IL METODO STATIS
Structuration des TAbleaux a Trois Indices de la Statistique

Analisi Statistica dei Dati Multidimensionali[1]

[1]Corso di Laurea in Scienze Statistiche e Attuariali

Facoltà di Scienze Economiche e Aziendali
Università degli Studi del Sannio

Introduzione

- Negli ultimi anni, si è assistito da parte degli studiosi ad una maggiore attenzione allo sviluppo di metodiche statistiche di tipo multivariato per lo studio dei fenomeni economici evolutivi miranti a coglierne gli aspetti strutturali e dinamici. Tali fenomeni risultano caratterizzati da più insiemi di misurazioni di diverse variabili rilevate su uno stesso (o diversi) insieme di unità statistiche in diverse occasioni (tempo, luoghi, situazioni, etc ...) e rappresentati, in genere, in forma tabellare.
- Si supponga un fenomeno caratterizzato da K tabelle di dati quantitativi rilevate in differenti tempi o in differenti occasioni. In base alle caratteristiche dei dati rilevati possiamo individuare tre differenti situazioni che ci consentono di mettere in evidenza diversi aspetti strutturali e dinamici del fenomeno.

Differenti situazioni: **STESSE UNITÀ STATISTICHE**

- Questa situazione è caratterizzata dalle unità statistiche che restano le stesse al variare delle occasioni o degli istanti di tempo in cui vengono misurate le variabili. Tale strutturazione dei dati ci consente di evidenziare l'evoluzione delle posizioni relative delle unità statistiche come quelle delle occasioni.
- Il fenomeno è quindi caratterizzato da n unità statistiche, K occasioni e p_k $(k = 1, \ldots K)$ variabili reali osservate sulle n unità all'occasione k $(n \geq p_k, \forall k)$ ed i dati vengono raccolti nelle K matrici $\mathbf{X}_k$ di dimensioni $n \times p_k$.

Differenti situazioni: **STESSE UNITÀ STATISTICHE**

- Sia $\Re^n$ lo spazio vettoriale delle funzioni numeriche definite sull'insieme delle unità statistiche. Rispetto alla base canonica di $\Re^n$ ad ogni variabile corrisponde un vettore appartenente ad $\Re^n$.
- Tale spazio risulta munito di una metrica rappresentata da una matrice **M** diagonale definita positiva contenente i pesi attribuiti (di somma unitaria) alle unità statistiche.
- Ogni gruppo di variabili è pertanto rappresentato da un insieme di vettori.
- Sia, inoltre, $\Re^{p_k}$ lo spazio vettoriale delle funzioni numeriche definite sull'insieme dei caratteri per la k-esima tabella. Rispetto alla base canonica di $\Re^{p_k}$ ad ogni gruppo di variabili corrisponde una nube di unità statistiche. Possiamo definire una metrica $\mathbf{Q}_k$ per ciascun spazio delle variabili $\Re^{p_k}$ la cui struttura dipende dalla natura delle variabili considerate.

- Risultano, quindi, definiti i seguenti K studi

$$(\mathbf{X}_1, \mathbf{Q}_1, \mathbf{M}), \quad (\mathbf{X}_2, \mathbf{Q}_2, \mathbf{M}), \quad \ldots, \quad (\mathbf{X}_K, \mathbf{Q}_K, \mathbf{M})$$

- L'elemento caratterizzante di ciascun studio sarà fornito dalla matrice dei prodotti scalari

$$\boxed{\mathbf{W}_k\mathbf{M} = \mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^T\mathbf{M}}$$

  operatore dello spazio vettoriale $\Re^{n\times n}$ delle prossimità fra le unità statistiche.
- La scelta di $\mathbf{W}_k\mathbf{M}$ come elemento caratteristico di uno studio si giustifica con le informazioni che esso apporta sui legami esistenti fra le unità statistiche oggetto dell'indagine.

