Pollution sources detection via principal component analysis and rotation

Marie Chavent, Hervé Guegan, Vanessa Kuentz, Brigitte Patouille, Jérôme Saracco

Pollution sources detection via principal component analysis and rotation[*]

Marie Chavent[1], Hervé Guégan[2], Vanessa Kuentz[1], Brigitte Patouille[1], and
Jérôme Saracco[1,3]

[1] Institut de Mathématiques de Bordeaux (FR CNRS 2254)
Université Bordeaux 1
351 Cours de la libération
33405 Talence, France
(e-mail: `chavent@math.u-bordeaux1.fr`)

[2] ARCANE-CENBG
Le Haut Vigneau
BP 120, 33175 Gradignan Cedex, France
(e-mail: `arcane@cenbg.in2p3.fr`)

[3] GREThA,
Université Montesquieu - Bordeaux IV
Avenue Léon Duguit
33608 PESSAC Cedex
(e-mail: `Jerome.Saracco@u-bordeaux4.fr`)

Abstract. Air pollution is a widely preoccupation which needs the development of control strategies. To reach this goal, pollution sources have to be precisely identified. Principal component analysis is a possible response to this problem. Indeed this factorial method enables to detect sources, that is to have a qualitative description of them. In this work, techniques of rotation are a useful help for the association of variables with factors. We highlight the fact that the rotation must be applied to the standardized principal components, so as to keep good interpretation properties. This methodology has then been applied to a problem of air pollution on a french site.
Keywords: Factor Analysis, Rotation, Pollution data.

1   Introduction

It is of great importance to identify air pollution sources in the development of air quality control strategies. Receptor modeling, using measurements of aerosol chemical composition at a sample site, is often a reliable way to provide information regarding source characteristics [Hopke, 1991]. Some multivariate receptor models are based on the analysis of the correlations between measured concentrations of chemical species, assuming that highly correlated compounds come from the same source.

---

[*] Proceedings of the XIIth International Symposium on Applied Stochastic Model and Data Analysis, Chania (2007)

One commonly used multivariate receptor model is Principal Component Analysis (PCA) [Jolliffe, 2002]. PCA extracts the principal components accounting for the majority of variance of the data that are then qualitatively interpreted as possible sources. In a second step a rotation can be used to facilitate the qualitative interpretation of the principal components. However, to keep good interpretation properties, the rotation must not directly be applied to the principal components. In Factor Analysis techniques, rotations are usually used and well defined. We will show that the rotation has to be applied to the standardized principal components.

We apply this methodology to air pollution sources detection. We present some results obtained in the framework of the project PRIMEQUAL (Projet de Recherche Interorganisme pour une MEilleure QUalité de l'Air à l'échelle Locale) of the french Ministry of Ecology. The following three steps process has been implemented: collecting PM2.5 (particles that are 2.5 microns or less in diameter, also called fine particles) with sequential fine particle samplers on a french urban site, measuring of the chemical composition with PIXE (Particle Induced X-ray Emission) method, and finally applying a PCA with rotation to identify the sources.

2   PCA and Factor Analysis

Notations. We consider a numerical data matrix $X = (x_i^j)_{n,p}$ where n objects are described on $p < n$ variables $x^1, \ldots, x^p$. Let $\tilde{X} = (\tilde{x}_i^j)_{n,p}$ be the standardized data matrix: $\tilde{x}_i^j = \frac{x_i^j - \bar{x}^j}{s^j}$ with $\bar{x}^j$ and s^j the sample mean and the sample standard deviation of x^j.

Let R be the sample correlation matrix of $x^1, \ldots, x^p$: $R = \tilde{X}'M\tilde{X}$ where $M = \frac{1}{m}I_n$ with $m = n$ or $n-1$ depending on the choice of the denominator of s^j. The correlation matrix can also be written $R = Z'Z$ with $Z = M^{1/2}\tilde{X}$.

Let us denote by $r \leq p$ the rank of Z and consider the singular value decomposition of Z:

$$Z = U\Lambda^{1/2}V' \tag{1}$$

where:

- Λ is the (r, r) diagonal matrix of the r nonnull eigenvalues λ_k, $k = 1, ..., r$ of the matrix $Z'Z$ (or ZZ'), ordered from largest to smallest.
- U is the (n, r) orthonormal matrix of the r eigenvectors u^k, $k = 1, ..., r$ of ZZ' associated with the first r eigenvalues.
- V is the (p, r) orthonormal matrix of the r eigenvectors v^k, $k = 1, ..., r$ of $Z'Z = R$ associated with the first r eigenvalues.

