

A13



Vai al contenuto multimediale

Biagio Palumbo
Antonio Lepore
Christian Capezza
Fabio Centofanti

**Introduzione ai modelli
di analisi multivariata a variabili latenti**

Applicazioni per l'ingegneria





Aracne editrice

www.aracneeditrice.it
info@aracneeditrice.it

Copyright © MMXVIII
Giacchino Onorati editore S.r.l. – unipersonale

www.giacchinoonoratieditore.it
info@giacchinoonoratieditore.it

via Vittorio Veneto, 20
00020 Canterano (RM)
(06) 45551463

ISBN 978-88-255-1333-2

*I diritti di traduzione, di memorizzazione elettronica,
di riproduzione e di adattamento anche parziale,
con qualsiasi mezzo, sono riservati per tutti i Paesi.*

*Non sono assolutamente consentite le fotocopie
senza il permesso scritto dell'Editore.*

I edizione: marzo 2018

Indice

7	<i>Acronimi</i>
9	<i>Simboli e notazione</i>
11	<i>Prefazione</i>
13	Capitolo I <i>Introduzione</i>
	1.1. Analisi di regressione multivariata lineare, 13 – 1.2. Metodo dei minimi quadrati, 15 – 1.3. Limiti della regressione multivariata lineare, 16 – 1.4. Metodi a variabili latenti per la riduzione della dimensionalità, 18 – 1.4.1. <i>Modelli bilineari a variabili latenti</i> , 20.
21	Capitolo II <i>Calibrazione</i>
	2.1. PCR, 22 – 2.1.1. <i>PCA</i> , 22 – 2.1.2. <i>Caso studio reale: monitoraggio delle emissioni di CO₂ di una nave</i> , 26 – 2.1.3. <i>Previsione</i> , 28 – 2.1.4. <i>Algoritmi</i> , 29 – 2.1.5. <i>Proprietà della PCR</i> , 29 – 2.2. PLS, 33 – 2.2.1. <i>PLS1/PLS2</i> , 34 – 2.2.2. <i>NIPALS</i> , 38 – 2.2.3. <i>SIMPLS</i> , 39 – 2.2.4. <i>Altri approcci NIPALS-PLS</i> , 40.
41	Capitolo III <i>Ispezione del modello</i>
	3.1. Outlier, 41 – 3.1.1. <i>Osservazioni anomale in calibrazione</i> , 42 – 3.1.2. <i>Osservazioni anomale in previsione</i> , 45 – 3.2. Validazione e selezione del modello, 46 – 3.2.1. <i>Validazione esterna</i> , 46 – 3.2.2. <i>Validazione interna</i> , 47 – 3.2.3. <i>Scelta del numero di componenti</i> , 49 – 3.3. Pretrattamento dei dati, 51 – 3.3.1. <i>Normalizzazione</i> , 51 – 3.3.2. <i>Trasformazioni</i> , 53.
57	Capitolo IV <i>Monitoraggio</i>
	4.1. Carta T^2 di Hotelling per gli score, 59 – 4.1.1. <i>Contributi delle variabili a T^2</i> , 62 – 4.2. Carta dell'errore quadratico di previsione delle variabili indipendenti, 63 – 4.2.1. <i>Contributi delle variabili a SPE</i> , 65 – 4.3. Errore di previsione della variabile di risposta, 66.

73 *Bibliografia*

77 *Appendice. Richiami di Algebra lineare*

87 *Ringraziamenti*

Acronimi

ARL	average run length
ATS	average time to signal
BLUE	best linear unbiased estimator
LCL	lower control limit
MSE	mean squared error
MSEE	mean squared estimation error
MSEP	mean squared error of prediction
NIPALS	nonlinear iterative partial least squares
OLS	ordinary least squares
PCA	principal component analysis
PCR	principal component regression
PLS	partial least squares
PRESS	predicted residual sum of squares
RMSEP	root mean squared error of prediction
SIMPLS	statistically inspired modification of PLS
SPC	statistical process control
SPE	squared prediction error
SVD	single value decomposition
UCL	upper control limit