Differenti situazioni: **STESSE UNITÀ STATISTICHE**

- In questa situazione due studi risultano equivalenti se

$$\boxed{\mathbf{W}_k\mathbf{M} = \mathbf{W}_{k'}\mathbf{M}} \text{ con } k \neq k^T$$

oppure se esiste uno scalare $v > 0$ tale che

$$\mathbf{W}_k\mathbf{M} = v\mathbf{W}_{k'}\mathbf{M}$$

- Nel caso particolare

$$\mathbf{Q}_k = (\mathbf{X}_k^T\mathbf{M}\mathbf{X}_k)^{-1}$$

(metrica di Malahanobis) l'elemento caratterizzante sarà dato dall'operatore di proiezione **M**-ortogonale

$$\boxed{\mathbf{A}_k\mathbf{M} = \mathbf{X}_k(\mathbf{X}_k^T\mathbf{M}\mathbf{X}_k)^{-1}\mathbf{X}_k^T\mathbf{M}}$$

rispetto al sottospazio di $\Re^n$ generato dai vettori colonna della matrice $\mathbf{X}_k$.

- Questa situazione è caratterizzata dalle variabili che restano le stesse al variare delle occasioni in cui vengono rilevate. Tale organizzazione dei dati ci consente di evidenziare l'evoluzione delle variabili come delle occasioni.
- Il fenomeno risulta, allora, caratterizzato da p variabili reali, K occasioni nelle quali le variabili vengono osservate su n_k unità statistiche ($k = 1, \ldots K$). Tali dati verranno raccolti, allora, nelle K matrici $\mathbf{X}_k$ di dimensioni $n_k \times p$.
- Si può munire lo spazio $\Re^p$ delle variabili di una metrica rappresentata da una matrice $\mathbf{Q}$ diagonale definita positiva così come ogni spazio $\Re^{n_k}$ delle unità statistiche risulta munito di una metrica rappresentata da una matrice $\mathbf{M}_k$ diagonale contenente i pesi attribuiti (di somma unitaria) alle n_k unità statistiche.

- Risultano, quindi, definiti i seguenti K studi

$$(\mathbf{X}_1, \mathbf{Q}, \mathbf{M}_1), \quad (\mathbf{X}_2, \mathbf{Q}, \mathbf{M}_2), \quad \ldots, \quad (\mathbf{X}_K, \mathbf{Q}, \mathbf{M}_K)$$

In questa situazione si prenderà come elemento caratterizzante di ciascun studio la matrice di varianza e covarianza

$$\boxed{\mathbf{V}_k = \mathbf{X}_k^T \mathbf{M}_k \mathbf{X}_k}$$

operatore dello spazio vettoriale $\Re^{p \times p}$ che gode delle stesse proprietà dell'operatore $\mathbf{W}_k\mathbf{M}$:

due studi risultano equivalenti se $\mathbf{V}_k = \mathbf{V}_{k'}$ (con $k \neq k'$) oppure se esiste uno scalare $v > 0$ tale che $\mathbf{V}_k = v\mathbf{V}_{k'}$.

- Questa situazione contempla le due precedenti e risulta caratterizzata dalla presenza delle stesse variabili e delle stesse unità statistiche per le varie occasioni. I dati sono contenuti nelle K matrici $\mathbf{X}_k$ aventi tutte le stesse dimensioni $n \times p$.
- Lo spazio $\Re^p$ delle variabili risulta munito di una metrica rappresentata dalla matrice $\mathbf{Q}$ mentre quello delle unità da una metrica rappresentata dalla matrice $\mathbf{M}$.
- Risultano così definiti i K studi

$$(\mathbf{X}_1, \mathbf{Q}, \mathbf{M}), \quad (\mathbf{X}_2, \mathbf{Q}, \mathbf{M}), \quad \ldots, \quad (\mathbf{X}_K, \mathbf{Q}, \mathbf{M})$$

  i cui elementi caratteristici possono essere $\mathbf{W}_k\mathbf{M}$, $\mathbf{A}_k\mathbf{M}$ oppure $\mathbf{V}_k$.
- In questo tipo di situazione si può essere interessati all'evoluzione delle posizioni relative delle unità statistiche come a quella delle variabili oppure all'evoluzione di entrambi.