From the singular value decomposition of Z, we deduce the following decomposition of $\tilde{X}$:

$$\tilde{X} = M^{-1/2}U\Lambda^{1/2}V' \tag{2}$$

An overview of PCA and Factor Analysis. Let $q \leq r$. PCA and Factor Analysis operate by writing the standardized matrix $\tilde{X}$ as:

$$\tilde{X} = G_q B'_q + E_q, \tag{3}$$

where G_q is a (n,q) matrix corresponding to the factors or principal components, whereas the (p,q) matrix B_q provides information that relates the components to the original variables $x^1, \ldots, x^p$. The (n,p) matrix E_q is the rest of the approximation of $\tilde{X}$ by $\hat{\tilde{X}}_q = G_q B'_q$. Note that if $q = r$, we have $\hat{\tilde{X}}_q = \tilde{X}$ and then $E_q = 0$.

- In PCA, when $q = r$, equation (2) is written:

$$\tilde{X} = \Psi V' \tag{4}$$

  with $\Psi = M^{-1/2} U \Lambda^{1/2}$.
  The columns of Ψ, called the principal component scores matrix, are the r principal components $\psi^k = \sqrt{m}\sqrt{\lambda_k} u^k$, $k = 1, \ldots, r$. Since U and V are orthonormal, we have $\psi^k = \tilde{X} v^k$ for $k = 1, \ldots, r$ and $var(\psi^k) = \lambda_k$. In other terms, the kth principal component ψ^k is a linear combination of the p columns of $\tilde{X}$. The coefficients of v^k are called the principal component scoring coefficients.
  In practice, the user retains only the first $q < r$ eigenvalues of Λ, and the corresponding approximation of $\tilde{X}$ is then:

$$\hat{\tilde{X}}_q = \Psi_q V'_q$$

  where Ψ_q and V_q are the matrices Ψ and V reduced to their first q columns.

- In Factor Analysis (with the PCA estimation method), when $q = r$, equation (2) is written:

$$\tilde{X} = FA' \tag{5}$$

  with $F = M^{-1/2} U$ and $A = V \Lambda^{1/2}$.
  The columns of F, called the factor scores matrix, are the r factors $f^k = \sqrt{m} u^k$, $k = 1, \ldots, r$. The matrix $A = (a_j^k)_{p,r}$ is the loading matrix, also called factor pattern matrix. The coefficient a_j^k is equal to the correlation between the variable x^j and the kth factor f^k. It is also equal to the correlation between x^j and the kth principal component ψ^k.
  Since U and V are orthonormal, we have $f^k = \tilde{X} \frac{v^k}{\sqrt{\lambda_k}}$ for $k = 1, \ldots, r$ and $var(f^k) = 1$. Note that the kth factor f^k is a linear combination of the p columns of $\tilde{X}$. The coefficients of $\frac{v^k}{\sqrt{\lambda_k}}$ are called the factors scoring coefficients.
  Moreover, one can observe that $f^k = \frac{\psi^k}{\sqrt{\lambda_k}}$. Then the factors f^k can also be seen as the standardized principal components, and the factor

scoring coefficients are the standardized principal components scoring coefficients.

When the user retains only the first $q < r$ eigenvalues of Λ, the corresponding approximation of $\tilde{X}$ is then:

$$\hat{\tilde{X}}_q = F_q A'_q$$

where F_q and A_q are the matrices F and A reduced to their first q columns.

3   On the good use of rotation

Let T be an orthogonal transformation matrix, $TT' = T'T = I_q$, corresponding to an orthogonal rotation of the q axes in a p-dimensional space.

Applying directly this orthogonal transformation to the principal components obtained by PCA gives:

$$\hat{\tilde{X}}_q = \Psi_q T (V_q T)'.$$

The q rotated principal components are the q columns of the matrix $\breve{\Psi}_q = \Psi_q T$. These rotated principal components are no more mutually orthogonal and then they are no longer principal components. For this reason, the rotation must be applied to the standardized principal components:

$$\hat{\tilde{X}}_q = F_q T (A_q T)'.$$

The q rotated standardized principal components (factors) are the q columns of the matrix $\breve{F}_q = F_q T$. The rotated standardized principal components have the property to be mutually orthogonal and of variance equal to 1.

