Modelli per dati di rete

Introduzione

- I dati di rete forniscono informazioni sulle relazioni tra oggetti, individui o entità che chiamiamo nodi
- Genericamente si tratta di relazioni tra coppie di nodi
- Una coppia di nodi viene chiamata diade
- Una quantità misurata o osservata per molte diadi viene chiamata variabile diadica
- Esempi: scambi di merce tra paesi, comunicazioni tra persone, connessioni tra regioni del cervello ...

Introduzione

- Una variabile diadica misurata su una popolazione di n nodi può essere sintetizzata con una matrice di adiacenza pesata (sociomatrice) $\mathbf{Y}$ di dimensioni $n \times n$
- $y_{i,j}$ misura la relazione tra i nodi i e j, nella direzione da i a j, $i \rightarrow j$
- Esempio: dataset che descrive connessioni tra persone tramite email
- $y_{i,j}$ rappresenta quanti messaggi i ha inviato a j

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,3} & NA & y_{1,5} & \dots \\ y_{2,1} & y_{2,2} & y_{2,3} & y_{2,4} & y_{2,5} & \dots \\ y_{3,1} & y_{3,2} & y_{3,3} & y_{3,4} & NA & \dots \\ y_{4,1} & y_{4,2} & y_{4,3} & y_{4,4} & y_{4,5} & \dots \\ y_{5,1} & y_{5,2} & y_{5,3} & y_{5,4} & y_{5,5} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Media di una sociomatrice

È semplice calcolare la media di una 'sociomatrice'

$$Y = \begin{bmatrix} NA & 0 & 2 & 6 \\ 1 & NA & 0 & 0 \\ 0 & 0 & NA & 0 \\ 3 & 0 & 2 & NA \end{bmatrix}$$

Calcoliamo la media:

$$\begin{aligned} \sum_{i \neq j} y_{i,j} &= 14 \\ n(n-1) &= 12 \end{aligned}$$

Dunque $\hat{\mu} = 14/12 \approx 1.67$.

Medie di riga e di colonna

Sia $y_{i,j}$ una relazione da i a j

Media totale: media di tutte le osservazioni

$$\bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{n(n-1)} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}$$

Media di riga: media di tutte le osservazioni per ogni riga

$$\bar{y}_{i\cdot} = \frac{y_{i\cdot}}{n-1} = \frac{1}{n-1} \sum_{j:j \neq i} y_{i,j}$$

Media di colonna: media di tutte le osservazioni per ogni colonna

$$\bar{y}_{\cdot j} = \frac{y_{\cdot j}}{n-1} = \frac{1}{n-1} \sum_{i:i \neq j} y_{i,j}$$

Eterogeneità a livello di riga

- Spesso i valori di una variabile in una certa riga sono tra loro correlati $\rightarrow$ valori alti e bassi non sono egualmente distribuiti tra le righe
- Questo comporta eterogeneità tra le medie di riga di una sociomatrice
- Relazioni all'interno di una stessa riga sono caratterizzate dallo stesso 'mittente' i
- Se l'individuo 1 è più 'socievole' dell'individuo 2, ci aspettiamo valori per la riga di 1 più elevati di quelli della riga di 2
- Eterogeneità nella 'socialità' dei nodi comporta variabilità nelle medie di riga

Eterogeneità a livello di colonna

- Ragionamento analogo vale per le colonne
- Relazioni all'interno di una stessa colonna sono caratterizzate dallo stesso 'ricevente' j
- Eterogeneità nella 'popolarità' dei nodi comporta variabilità nelle medie di colonna

Scomposizione additiva di una sociomatrice

Scomposizione ANOVA

La variabilità di $y_{i,j}$ intorno alla media μ è data da effetti additivi di riga e colonna, a_i e b_j

$$y_{i,j} = \mu + a_i + b_j + \varepsilon_{i,j}$$

- Media totale: $\mu = \bar{y}_{..}$
- Eterogeneità di a_i, $b_j \rightarrow$ eterogeneità nelle medie di riga e di colonna

