Probablità, Statistica e Processi Stocastici

Franco Flandoli, Università di Pisa

Corso per la Scuola di Dottorato in Ingegneria

Riepilogo su PCA (affiancare scheda R)

- Si parte da n variabili aleatorie $X_1, ..., X_n$. In pratica, si parte da una tabella di dati con n colonne (le variabili) e N individui (es. le regioni italiane ecc.)
- Il comando `princomp` esegue PCA, cioè calcola la matrice $n \times n$ di covarianza (empirica) Q, la sua decomposizione spettrale $\mathbf{e}_1, ..., \mathbf{e}_n$, $\lambda_1 \geq ... \geq \lambda_n$;
- poi calcola rispetto alla nuova base $\mathbf{e}_1, ..., \mathbf{e}_n$, la proiezione dei dati (nel caso empirico) o del vettore $\mathbf{X} = (X_1, ..., X_n)$ (nella visione teorica)

$$V_i = \langle \mathbf{X}, \mathbf{e}_i \rangle$$

- Il comando `biplot` rappresenta poi graficamente il piano principale $\mathbf{e}_1, \mathbf{e}_2$ e le proiezioni dei punti (ed anche una rappresentazione schematica delle proiezioni delle variabili $X_1, ..., X_n$).

Le variabili nascoste in PCA ed i loadings (affiancare scheda R)

Dal fatto che due basi di $\mathbb{R}^n$ sono legate da opportune combinazioni lineari discende con facili calcoli che lo sono anche le variabili:

$$\begin{aligned} X_1 &= a_{11} V_1 + \ldots + a_{1n} V_n \\ &\ldots \\ X_n &= a_{n1} V_1 + \ldots + a_{nn} V_n \end{aligned}$$

dove i coefficienti a_{ij} sono detti "loadings".

Le variabili nascoste ed i loadings

Se ora isoliamo le prime due variabili V_1, V_2

$$\begin{aligned} X_1 &= a_{11} V_1 + a_{12} V_2 + \epsilon_1 \\ &\dots \\ X_n &= a_{n1} V_1 + a_{n2} V_2 + \epsilon_n \end{aligned}$$

dove $\epsilon_1 = a_{13} V_3 + \dots + a_{1n} V_n$ ecc. abbiamo trovato una relazione lineare approssimata tra le due variabili V_1, V_2 e le grandezze di partenza, che vengono quindi "spiegate" da tali variabili. Abbiamo una riduzione della dimensione del problema, del numero di variabili necessarie a descrivere il nostro fenomeno.
Gli errori sono piccoli:

$$Var\left[\epsilon_1\right] = a_{13}^2 \lambda_3 + \dots + a_{1n}^2 \lambda_n \leq \lambda_3$$

(vale $a_{11}^2 + \dots + a_{1n}^2 = 1$). La tecnica denominata Factor Analysis è una versione più avanzata di quanto appena detto.

Le variabili nascoste ed i loadings

Quindi, se ad esempio avessimo un modello di regressione lineare multipla (RLM) della forma

$$Y = b_1 X_1 + ... + b_n X_n + \epsilon$$

potremmo semplificarlo nella forma

$$Y = c_1 V_1 + c_2 V_2 + \epsilon'.$$

E' possibile che in qualche esempio pratico questa relazione sia più precisa del modello complesso, se le variabili nascoste in realtà catturano meglio la "fisica" del problema.
I coefficienti c_i si cacolano dai coefficienti b_i e dai loadings.
Questa tacnica può essere utile in caso di *allineamenti tra fattori*, nella RLM.

Detour sulla regressione lineare multipla (vedi scheda R)

Base di tantissimi modelli è il semplice modello lineare con tanti input $X_1, ..., X_n$ (detti fattori, predittori) ed un output Y della forma

$$Y = a_1 X_1 + ... + a_n X_n + b + \epsilon.$$

Scritto così è inteso in senso teorico, come legame tra variabili. Se invece abbiamo dei dati, scriveremo la relazione

$$y_i = a_1 x_{1,i} + ... + a_n x_{n,i} + b + \epsilon_i$$

dove $i = 1, ..., N$ è l'indice che distingue gli "individui", le "unità sperimentali".
L'individuo i-esimo è descritto dai valori sperimentali

$$x_{1,i}, ..., x_{n,i}, y_i.$$

Il numero ϵ_i è il *residuo* dell'individuo i-esimo.

