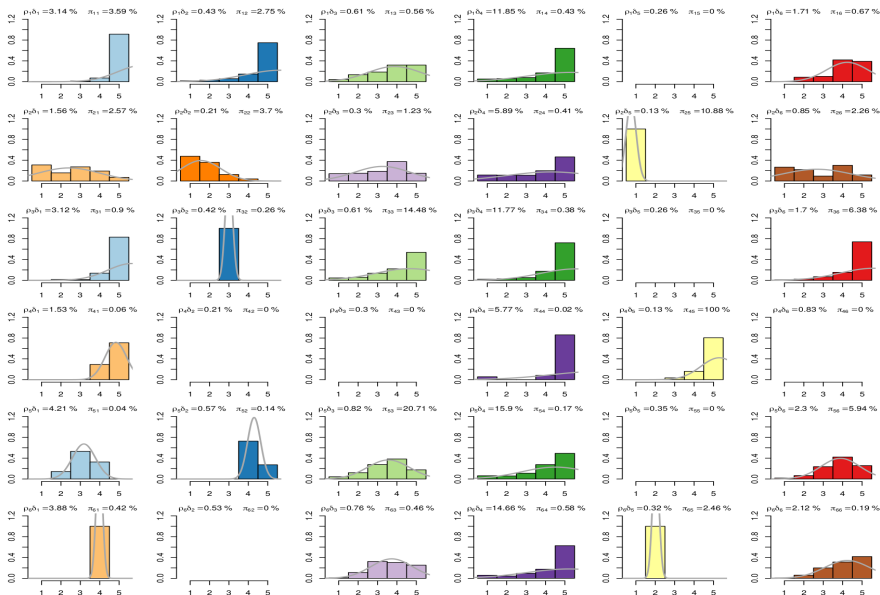
Matrice di dati con $N = 1644$, $P = 1733$ e percentuale di dati mancanti pari a 98.85%

- Tecniche di clustering darebbero risultati non affidabili e a basso contenuto informativo
- Dati ordinali $\mathbf{X} = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ con $x_{ij} \in \{1, \dots, 5\}$:
 - Per $p(\mathbf{X}|\mathbf{z}, \mathbf{w}; \theta)$ viene assunto un modello generatore basato su una variabile gaussiana latente (v. modello probit cumulativo)

Amazon - Risultati

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

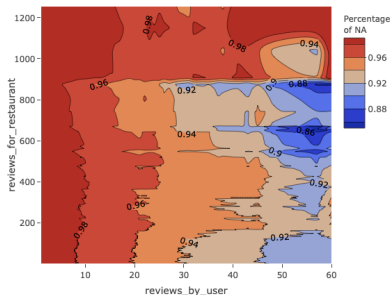
Amazon - Risultati



Work in progress

🦉 **Scopo:** costruire un sistema di raccomandazione da integrare nei servizi di TripAdvisor che integri informazioni aggiuntive su prodotti e clienti

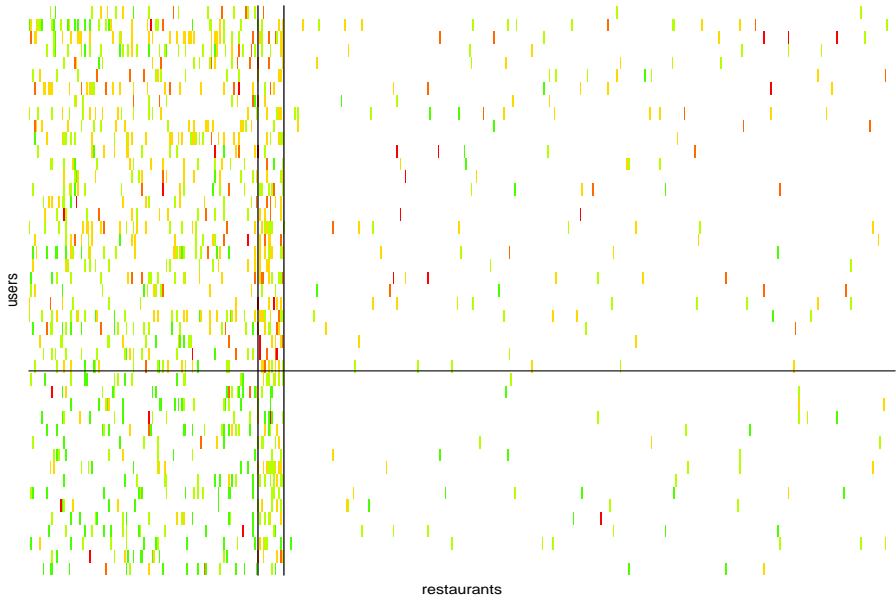
🦉 **Struttura dati:** recensioni TripAdvisor per tutti bar e ristoranti in provincia di Padova dal 2011 al 2018



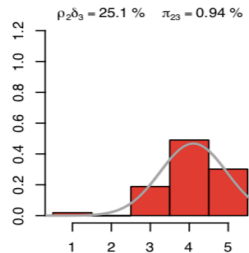
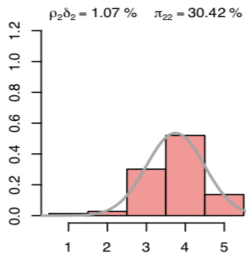
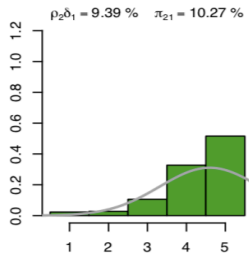
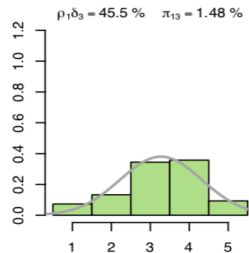
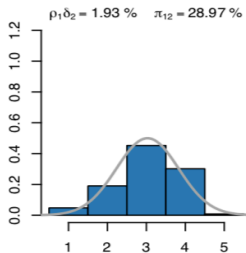
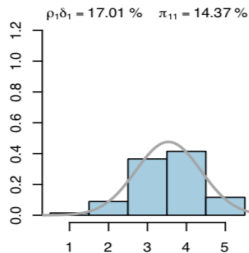
🦉 42263 utenti, 709 ristoranti

🦉 97555 recensioni, perc valori mancanti pari al 99.7%

Work in progress



Work in progress



- **R packages:**

- `blockcluster`: dati continui, binari, categoriali
- `funLBM`: dati funzionali
- `ordinalClust`, `ordinalLBM`: dati ordinali

- **Articoli/libri:**

- Govaert, G. & Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons
- Bouveyron, C. et al (2018). *The functional latent block model for the co-clustering of electricity consumption curves*. JRSS-C, 67(4), 897-915.
- Bergé, L.R. et al (2019). *The latent topic block model for the co-clustering of textual interaction data*. CSDA
- Corneli, M. et al (2019). *Co-clustering of ordinal data via latent continuous random variables and a classification EM algorithm*. <https://hal.archives-ouvertes.fr/hal-01978174>