Simboli e notazione

A	Matrice diagonale $A \times A$ con elementi diagonali uguali a $l_1, \dots, l_A$
$\Sigma^{\mathbb{Z}}$	Matrice di covarianza, $\{\sigma_{i,j}^{\mathbb{Z}}\}$, di dimensioni $p \times p$ del vettore di variabili aleatorie $\mathbb{Z}$
B	Matrice di dimensioni $(p + 1) \times m$ le cui righe sono i vettori $\beta_{k_i}, k = 0, 1, \dots, p$
E	Matrice dell'errore di dimensioni $n \times p$ della regressione di $\mathbb{X}$ su $\mathbb{T}$
F	Matrice dell'errore di dimensioni $n \times m$ della regressione di $\mathbb{Y}$ su $\mathbb{T}$
G	Matrice dell'errore di dimensioni $n \times m$ della regressione di $\mathbb{Y}$ su $\mathbb{X}$
P	Matrice di dimensioni $p \times A$ le cui colonna sono i vettori $\mathbf{p}_1, \dots, \mathbf{p}_A$
$\mathbf{p}_i$	i -esimo vettore dei loading
Q	Matrice di dimensioni $m \times A$ dei coefficienti di regressione di $\mathbb{Y}$ su $\mathbb{T}$
$\mathbf{q}_i$	i -esima colonna di Q
T	Matrice di dimensioni $n \times A$ le cui colonne sono i vettori $\mathbf{t}_1, \dots, \mathbf{t}_A$
$\mathbf{t}_i$	i -esimo vettore degli score
V	Matrice di dimensioni $p \times A$ tale che $\mathbb{T}^{\top} = \mathbb{X}^{\top} \mathbf{V}$
$\mathbf{v}_i$	i -esima colonna di V
X	Matrice di dimensioni $n \times (p + 1)$ che ha per righe i vettori $\begin{bmatrix} 1 & \mathbf{x}_{(i)}^{\top} \end{bmatrix}$
$\mathbf{x}_{(i)}$	Determinazione di $\mathbb{X}_i$
Y	Matrice di dimensioni $n \times m$ che ha per righe $\mathbf{y}_{(i)}^{\top}$
$\mathbf{y}_{(i)}$	Determinazione di $\mathbb{Y}_i$
$\mathbf{z}_{(i)}^{\top}$	Vettore corrispondente alla i -esima riga della generica matrice Z
β_0	Vettore dei termini noti nella regressione di $\mathbb{Y}$ su $\mathbb{X}$

β_k	Vettore dei coefficienti di regressione relativi alla k -esima variabile indipendente nella regressione di $\mathbb{Y}$ su $\mathbb{X}$
$\mu^{\mathbb{Z}}$	Media del vettore di variabili aleatorie $\mathbb{Z}$
ε	Vettore degli errori con $E(\varepsilon) = 0$ e matrice di covarianza $\Sigma^{\mathbb{Y} \mathbb{X}}$
ε_i	i -esima osservazione di ε
$\hat{\theta}$	Stimatore di θ
$\mathbb{T}$	Vettore delle A variabili latenti
$\mathbb{X}$	Vettore dei p regressori (delle variabili indipendenti)
$\mathbb{X}_i$	$= (X_{i1}, \dots, X_{ip})^\top$, i -esima osservazione del vettore dei regressori $\mathbb{X}$
$\mathbb{Y}$	Vettore delle m variabili di risposta (variabili dipendenti)
$\mathbb{Y}_i$	$= (Y_{i1}, \dots, Y_{ip})^\top$, i -esima osservazione del vettore delle variabili di risposta $\mathbb{Y}$
$\mu_i^{\mathbb{Z}}$	Media della i -esima componente del vettore di variabili aleatorie $\mathbb{Z}$
$\bar{\mathbf{z}}$	Vettore delle medie campionare di $\mathbb{Z}$
$\sigma_{ij}^{\mathbb{Z}}$	Covarianza tra i -esima e j -esima componente del vettore di variabili aleatorie $\mathbb{Z}$
A	Numero di variabili latenti
l_a	a -esimo autovalore di $\mathbf{X}^\top \mathbf{X}$
m	Numero di variabili dipendenti
n	Numero di osservazioni
p	Numero di variabili indipendenti