Regressione lineare multipla (affiancare scheda R)

I coefficienti del modello, $a_1, ..., .a_n, b$, vengono calcolati con metodo dei minimi quadrati: sono quelli che minimizzano lo scarto quadratico medio dei residui

$$\min_{a_1,...,.a_n,b} \sum_{i=1}^{N} \epsilon_i^2.$$

Ci sono formule esplicite per questi coefficienti, basate sull'inversa di una certa matrice (il software non deve innescare un vero procedimento di minimizzazione).

Scopi della regressione lineare multipla

Due dei principali scopi di un modello regressivo

$$Y = a_1 X_1 + ... + a_n X_n + b + \epsilon$$

1. *fare previsioni*: se di un nuovo "individuo" conosciamo solo i valori $x_{1,i}, ..., x_{n,i}$, possiamo stimare il suo valore $\widehat{y}_i$ tramite la formula $\widehat{y}_i = a_1 x_{1,i} + ... + a_n x_{n,i} + b$
2. analizzare una realtà, es. *capire l'importanza dei fattori* (può servire per indirizzare gli investimenti), quanto essi influiscano sull'output. Ad esempio, se tutte le quantità sono positive, il numero

$$\frac{a_1 x_{1,i}}{a_1 x_{1,i} + ... + a_n x_{n,i} + b} \sim a_1 \frac{x_{1,i}}{y_i}$$

descrive quanto conta (in %) il fattore n. 1, per l'individuo i-esimo.

2.bis Oppure si può calcolare un indice medio, tipo

$$a_1 \frac{\sigma_{X_1}}{\sigma_Y}$$

(se la tabella è standardizzata, esso è $= a_1$).

L'esempio della scheda R

- Nella scheda allegata cerchiamo di contruire un modello regressivo tra le variabili PLIC, SC, SA.SC, TMI (viste come input, fattori) e la variabile TD (vista come output).
- La tabella è standardizzata, quindi l'importanza media dei fattori coincide con i coefficienti del modello.
- Possibile che TMI, correlato 0.48 a TD, abbia importanza solo 0.0089?
- O ancor peggio, che a fronte delle correlazioni -0.85 e 0.9, tra SC e SA.SC con TD (correlazioni quasi uguali), i coefficienti della regressione siano -0.3 e 0.64 (cioè SA.SC ha importanza doppia rispetto a SC)?
- (Fatte separatamente eliminando uno dei due, le due regressioni danno importanze pari a -0.79 e 0.89, quasi uguali alle correlazioni)
- (la somma invece, 0.64+0.3, è più simile 0.9)

Difetti causati dagli allineamenti

- Si dice che due fattori, es. X_1 e X_2, sono allineati, se sono molto correlati.
- Detto alla buona, sono un po' una copia l'uno dell'altro, quindi tenerli entrambi è, in qualche misura, inutile.
- La presenza di fattori allineati non influisce più di tanto se lo scopo della regressione è previsivo: si spreca solo della fatica, perché si calcolano dei doppioni.
- Invece la loro presenza è *distruttiva* se lo scopo è capire l'importanza dei fattori. Tra poco cercheremo di capire il perché.
- Conviene quindi ricondursi a pochi fattori più scorrelati possibile. Un modo è eliminare i doppioni (o i gruppi con più di due molto correlati).
- Un altro è sostituire ai fattori originari, le componenti principali offerta da PCA (si ricordi che sono scorrelate).
- (Questo però sposta il problema, offre l'importanza di V_1 ecc....)

Difetti causati dagli allineamenti

Vediamo di capire perché una coppia di fattori allineati ha conseguenze distruttive sull'utilizzo di un modello di regressione per lo scopo di capire l'importanza dei fattori.
Pensiamo ad un caso limite: $X_1 = X_2$ (correlazione $= 1$), scorrelati da un altro fattore X_3. Supponiamo che, se usassimo solo X_1 e X_3 il modello giusto sia

$$Y = 5 \cdot X_1 + 2 \cdot X_3.$$

Allora anche il seguente modello va altrettanto bene:

$$Y = 15 \cdot X_1 - 10 \cdot X_2 + 2 \cdot X_3$$

e come lui infiniti altri (es. $Y = -7 \cdot X_1 + 12 \cdot X_2 + 2 \cdot X_3$). Il SW "non sa" quale scegliere.

