# Capitolo 4

# Identificazione e caratterizzazione delle sorgenti di aerosol attraverso l'analisi fattoriale

L'identificazione delle sorgenti è uno degli obiettivi principali nello studio del particolato atmosferico, indispensabile per l'elaborazione di una strategia di abbattimento degli inquinanti. Dato che le particelle di aerosol mantengono in parte la composizione elementale della sorgente emettitrice, anche a lunghe distanze, la rivelazione simultanea di gruppi di elementi può essere di grande aiuto nell'individuazione delle sorgenti del particolato.

Se le concentrazioni di due o più elementi hanno un andamento temporale simile, significa che questi sono probabilmente soggetti agli stessi processi di produzione e/o di trasporto; pertanto le frazioni di particolato campionato dovute a sorgenti indipendenti possono essere identificate da opportune combinazioni lineari delle concentrazioni elementali, dette fattori, caratterizzate dall'essere tra loro scorrelate.

I modelli a recettore, come la *Principal Component Analysis* (PCA), hanno lo scopo di identificare e quantificare l'impatto delle sorgenti inquinanti a partire dall'analisi delle concentrazioni presenti nel sito di campionamento.

In questo capitolo sono riportati una breve descrizione della PCA e i risultati ottenuti dalla sua applicazione ai dati raccolti in questo lavoro.

## § 4.1 *Principal Component Analysis* (PCA)

La PCA risulta utile quando dobbiamo analizzare molti dati, tipicamente i valori assunti da diverse grandezze fisiche (variabili) su molti campioni. La tecnica consiste nel sostituire un grande numero di variabili correlate con un piccolo numero di variabili scorrelate, dette fattori.

Un set di dati può essere espresso in forma matriciale come:

$$X = \{x_{ij}\} \qquad i = 1, .., l; \quad j = 1, .., m \qquad (4.1)$$

con $x_{ij}$ uguale al valore della variabile $i$ nel campione $j$. Nel nostro caso, le variabili sono le concentrazioni in aria dei diversi elementi e i campioni sono i giorni di campionamento.

Il modello si basa sulla possibilità di scrivere:

$$z_{ij} = \sum_{k=1}^{n} b_{ik} \cdot g_{kj} \qquad Z = B \cdot G \qquad (4.2)$$

dove :

- $Z = \left\{ z_{ij} = \dfrac{x_{ij} - < x_i >}{\sigma_i} \right\}$ è la matrice delle variabili standardizzate; $<x_i>$ e $\sigma_i$ sono la media e la deviazione standard della concentrazione dell'i-esimo elemento; le nuove variabili hanno, quindi, media nulla e varianza unitaria; questa trasformazione di variabili fa sì che ciascun elemento abbia nei calcoli successivi ugual peso.
- $G = \{g_{kj}\}$ è la matrice dei factor scores che rappresentano l'andamento temporale delle sorgenti a meno di un fattore moltiplicativo e di un fattore additivo. I fattori sono standardizzati e scorrelati, ovvero:

$$R(g_k, g_{k'}) \equiv \frac{\sum_j (g_{kj} - < g_k >) \cdot (g_{k'j} - < g_{k'} >)}{N+1} = \begin{cases} 1 \; se \; k = k' \\ 0 \; se \; k \neq k' \end{cases} \tag{4.3}$$

- $B = \{b_{ik}\}$ è una matrice di costanti (matrice dei factor loadings); $b_{ik}$ è il coefficiente di correlazione fra la concentrazione dell'elemento *i* e il fattore *k*.

Le matrici *B* e *G* sono ottenute calcolando gli autovettori e autovalori della matrice di correlazione delle concentrazioni standardizzate $R(z_i, z_{i'})$:

$$b_{ik} = \sqrt{\lambda_k} \cdot \lambda_{ik} \tag{4.4}$$

$$g_{kj} = \frac{\sum_{i=1}^{n} \lambda_{ki} \cdot z_{ij}}{\sqrt{\lambda_k}} \tag{4.5}$$

dove $\lambda_k$ è l'autovalore *k*-esimo, $\lambda_{ki}$ è la componente *i* del *k*-esimo autovettore (normalizzato) e l'indice *k* è ordinato in ordine decrescente di autovalore, ovvero $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. La varianza totale[1] del sistema è la somma di tutti gli autovalori. La semplificazione del modello consiste nel prendere solo i primi *p* autovettori (con $p < n$) con autovalore maggiore[2], spiegando così una frazione della varianza totale pari a:

[1] La varianza totale di un sistema è definita come la traccia della matrice di covarianza; ma, visto che le variabili $z_{ij}$ sono standardizzate, allora la matrice di covarianza coincide con la matrice di correlazione, quindi:

$$T = \sum_{i=1}^{n} R(z_i, z_i) = n$$

Visto che la traccia di una matrice è invariante per diagonalizzazione, T è anche uguale alla somma degli autovalori $\lambda_k$ della matrice stessa.