Situazione	Studio e dimens. matrice $\mathbf{X}_k$	Elemento caratteristico	Condizione d'equivalenza (con $k \neq k'$)
Stesse unità statistiche	$(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{M})$ $n \times p_k$	$\mathbf{W}_k\mathbf{M}$; $\mathbf{A}_k\mathbf{M}$	$a)$ $\begin{cases} \mathbf{W}_k\mathbf{M} = \mathbf{W}_{k'}\mathbf{M} \\ \mathbf{W}_k\mathbf{M} = v\mathbf{W}_{k'}\mathbf{M} \\ \mathbf{A}_k\mathbf{M} = \mathbf{A}_{k'}\mathbf{M} \end{cases}$
Stesse variabili	$(\mathbf{X}_k, \mathbf{Q}, \mathbf{M}_k)$ $n_k \times p$	$\mathbf{V}_k$	$b)$ $\begin{cases} \mathbf{V}_k = \mathbf{V}_{k'} \\ \mathbf{V}_k = v\mathbf{V}_{k'} \end{cases}$
Stesse variabili ed unità statistiche	$(\mathbf{X}, \mathbf{Q}, \mathbf{M})$ $n \times p$	$\mathbf{W}_k\mathbf{M}$; $\mathbf{V}_k$; $\mathbf{A}_k\mathbf{M}$	$a)$ e $b)$

Table: Tabella riassuntiva delle situazioni riscontrabili

- Il metodo Statis (Hermeir Des Plantes, 1976; Escoufier, 1979) è un metodo esplorativo dell'Analisi dei Dati longitudinali;
- L'obiettivo è quello di confrontare, in un contesto di tipo fattoriale, K studi statistici fornendo una rappresentazione grafica dell'evoluzione dell'insieme delle unità statistiche (delle variabili) al variare di una terza dimensione.
- Il metodo Statis, nella sua formulazione originaria, consta di tre fasi: **Interstruttura**, **Compromesso** ed **Intrastruttura**.
- Queste tre fasi si presentano tra loro legate e nell'interpretazione dei risultati vanno considerate congiuntamente.

- L'idea di base del metodo è quella della ricerca di una struttura comune agli studi (**Interstruttura**) attraverso la loro rappresentazione su un opportuno spazio vettoriale.
- La similarità fra coppie di studi viene espressa in funzione dei rispettivi elementi caratteristici $\mathbf{G}_k$ tramite il Coefficiente di Correlazione Vettoriale

$$RV(\mathbf{G}_k, \mathbf{G}_{k'}) = \frac{tr(\mathbf{G}_k\mathbf{M}\mathbf{G}_{k'}\mathbf{M})}{\sqrt{tr(\mathbf{G}_k\mathbf{M})^2 tr(\mathbf{G}_{k'}\mathbf{M})^2}}$$

(con $0 \leqslant RV(\mathbf{G}_k, \mathbf{G}_{k'}) \leqslant 1$) o il Coefficiente di Covarianza Vettoriale

$$COVV(\mathbf{G}_k, \mathbf{G}_{k'}) = tr(\mathbf{G}_k\mathbf{M}\mathbf{G}_{k'}\mathbf{M})$$

per $k \neq k'$.

- L'elemento caratteristico per ciascun studio sarà dato dalla matrice **G**, di dimensione $k \times k$
- ⋆ dei prodotti scalari nel senso della traccia (*COVV*)

$$\mathbf{G}_{k,k'} = tr(\mathbf{G}_k \mathbf{M} \mathbf{G}_{k'} \mathbf{M})$$

con $k \neq k'$ fra gli elementi caratteristici $\mathbf{G}_k$ che potranno essere, a seconda della situazione, degli operatori dei prodotti scalari ($\mathbf{W}_k\mathbf{M}$), delle matrici di varianza e covarianza ($\mathbf{V}_k$) o dei proiettori ($\mathbf{A}_k\mathbf{M}$),

- ⋆ o delle similarità fra coppie di studi (*RV*)

$$\mathbf{G}_{k,k'} = \frac{tr(\mathbf{G}_k \mathbf{M} \mathbf{G}_{k'} \mathbf{M})}{\sqrt{tr(\mathbf{G}_k \mathbf{M})^2 tr(\mathbf{G}_{k'} \mathbf{M})^2}}$$

- Gli autovettori di **G** forniranno una rappresentazione globale, analoga ad una rappresentazione dell'Analisi in Componenti Principali, sulla quale la prossimità di due punti è significativa della prossimità degli elementi caratteristici nel senso della metrica utilizzata.
- Le distanze dei punti-studio dall'origine vengono interpretati in riferimento al cerchio di correlazione di una ACP:
- Uno studio è ben rappresentato se il suo punto rappresentativo è vicino al cerchio di raggio unitario;
- Gli angoli tra i vettori congiungenti i punti con l'origine hanno dei coseni che nel caso di rappresentazione collineare sono uguali ai coefficienti *RV*.