Difetti causati dagli allineamenti

Confrontiamo i due modelli

$$Y = 5 \cdot X_1 + 2 \cdot X_3$$

$$Y = 15 \cdot X_1 - 10 \cdot X_2 + 2 \cdot X_3.$$

Per scopi previsivi sono simili. Ma dal secondo dedurremmo che l'importanza relativa di X_1 (per un individuo) è

$$\frac{15 \cdot x_{1,i}}{y_i}$$

mentre in realtà è solo

$$\frac{5 \cdot x_{1,i}}{y_i}.$$

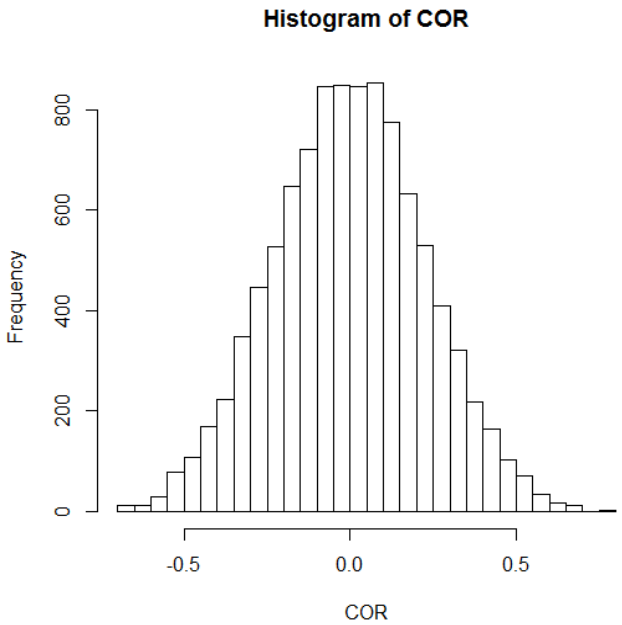
Circa il secondo fattore, si troverebbe addirittura un'importanza negativa, contrariamente al fatto che esso influisca positivamente su Y.

Ammissione di vaghezza

- Abbiamo visto che, nella RLM, la presenza di fattori allineati è perniciosa (specialmente in fase di interpretazione del modello, es. valutazione dell'importanza dei fattori)
- Possiamo eliminare uno dei due fattori di una coppia allineata, ma perdiamo informazione, se il fattore eliminato non è proprio una copia dell'altro (si pensi ad una correlazione ~ 0.5)
- Possiamo applicare PCA, ma poi abbiamo l'importanza delle componenti principali invece che dei fattori di partenza. E' quindi l'idea migliore? Forse è solo una delle varie cose da tentare.
- Possiamo semplicemente vedere i valori della matrice di correlazione. Essi non catturano l'importanza dei fattori *all'interno del modello*, quindi sono rozzi, ma danno un'idea della maggior o minor rilevanza.
- Vari documenti applicativi in rete terminano dicendo: cercate nuovi fattori, meno allineati! (bella forza)

Detour sull'entità del coefficiente di correlazione

Dobbiamo spesso capire se un coefficiente di correlazione indica un'elevato legame, un po' di legame, o è compatibile con assenza di legame. Si può allora tracciare l'istogramma della correlazione tra coppie indipendenti della lunghezza desiderata e giudicare ad occhio. L'esempio mostra il caso $N = 20$, dove correlazioni dell'ordine di 0.3 sono compatibili con l'indipendenza, mentre 0.7 no.



Serie storiche (verso fPCA)

- La tecnica chiamata fPCA (functional PCA) esamina serie storiche utilizzando paradigmi propri di PCA.
- E' utile premettere un po' di terminologia riguardante le serie storiche, per capire meglio.
- Quindi invertiamo leggermente l'ordine del programma, cioè anticipiamo alcune idee sulle serie storiche.
- Una serie storica è una sequenza di numeri $x_1, ..., x_n$ in cui l'indice $1, ..., n$ corrisponde al tempo, discetizzato come serve a seconda del problema (giorni, mesi anni ecc.).
- Ad esempio, il volume mensile di esportazioni italiane di automobili è una serie storica, reperibile sul sito Eurostat.

Esercizio riassuntivo

Svolgiamo insieme un esercizio riassuntivo sui vettori gaussiani.

1. Creiamo un vettore di punti gaussiani generati con R, che corrispondano ad una gaussiana dilatata in una direzione e ruotata di $\pi/10$
2. Di tali punti calcoliamo la matrice di covarianza empirica, che chiameremo Q
3. Di Q calcoliamo la radice quadrata $\sqrt{Q}$
4. Usando $\sqrt{Q}$, simuleremo punti gaussiani aventi covarianza Q
5. osserveremo ad occhio la somiglianza coi punti creati all'inizio (volendo ci sarebbero dei test statistici)

L'esercizio si trova svolto in una scheda a parte.