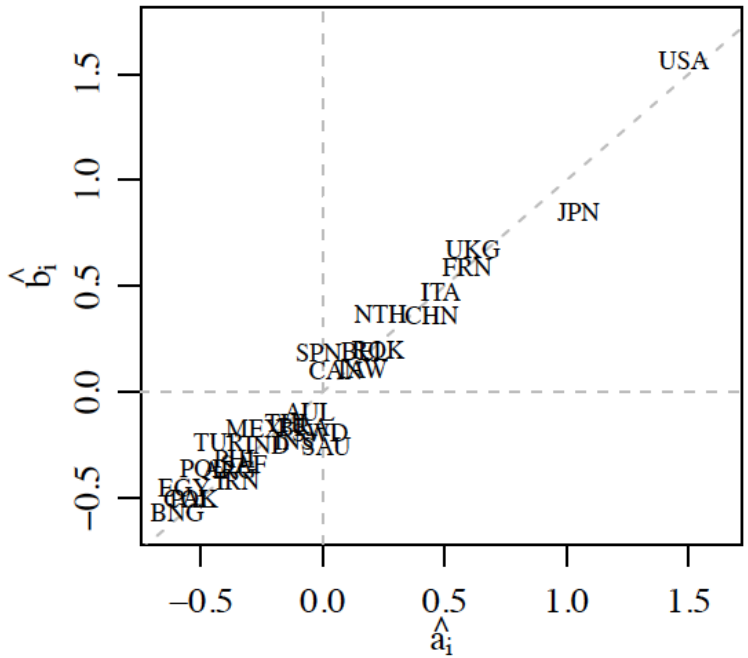
Scomposizione additiva di una sociomatrice

- Analisi ANOVA ignora che ogni nodo è contemporaneamente mittente e ricevente
- Ogni nodo i è coinvolto in due effetti additivi: effetto di riga a_i ed effetto di colonna b_i
- Dato che ogni coppia di effetti (a_i, b_i) condivide un nodo, ci aspettiamo correlazione tra i vettori $(a_1, \ldots, a_n)$ e $(b_1, \ldots, b_n)$
- Spesso ci interessa capire se nodi 'socievoli' sono anche 'popolari'
- Inoltre per ogni coppia $\{i, j\}$ ci sono due osservazioni $y_{i,j}$ e $y_{j,i}$
- Ragionevole aspettarsi che $y_{i,j}$ e $y_{j,i}$ siano correlate

Esempio

- Sociomatrice dei dati di esportazioni di $n = 30$ paesi, $i = 1, \ldots, 30$
- $y_{i,j}$ volume di esportazioni nell'anno 2018 dal paese i al paese j
- Per ogni paese, $\hat{a}_i$ è la media della riga i-esima meno la media totale $\hat{\mu}$, $\hat{b}_i$ è la media della colonna i-esima meno la media totale $\hat{\mu}$

Esempio

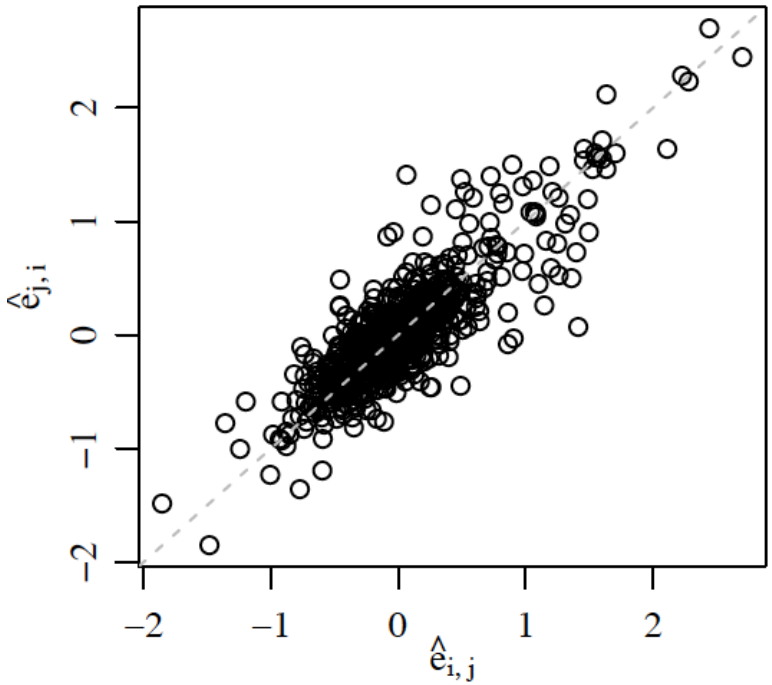


Forte correlazione tra effetti di riga e di colonna

$\rightarrow$ paesi con grandi volumi di esportazione saranno anche grandi importatori