[2] Ci sono diversi criteri per stabilire il numero di fattori da trattenere: il più semplice consiste nel scegliere quegli autovettori con autovalore corrispondente >1 visto che equivale a trattenere solo i fattori che "si portano dietro" più varianza delle variabili originarie; oppure fissare un valore per F e scegliere tanti autovettori fintanto che la somma degli autovalori corrispondenti non raggiunga F.

$$F = \frac{\lambda_1 + \lambda_2 + ... + \lambda_p}{\lambda_1 + \lambda_2 + ... + \lambda_n} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_p}{n} \tag{4.6}$$

La varianza della concentrazione dell'i-esimo elemento spiegata dai *p* fattori, detta comunalità, è:

$$h_i^2 = \sum_{k=1}^{p} b_{ik}^2 \tag{4.7}$$

La prima informazione fornita dal modello è, quindi, il numero di fattori necessario per spiegare una certa frazione della varianza totale e di ciascun elemento.

Dato che i *factor loadings* sono i coefficienti di correlazione tra i fattori e le concentrazioni elementari, utilizzando le informazioni presenti in letteratura sulla composizione delle sorgenti dell'aerosol, è possibile ipotizzare una relazione tra fattori e sorgenti.

Ritornando alla relazione (4.2), la matrice *B* non è in generale unica, ma è definita a meno di una trasformazione ortonormale *U* ($U \cdot U^T = 1$):

$$Z = B \cdot G = B \cdot U \cdot U^T \cdot G \equiv B^R \cdot G^R \tag{4.8}$$

$$B^R = B \cdot U \qquad G^R = U^T \cdot G \tag{4.9}$$

Data l'ortonormalità della trasformazione, i fattori ruotati, ovvero gli elementi della matrice $G^R$ sono sempre a varianza unitaria e scorrelati tra loro. Gli elementi di $B^R$, ovvero i *factor loadings* ruotati, hanno ancora il significato di correlazione tra le vecchie variabili, le concentrazioni elementali $z_i$ e quelle nuove, che in questo caso sono i fattori ruotati.

La non unicità di *B* permette di effettuare una trasformazione dei fattori che renda più semplice la loro associazione alle sorgenti inquinanti, come la rotazione

VARIMAX: questa trasformazione ruota i fattori in modo da massimizzare il numero dei *factor loadings* con valori prossimi a 0 e a 1, mantenendo inalterate la varianza totale e quelle dei singoli elementi spiegate dal modello. Dopo questa trasformazione gli elementi prodotti da una sola sorgente si trovano con peso più elevato all'interno di un solo fattore e quindi questo fattore può essere associato a tale sorgente.

Riassumendo, attraverso la PCA, si ottengono i *factor loadings* ($B^R$) ruotati che permettono di identificare le sorgenti del particolato e i *factor scores* ($G^R$) ruotati che rappresentano l'andamento temporale delle sorgenti identificate ma non in modo assoluto; infatti i *factor scores* sono standardizzati (media nulla e varianza unitaria) e quindi rappresentano l'andamento temporale delle sorgenti a meno di un fattore moltiplicativo e di un fattore additivo.

Ulteriori informazioni possono essere ottenute passando ad un'analisi di tipo "assoluto". In termini assoluti, la (4.2) può essere riscritta come:

$$x_{ij} \cong \sum_{k=1}^{p} a_{ik} \cdot f_{kj} \qquad X \cong A \cdot F \tag{4.10}$$

dove:

- $A = \{a_{ik}\}$ è la matrice dei <u>profili</u> delle sorgenti: $a_{ij}$ è la frazione dell'*i*-esimo elemento prodotto dalla sorgente *k* nel sito di campionamento;
- $F = \{f_{kj}\}$ è la matrice dei <u>pesi</u> delle sorgenti: $f_{kj}$ è la concentrazione di particolato campionato nel giorno *j* prodotta dalla sorgente *k*.

I profili e i pesi delle sorgenti forniscono, quindi, gli "apporzionamenti" elementali, che sono i contributi di ogni sorgente alla concentrazione misurata di ogni elemento: $c_{ijk} \equiv a_{ik} \cdot f_{kj}$ è il contributo ($\mu g/m^3$) della sorgente *k* all'elemento *i* nel giorno *j*.

Le relazioni che legano le equazioni (4.10) e (4.8) sono date da:

- $z_{ij} = \dfrac{x_{ij} - < x_i >}{\sigma_i}$ (4.11)

- $g_{kj}{}^{R} = \dfrac{f_{kj} - < f_k >}{\sigma_k}$ (4.12)

- $b_{ik}^{R} = \dfrac{a_{ik} \cdot \sigma_k}{\sigma_i}$ (4.13)

dove $< f_k >$ e $\sigma_k$ sono il valor medio e la *standard deviation* della sorgente *k*-esima che, a priori, non sono note.