- Un coefficiente $RV(\mathbf{G}_k, \mathbf{G}_{k'})$ prossimo all'unità, con $\mathbf{G}_k = \mathbf{W}_k\mathbf{M}$, indica che si ha la stessa struttura delle unità statistiche all'interno delle tabelle $\mathbf{X}_k$ e $\mathbf{X}_{k'}$ e che le posizioni reciproche delle unità sono stabili. Per i prodotti scalari di tipo *COVV* la rapresentazione ottenuta è analoga a quella delle variabili di una ACP su una matrice di covarianza.
- Il problema della ricerca di un sottospazio di dimensioni ridotte in modo tale che le prossimità fra gli elementi caratteristici siano rispettate si traduce (Rao, 1964) nella ricerca di una matrice $\mathbf{B}$ di rango h $(h < K)$ tale che $\|\mathbf{G} - \mathbf{B}\|^2$ sia minimo e dove $\|.\|$ indica la norma euclidea.
- Il minimo è raggiunto per $\mathbf{B} = \sum_{\alpha=1}^{h} \lambda_\alpha u_\alpha u'_\alpha$ con λ_α e u_α, rispettivamente, autovalori ed autovettori della matrice $\mathbf{G}$. Le coordinate degli studi sui primi h assi saranno fornite da:$\sqrt{\lambda_1}u_1, \sqrt{\lambda_2}u_2, \ldots, \sqrt{\lambda_h}u_h$.

Fase del COMPROMESSO

- Nella fase del compromesso si vuole individuare uno studio di riferimento che sia rappresentativo dell'insieme dei K studi.
- L'obiettivo è allora quello di trovare un elemento caratteristico fittizio, della stessa natura di quelli considerati nella fase dell'Interstruttura, che abbia la proprietà di riassumere gli elementi caratteristici considerati.
- In tale ambito diverse proposte sono state effettuate in letteratura (Escoufier, 1980; Ralambondrainy, 1984).

- Escoufier (1980) individua un operatore compromesso $\mathbf{W}_c\mathbf{M}$ o $\mathbf{V}_c\mathbf{Q}$ di norma massima e tale da massimizzare la somma dei quadrati degli operatori scalari fra questo operatore e tutti gli altri.
- Ciò porta a considerare il seguente problema (con $\mathbf{G}_k = \mathbf{W}_k\mathbf{M}$)

$$\left| \begin{array}{l} \max tr\,(\mathbf{W}_c\mathbf{M})^2 \text{ e } \max \sum_{k=1}^{K} tr\,(\mathbf{W}_k\mathbf{M}\mathbf{W}_c\mathbf{M})^2 \\ \mathbf{W}_c\mathbf{M} = \sum_{k=1}^{K} \mu_k \mathbf{W}_k\mathbf{M} \text{ dove } \sum_{k=1}^{K} \mu_k^2 = 1 \text{ con } \mu_k \geq 0 \end{array} \right.$$

- La soluzione è fornita dall'autovettore associato all'autovalore dominante $\mathbf{u}_1$ della matrice semi-definita positiva $\mathbf{G}$. Gli elementi avranno tutti lo stesso segno e quindi si possono scegliere positivi (Teorema di Frobenius).
- I coefficienti μ_k saranno gli elementi dell'autovettore così individuato.

Fase del COMPROMESSO

- È evidente il criterio di come Statis individua il referenziale comune dei vari studi: ricerca una opportuna combinazione lineare degli operatori associati agli studi

$$\mathbf{W}_c\mathbf{M} = \sum_{k=1}^{K} \mu_k \mathbf{W}_k\mathbf{M}$$

  con pesi pari agli elementi dell'autovettore $\mathbf{u}_1$.
- Il piano fattoriale costituito dai primi due autovettori $\mathbf{u}_1$ ed $\mathbf{u}_2$ della matrice $\mathbf{W}_c\mathbf{M}$ ci consente una rappresentazione piana delle n unità statistiche compromesso.