Esempio

$$\hat{\varepsilon}_{i,j} = y_{i,j} - (\hat{\mu} + \hat{a}_i + \hat{b}_j)$$



Forte correlazione diadica tra $\hat{\varepsilon}_{i,j}$ e $\hat{\varepsilon}_{j,i}$

$y_{i,j}$ e $y_{j,i}$ sono correlate anche al netto di effetti di riga e colonna

Come tenerne conto?

Social Relations Model

Una versione più completa dell'ANOVA è il Social Relations Model, SRM

$$\begin{aligned} y_{i,j} &= \mu + a_i + b_j + \varepsilon_{i,j} \\ \{(a_1, b_1) \ldots, (a_n, b_n)\} &\sim N(0, \Sigma_{ab}) \quad i.i.d. \\ \{(\varepsilon_{i,j}, \varepsilon_{j,i})\} &\sim N(0, \Sigma_{\varepsilon}) \quad i.i.d. \end{aligned}$$

dove

$$\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}$$

e

$$\Sigma_{\varepsilon} = \sigma_{\varepsilon}^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Social Relations Model

- Condizionatamente agli effetti di riga $(a_1, \ldots, a_n)$, la media della riga i è $\mu + a_i$ e la variabilità delle medie di riga è data da σ_a^2
- Quindi σ_a^2 descrive l'eterogeneità delle medie di riga
- Analogamente, σ_b^2 descrive l'eterogeneità delle medie di colonna
- La covarianza σ_{ab} descrive la correlazione tra medie di riga e di colonna
- Variabilità addizionale descritta da σ_ε^2 e correlazione diadica catturata da ρ (oltre a quella descritta da σ_{ab})

Inferenza sui parametri

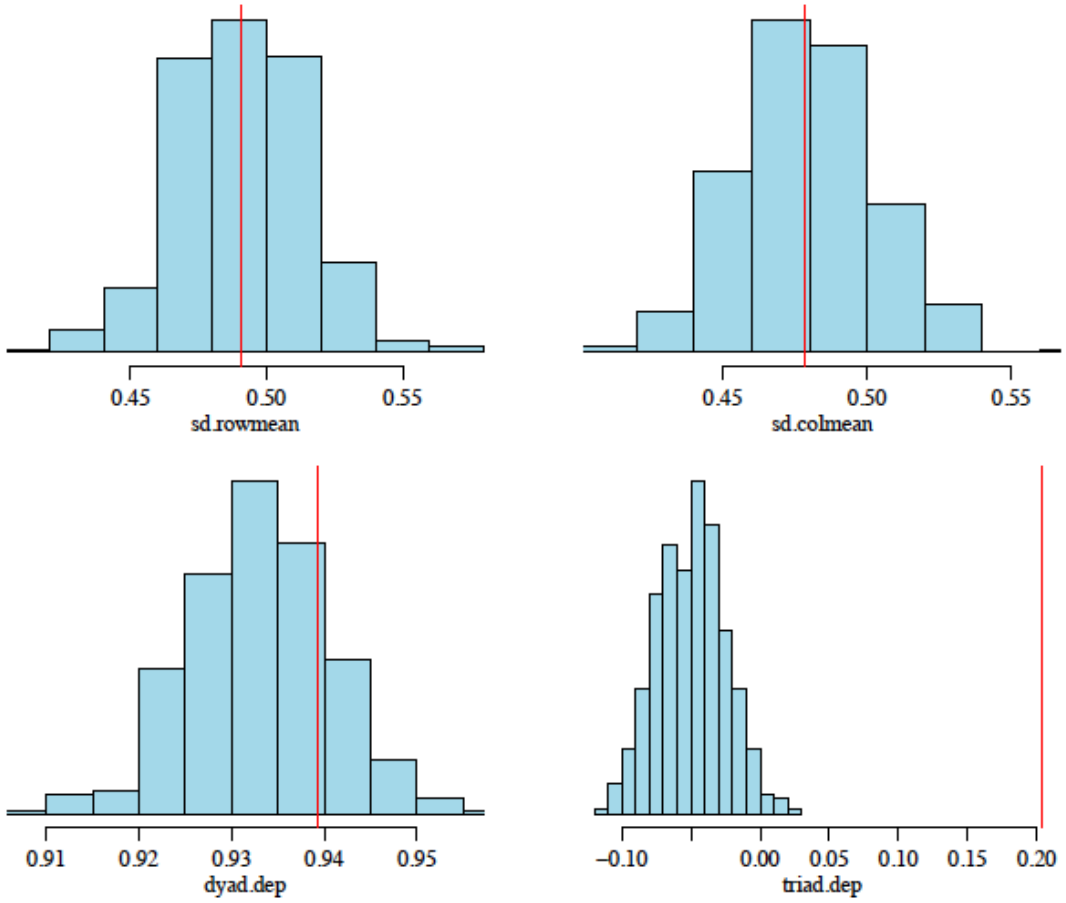
- Per un modello SRM con errori normali, è possibile ottenere una stima via massima verosimiglianza.
- tuttavia, quando abbiamo a che fare con dati binari e ordinali, la funzione di verosimiglianza richiede il calcolo di integrali difficilmente trattabili (dovuti alla combinazione delle trasformazioni e delle dipendenze indotte dal modello SRM). Si possono usare approssimazioni di vario tipo...
- Invece, è relativamente semplice ottenere un algoritmo di Gibbs sampling per simulare dalla distribuzione a posteriori (una volta definite delle ragionevoli distribuzioni a priori).
- È quindi utile/sensato adottare un approccio bayesiano alla stima (non quindi per scelta "filosofica", ma per comodità).
- In questo caso, siccome ci basiamo su un algoritmo numerico (Gibbs sampling - MCMC), non avremo una forma analitica per le distribuzioni a posteriori, ma ci baseremo su simulazioni delle distribuzioni a posteriori.
- La stima bayesiana di modelli di tipo SRM è implementata nel pacchetto `amen` dell'ambiente `R`.