Prefazione

Il presente volume nasce da un'attività di ricerca promossa in ambito industriale, finalizzata alla previsione ed al monitoraggio delle emissioni di CO₂ in atmosfera dovute al consumo di carburante durante la navigazione di una nave da carico e passeggeri nel Mar Mediterraneo. La mancanza in letteratura di una trattazione completa e sintetica in lingua italiana dei principali modelli di analisi multivariata a variabili latenti, utili a tale scopo, ha spinto gli autori a scrivere questo testo. I modelli presentati vengono implementati con specifico riferimento al contesto industriale preso in considerazione, ma la loro validità supera lo specifico campo di applicazione proposto. Il libro consente infatti ai lettori di familiarizzare con alcuni metodi statistici utili, in ottica Big Data & Analytics per Industria 4.0, al monitoraggio di sistemi complessi in ambito ingegneristico.

Il volume si apre con l'esposizione dei principali limiti connessi all'uso dei modelli di regressione multivariati e del metodo di stima dei minimi quadrati ordinari, nell'analisi di dati ad elevata dimensionalità, motivando in tal modo la trattazione dei modelli affrontata nei capitoli successivi. Il secondo capitolo introduce, in modo sintetico, i più comuni metodi di calibrazione per la riduzione della dimensionalità basati sui modelli bilineari a variabili latenti. Il terzo capitolo esamina, invece, alcune procedure classiche per verificare che il modello ottenuto in fase di calibrazione sia sufficientemente affidabile nell'effettuare previsioni, e tratta le principali tecniche per affrontare il problema di eventuali valori anomali, non rappresentativi della popolazione in esame. L'ultimo capitolo illustra i metodi di controllo statistico di processo più utilizzati nell'ambito dei metodi multivariati a variabili latenti. In particolare l'attenzione viene rivolta alla costruzione dei limiti per le carte da utilizzare in Fase I (off-line), coerenti con le tecniche già utilizzate per l'ispezione del modello, e in Fase II (on-line), per il monitoraggio di nuove osservazioni. Per dimostrarne le potenzialità, i metodi e le tecniche proposte vengo-

no applicate a un caso studio in ambito navale tramite l'ausilio del software statistico open source R.

Introduzione

Questo capitolo introduce il classico metodo di stima dei minimi quadrati per l'analisi di regressione lineare di modelli multivariati, e ne evidenzia i principali limiti. In tal modo viene motivata la trattazione, nei paragrafi successivi, dei metodi a variabili latenti per la riduzione della dimensionalità di un insieme di dati, che hanno lo scopo principale di trasformare un insieme di variabili, di un dato problema, in un nuovo insieme, costituito da un numero inferiore di variabili.

1.1. Analisi di regressione multivariata lineare

Si indichi con $\mathbb{X} = (X_1, \dots, X_p)^\top$ un vettore¹ di variabili aleatorie (di seguito chiamate variabili indipendenti, o regressori, o covariate) di media $\boldsymbol{\mu}^{\mathbb{X}} = (\mu_1^{\mathbb{X}}, \dots, \mu_p^{\mathbb{X}})^\top$, dove $\mu_k^{\mathbb{X}} = E(X_k)$ per ogni $k = 1, \dots, p$, e matrice di covarianza $\boldsymbol{\Sigma}^{\mathbb{X}} = \left\{ \sigma_{ij}^{\mathbb{X}} \right\}_{i,j=1,\dots,p}$ dove $\sigma_{ij}^{\mathbb{X}} = \text{Cov}(X_i, X_j)$. Con $E(\cdot)$ e $\text{Cov}(\cdot)$ si denotano, rispettivamente, gli operatori valore atteso e covarianza. Si indichi con $\mathbb{Y} = (Y_1, \dots, Y_m)^\top$ un vettore di variabili aleatorie (di seguito chiamate variabili dipendenti o di risposta) di media $\boldsymbol{\mu}^{\mathbb{Y}} = (\mu_1^{\mathbb{Y}}, \dots, \mu_m^{\mathbb{Y}})^\top$, dove $\mu_j^{\mathbb{Y}} = E(Y_j)$ per ogni $j = 1, \dots, m$, e matrice di covarianza $\boldsymbol{\Sigma}^{\mathbb{Y}} = \left\{ \sigma_{ij}^{\mathbb{Y}} \right\}_{i,j=1,\dots,m}$ dove $\sigma_{ij}^{\mathbb{Y}} = \text{Cov}(Y_i, Y_j)$.