Fase dell' INTRASTRUTTURA

- Nella fase dell'intrastruttura si ricerca una rappresentazione completa di tutte le unità statistiche e di tutte le variabili nello spazio del compresso precedentemente individuato.
- Tale fase (con $\mathbf{G}_k = \mathbf{W}_k\mathbf{M}$) consiste nella proiezione delle unità statistiche e delle variabili nel sottospazio generato dagli autovettori $\mathbf{U}$ della matrice $\mathbf{W}_c\mathbf{M}$ tale che $\mathbf{U}^T\mathbf{M}\mathbf{U} = \Gamma$ dove Γ è la matrice diagonale diagonale degli autovalori $\lambda_1, \ldots, \lambda_n$.
- Al sottospazio generato dagli autovettori $\mathbf{U}$ risulta associato l'operatore di proiezione ortogonale $\mathbf{P} = \mathbf{U}\Gamma^{-1}\mathbf{U}^T\mathbf{M}$.

Fase dell' INTRASTRUTTURA

- In letteratura sono state avanzate diverse proposte, per la determinazione delle coordinate delle unità statistiche.
- Seguendo l'impostazione data da Place (1980), le coordinate delle $n \times K$ unità statistiche nello spazio del compromesso sono fornite dalle colonne della matrice $\mathbf{PE}_k$ dove $\mathbf{E}_k$ $(k = 1, \ldots, K)$ è la matrice contenente le componenti principali della matrice $\mathbf{G}_k\mathbf{M}$.
- Questo criterio, però, porta, per la situazione 1, ad una difficile interpretazione dell'evoluzione delle unità statistiche mentre sembra indicato per evidenziare i legami tra le unità statistiche e le variabili attraverso una rappresentazione congiunta mentre, nella situazione 2, conduce ad una difficile interpretazione delle variabili.

Fase dell' INTRASTRUTTURA

- Un diverso approccio è quello seguito da Glacon (1981) che ha il merito di ottenere sul grafico le unità o le variabili compromesso centrate rispetto alle unità o alle variabili dei K studi.
- Tale approccio prevede di proiettare nello spazio, generato dagli elementi di ogni studio, gli autovettori della matrice compromesso:

$$\mathbf{U}\Gamma^{-1/2}$$

Fase dell' INTRASTRUTTURA

Dati

n = num. di individui
p = num. di variabili
K = num. di tabelle $(k = 1, \ldots, K)$
M = mat. diag. pesi individui

Tabelle

$X_k \quad (n \times p_k)$
$X_k \quad (n_k \times p)$
$X_k \quad (n \times p)$

$\Downarrow$

OPERATORE G_k

matr. prodotti scalari
matr. varianza e cov.
oper. di proiezione

TABELLA DEI PROD. SCALARI FRA OPERATORI

$$G_{(K\times K)} = \left[\quad g_{kk'} \quad \right]$$

$$g_{kk'} = tr(G_k M G_{k'} M)$$

$$g_{kk'} = \frac{tr(G_k M G_{k'} M)}{[tr\,(G_k M)^2 tr\,(G_{k'} M)^2]^{1/2}}$$

$\Downarrow$

Fase dell' INTRASTRUTTURA

Diagonalizzazione di G

INTERSTRUTTURA=Rappresentazione piana dei K studi

$\Downarrow$

Operatore Compromesso

$$W_c M = \sum_{k=1}^{K} \eta_k W_k M$$
$$V_c Q = \sum_{k=1}^{K} \eta_k V_k Q$$

$u_1 = [\eta_1, \ldots, \eta_K]$ autovettore di G

$\Downarrow$

Diagonalizzazione di $W_c M$ o di $V_c Q$

Rappresentazione piana delle n unità statistiche compromesso
Rappresentazione piana delle p variabili compromesso

INTRASTRUTTURA $=$ Rappresentazione piana delle $n \times K$ unità statistiche
Rappresentazione piana delle $p \times K$ variabili

Tab. 2: Fasi del metodo Statis

Uno studio sul commercio internazionale di macchinari

- Negli ultimi decenni si è assistito, nell'ambito degli scambi commerciali fra le diverse nazioni, ad un duplice fenomeno: una crescente espansione degli scambi internazionali ed una sempre maggiore specializzazione dei paesi.
- Lo sviluppo della specializzazione commerciale è sicuramente spiegabile dalla specializzazione tecnologica raggiunta dal paese in quanto il progresso tecnologico e la competitività sui mercati internazionali son legati fra loro secondo un processo iterativo e cumulativo.