Statistiche di sintesi

- $\hat{\sigma}_a$ deviazione standard delle medie di riga
- $\hat{\sigma}_b$ deviazione standard delle medie di colonna
- $\hat{\rho} = cor(\hat{\varepsilon}_{ij}\hat{\varepsilon}_{ji})$ dipendenza diadica
- $\sum_{i,j,k} \hat{\varepsilon}_{ij}\hat{\varepsilon}_{jk}\hat{\varepsilon}_{ik}$ dipendenza triadica

Social Relations Model

Statistiche di rete ed adattamento



Distribuzioni a posteriori e statistiche osservate (linea verticale rossa).

In generale, grandi discrepanze tra distribuzioni a posteriori e statistiche osservate

Social Relations Regression Model

- Spesso siamo interessati a misurare la relazione tra una variabile diadica (risposta) e altre variabili diadiche o di nodo.
- Il modello Social Relations Regression Model, SSRM, combina un modello di regressione lineare con la struttura di un SRM

$$y_{i,j} = \beta^\top \mathbf{x}_{i,j} + a_i + b_j + \varepsilon_{i,j}$$

dove $\mathbf{x}_{i,j}$ è un vettore p-dimensionale di regressori e β è il vettore di coefficienti di regressione.
Il vettore $\mathbf{x}_{i,j}$ può contenere variabili di nodo o diadiche.
Ad esempio $\mathbf{x}_{i,j} = \mathbf{x}_{r,i}, \mathbf{x}_{c,j}, \mathbf{x}_{d,i,j}$ contiene:

- caratteristiche del nodo i come mittente $\mathbf{x}_{r,i}$
- caratteristiche del nodo j come ricevente $\mathbf{x}_{c,j}$
- caratterstiche della coppia ordinata (i, j) $\mathbf{x}_{d,i,j}$

Esempio

- Consideriamo il dataset sulle esportazioni
- Si hanno altre variabili relative al paese: prodotto interno lordo (GDP), 'polity' (misura di apertura politica del paese), distanza geografica tra coppie di capitali
- Si è interessati a valutare se la variabile 'polity' abbia un ruolo nel determinare il volume di esportazioni, dopo aver controllato gli effetti di GDP e distanza geografica