L'esistenza di una relazione funzionale tra $\mathbb{X}$ e $\mathbb{Y}$ può essere indicata come

$$\mathbb{Y} \approx f(\mathbb{X}), \quad (1.1)$$

1. Tutti i vettori sono da intendersi come vettori colonna, in modo da facilitare la costruzione delle matrici per colonne. Pertanto, $\mathbf{x}_k$ indicherà la colonna k -esima di una matrice $\mathbf{X}$. Quando necessario, $\mathbf{x}_{(i)}^\top$ ne indicherà la riga i -esima.

dove il simbolo $\approx$ tiene conto dell'incertezza nella misura e nell'esatta identificazione di $f(\cdot)$.

Nel caso di regressione multivariata lineare, la relazione (1.1) può essere scritta come

$$\mathbb{Y} = \beta_0 + \sum_{k=1}^p X_k \beta_k + \varepsilon, \quad (1.2)$$

dove $\beta_0 = (\beta_{01}, \dots, \beta_{0m})^\top$ è il vettore dei parametri dei termini noti e $\beta_k = (\beta_{k1}, \dots, \beta_{km})^\top$ è il vettore degli m coefficienti di regressione della k -esima variabile indipendente. L'errore $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)^\top$ è un vettore di variabili aleatorie, di media $\mathbf{0}$ e matrice di covarianza $\Sigma^{\mathbb{Y}|\mathbb{X}}$, che tiene conto dell'inadeguatezza del modello nello spiegare il legame ipotizzato tra le variabili. Si assumerà che le variabili $\varepsilon_1, \dots, \varepsilon_m$ siano tra loro indipendenti, (i.e., $\Sigma^{\mathbb{Y}|\mathbb{X}}$ sia una matrice diagonale) e che siano indipendenti anche da $\mathbb{X}$ (i.e., $E(X_k \varepsilon_j) = 0$ per ogni $k = 1, \dots, p$ e per ogni $j = 1, \dots, m$). L'analisi di regressione multivariata ha lo scopo di verificare la relazione funzionale (1.2) per monitorare il legame tra la variabile di risposta e i regressori, e per effettuare previsioni. In particolare, mira ad identificare il criterio con cui determinare gli stimatori dei parametri β_k , $k = 0, 1, \dots, p$ del modello per descrivere un campione casuale $(\mathbb{Y}_i, \mathbb{X}_i)$, dove $\mathbb{Y}_i = (Y_{i1}, \dots, Y_{im})^\top$ è il vettore della i -esima osservazione delle variabili di risposta, $\mathbb{X}_i = (X_{i1}, \dots, X_{ip})^\top$ è il vettore della i -esima osservazione dei regressori.

Uno stimatore $\hat{\theta}$ di un parametro θ è una funzione che associa ad un campione di variabili aleatorie un valore nell'insieme di tutti i valori che θ può assumere. Essendo una funzione di variabili aleatorie, è esso stesso una variabile aleatoria. In corrispondenza di uno specifico campione di determinazioni, il valore assunto da $\hat{\theta}$ viene chiamato stima² di θ . Gli stimatori $\hat{\beta}_k$ dei parametri β_k sono funzione delle osservazioni dei regressori e delle variabili di risposta.

2. Per semplicità di notazione, essendo facilmente distinguibile dal contesto, anche la stima verrà indicata, se necessario, con $\hat{\theta}$.