Dalle equazioni (4.12) e (4.13) si ottiene:

- $a_{ik} = \dfrac{\sigma_i \cdot b_{ik}^{R}}{\sigma_k}$ (4.14)

- $f_{kj} = \sigma_k \cdot g_{kj}^{R} + < f_k >$ (4.15)

Determinare $a_{ik}$ e $f_{kj}$ non è, in generale, possibile perché il valor medio $< f_k >$ e la varianza $\sigma_k$ delle componenti non sono note a priori. Tuttavia, osserviamo che l'equazione (4.10) può essere riscritta come:

$$x_{ij} \cong \sum_{k=1}^{p} (a_{ik} \cdot \sigma_k) \cdot \frac{f_{kj}}{\sigma_k} \equiv \sum_{k=1}^{p} a_{ik}{}^{*} \cdot f_{kj}{}^{*} \qquad (X \cong A^{*} \cdot F^{*}) \tag{4.16}$$

Si può dimostrare [Swie96] $A^{*}$ e $F^{*}$ possono sempre essere espresse in termini di quantità note e che quindi è sempre possibile calcolarle:

- $A^{*} = A \cdot S_F = S_X \cdot B^{R}$ (4.17)

- $F^{*} = S^{-1}{}_F \cdot F = U^{T} \cdot \Lambda^{-1} \cdot B^{T} \cdot S_X \cdot X$ (4.18)

dove:

- $S_F$ = diag ($\sigma_k$) $\rightarrow S_F{}^{-1}$ = diag ($1/\sigma_k$);
- $S_X$ = diag ($\sigma_i$) $\rightarrow S_X{}^{-1}$ = diag ($1/\sigma_i$);
- $\Lambda$ = diag ($\lambda_k$) $\rightarrow \Lambda^{-1}$ = diag ($1/\lambda_k$);

- $B$ = matrice dei profili prima della rotazione VARIMAX.

Le matrici $A^*$ e $F^*$ forniscono tre informazioni utili:

- Gli andamenti temporali di ciascuna sorgente a meno di un fattore moltiplicativo:

$$f^*_{kj} = f_{kj}/\sigma_k \tag{4.19}$$

- I rapporti fra i vari elementi all'interno di ciascuna sorgente:

$$\frac{a^*_{ik}}{a^*_{ik'}} = \frac{a_{ik}}{a_{ik'}} \tag{4.20}$$

- Le $c_{ijk} \equiv a^*_{ik} \cdot f^*_{kj}$ che rappresentano le concentrazioni dell' $i$-esimo elemento nel giorno $j$ dovuta alla sorgente $k$.

Il passaggio da $A^*$ e $F^*$ ad $A$ e $F$ è possibile solo se conosciamo la massa del particolato campionato: infatti, se introduciamo nell'analisi statistica insieme agli elementi anche la massa del particolato, i suoi *factor loadings* ($b^R_{PM,k}$) permettono di ottenere la varianza $\sigma_k$ dei fattori:

$$1 = a_{PM,k} = \frac{\sigma_{PM} \cdot b^R_{PM,k}}{\sigma_k} \quad \Rightarrow \quad \sigma_k = \sigma_{PM} \cdot b^R_{PM,k} \tag{4.21}$$

ricordando che $a_{ik}$ rappresenta la frazione dell'$i$-esimo elemento nel particolato prodotto dalla sorgente $k$ (quindi $a_{PM,k} = 1$).

Un altro metodo è dovuto a Thurstone [Thur85]: questo metodo calcola la matrice $F^*$ aggirando il problema del calcolo matriciale (4.18) con un artificio matematico; infatti,

se introduciamo fra i dati da analizzare un giorno virtuale (indicato col pedice $j = 0$) in cui tutte le concentrazioni sono nulle ($x_{i0} = f_{i0} = 0$), otteniamo dalle equazioni (4.12):

$$g_{k0}^{R} = -\frac{< f_k >}{\sigma_k} \;\Rightarrow\; f_{ki}^{*} = f_{kj} \cdot \sigma_k^{-1} = g_{kj}^{R} - g_{k0}^{R} \tag{4.22}$$

In base al metodo di Thurstone, la matrice $F$ viene calcolata tramite una regressione lineare multipla in cui la massa totale di $PM_{10}$ è utilizzata come variabile dipendente e le $f_k^*$ come variabili indipendenti:

$$x_{PM_{10},j} = c + \sum\nolimits_k c_k \, f_{kj}^{*} \tag{4.23a}$$

$$\Rightarrow \quad f_{kj} = c_k \, f_{kj}^{*} \tag{4.23b}$$

Si possono poi ottenere i profili delle sorgenti e i contributi ($\mu g/m^3$) di ogni sorgente alle concentrazioni degli elementi misurati applicando $n$ regressioni lineari multiple (una per ogni elemento misurato) tra le concentrazioni degli elementi (variabile dipendente) e i pesi delle sorgenti (variabili indipendenti):

$$x_{ij} = a_{i0} + \sum\nolimits_k a_{ik} \, f_{kj} \tag{4.24a}$$

$$c_{ikj} = a_{ik} \cdot f_{kj} \tag{4.24b}$$