Viene dapprima stimato un modello lineare del tipo:

$$y_{i,j} = \beta_0 + \beta_{r,1}\text{polity}_i + \beta_{r,2}\text{GDP}_i + \beta_{c,1}\text{polity}_j + \beta_{c,2}\text{GDP}_j + \beta_d\text{distance}_{i,j} + \varepsilon_{i,j}$$

Esempio

Confronto tra modello lineare e SRRM

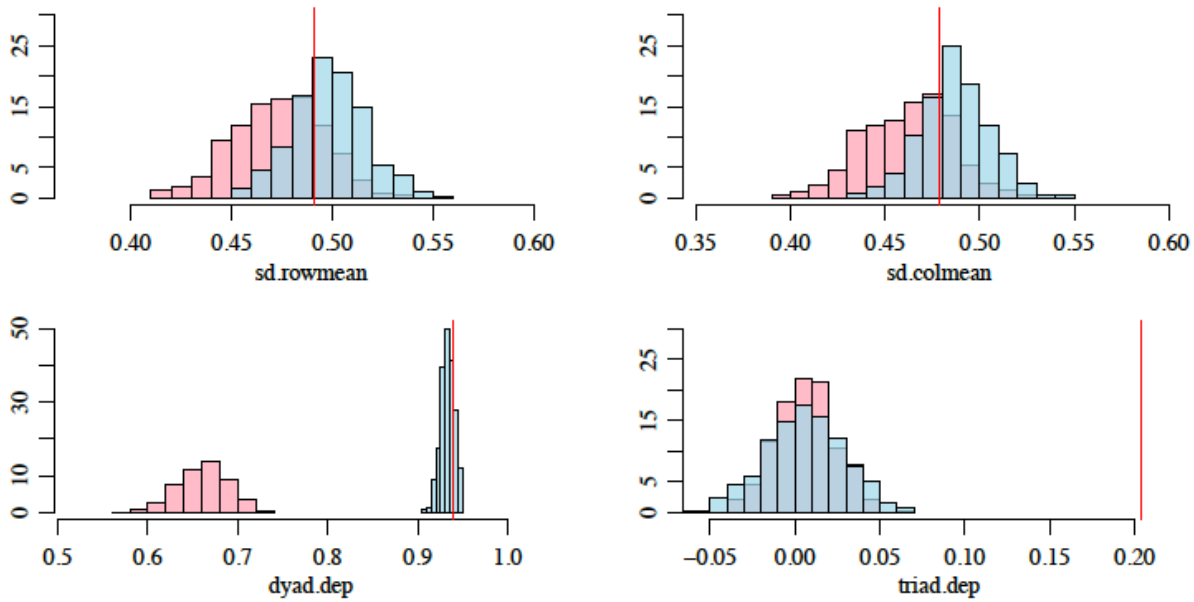
	LM			SRRM		
regressore	$\hat{\beta}$	se($\hat{\beta}$)	t-ratio	$\hat{\beta}$	se($\hat{\beta}$)	t-ratio
exporter polity	0.015	0.004	4.166	0.015	0.016	0.934
importer polity	0.022	0.004	6.070	0.022	0.016	1.419
exporter GDP	0.411	0.021	19.623	0.407	0.095	4.302
importer GDP	0.398	0.020	19.504	0.397	0.094	4.219
distance	-0.057	0.004	-13.360	-0.064	0.005	-11.704

A prima vista il modello lineare sembrerebbe essere una valida opzione . . . ma noi ci aspettiamo effetti di riga, colonna e dipendenza diadica per dati come questi!
Possiamo stimare quindi un modello SRRM

$$y_{i,j} = \beta_0 + \beta_{r,1}\text{polity}_i + \beta_{r,2}\text{GDP}_i + \beta_{c,1}\text{polity}_j + \beta_{c,2}\text{GDP}_j + \beta_d\text{distance}_{i,j} + a_i + b_j + \varepsilon_{i,j}$$

Esempio

Confronto tra modello lineare e SRRM



Distribuzioni a posteriori per regressione lineare (rosa) e SRRM (blu).
Guardando la correlazione diadica, il modello SRRM appare più adeguato.