- Ciò è interpretabile in quanto i miglioramenti avvenuti in paese sono conseguenze delle esperienze e delle conoscenze tecnologiche passate.
- Si può, quindi, ipotizzare che i paesi tecnologicamente più avanzati conservano, nel tempo, il loro vantaggio creando nei mercati delle forme monopolistiche.
- Per analizzare il fenomeno nella sua interezza, senza rinunciare alla dimensione temporale, è necessario ricorrere a tecniche statistiche multivariate che tendono a cogliere aspetti strutturali e dinamici e, a tal fine, analizzeremo la specializzazione internazionale nel settore dei macchinari avvalendoci dei metodi descritti.

I dati riguardano le esportazioni riferite a 58 paesi su 9 divisioni nel periodo di tempo 1987-1992. Le divisioni si riferiscono al settore 7 della classificazione OCSE:

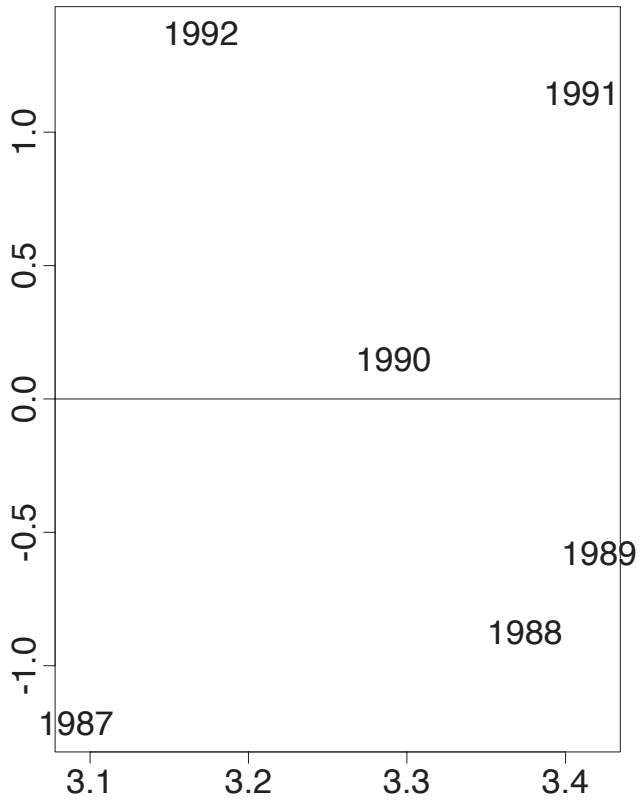
- macchinari per la generazione di energia e attrezzature (ENER)
- macchinari ed apparecchiature specializzate per particolari industrie (INDU)
- macchinari ed apparecchiature per la lavorazione dei metalli (META)
- altri macchinari industriali e pezzi di ricambio (RICA)
- macchinari ed attrezzature informatiche (INFO)
- apparecchiature per la riproduzione del suono e per le telecomunicazioni (TELE)
- macchinari ed apparecchiature elettriche (ELET)
- veicoli stradali (STRA)
- altri mezzi di trasporto (TRAS)

- L'indice di specializzazione utilizzato è stato calcolato mediante il rapporto dei rapporti tra l'esportazione del gruppo di prodotti j del paese i nell'anno t (x_{ijt}) con quelle del mondo w (x_{wjt}) e delle esportazioni del totale dei manufatti m nello stesso anno del paese i (x_{imt}) con quelle del mondo w (x_{wmt}):

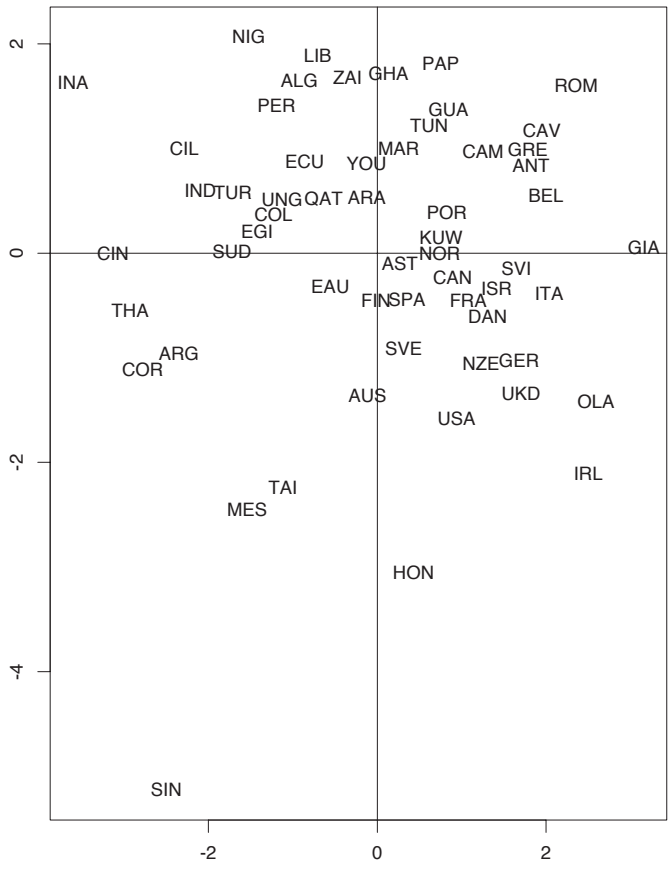
$$S_{ijt} = \frac{x_{ijt}/x_{imt}}{x_{wjt}/x_{wmt}}$$

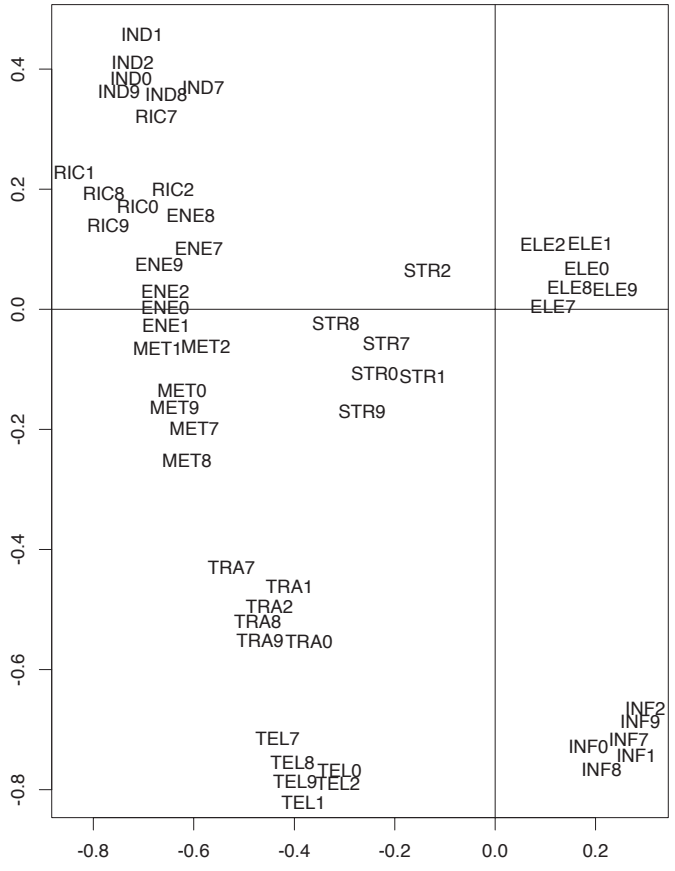
- L'indice S_{ijt} varia tra zero e più infinito e assume un valore pari ad uno quando non vi è più specializzazione. Infatti se S_{ijt} è prossimo all'unità il livello di esportazione del prodotto j nell'anno t per il paese i è, in percentuale, simile a quello medio mondiale.

Interstruttura



Compromesso





Intrastruttura