In order to be able to interpret the q rotated factors, it is important to remark that the coefficients $\breve{a}_j^k$ of the matrix $\breve{A}_q = A_q T$ are the correlations between the rotated factors $\breve{f}_q^{\,k}$ and the variables x^j.

From a practical point of view, the orthogonal transformation matrix T is then defined in order to construct a matrix $\breve{A}_q$ such that each variable x^j is clearly correlated to one of the rotated factor $\breve{f}_q^{\,k}$ (that is $\breve{a}_j^k$ close to 1) and then not correlated to the others rotated factors (that is $\breve{a}_j^{k*}$ close to 0 for $k* \neq k$). The most popular rotation technique is varimax. It seeks rotated loadings that maximize the variance of the squared loadings in each column of $\breve{A}_q$.

4   Application to air pollution sources detection

In air pollution receptor modeling, the (n, p) data matrix X consists in the measurements of p chemical species in n samples of fine particulates. In this

application, $n = 61$ samples of PM2.5 have been collected with sequential fine particle samplers by $AIRAQ^1$ in the urban french site of Anglet, every twelve hours, in december 2005. The concentrations in $ng\ m^{-3}$ of $p = 16$ chemical compounds (Al2O3, Si02, P, S04, Cl, K, Ca, Ti, Mn, Fe2O3, Ni, Cu, Zn, Br, Pb, C-Org) have been measured with the PIXE method by ARCANE-CENBG[2]. The coefficient x_i^j is then the concentration of the jth chemical compound in the ith sample.

In order to identify the sources of fine particulate emission in the samples, we have applied a PCA to the concentration matrix X, followed by an orthogonal rotation. We have then associated groups of correlated chemical compounds to air pollution sources. We give here some results obtained with this methodology .

The loading matrix $\breve{A}_5$ obtained after a varimax rotation of the standardized principal components (factors estimated by PCA in the Factor Analysis model) of the matrix X is given in Table 1.

	$\breve{f}_5^{\ 1}$	$\breve{f}_5^{\ 2}$	$\breve{f}_5^{\ 3}$	$\breve{f}_5^{\ 4}$	$\breve{f}_5^{\ 5}$
Al2O3	0.981	0.087	-0.042	0.070	-0.038
SiO2	0.979	0.012	-0.055	0.104	-0.074
P	0.972	0.090	-0.017	0.071	-0.092
SO4	-0.028	0.765	0.247	0.180	-0.345
Cl	-0.153	-0.274	-0.136	-0.181	**0.879**
K	0.597	0.716	0.111	0.233	0.031
Ca	0.608	0.091	-0.113	0.560	0.272
Mn	-0.279	0.119	0.604	0.582	-0.238
Fe2O3	0.198	0.282	0.289	0.848	-0.112
Cu	0.213	0.359	0.161	0.816	-0.149
Zn	-0.029	0.053	**0.977**	0.129	-0.044
Br	0.490	0.615	0.097	0.281	0.392
Pb	0.004	0.163	**0.969**	0.126	-0.054
C-Org	-0.018	0.893	0.021	0.222	-0.160

Table 1. Correlations between the chemical compounds and the rotated standardized principal components (factors).

The loading matrix $\breve{A}_5$ can be used to associate, if possible, sources to the rotated factors. Indeed we observe for each factor the strongly correlated coumpounds. For instance Zn and Pb are strongly correlated to $\breve{f}_5^{\ 3}$. Because Zn and Pb are known to have industrial origin, this rotated factor is associated to the industrial pollution source. In the same way the element

---

[1] Réseau de surveillance de la qualité de l'air en Aquitaine
[2] Atelier Régional de Caractérisation par Analyse Nuclaire Elémentaire - Centre d'Etudes Nucléaires de Bordeaux Gradignan

Cl is strongly correlated to $\breve{f}_5^{\;5}$, which is then associated with sea salt pollution. Possible associations between the five rotated factors and five pollution sources are given in Table 2.

Factor1	Soil dust
Factor2	Combustion
Factor3	Industry
Factor4	Vehicle
Factor5	Sea

Table 2. Factor-source associations

In order to confirm these associations we have confronted the rotated factors with external parameters such as meteorological data (temperatures and wind directions) and the periodicity night/day of the sampling. The rotated factors are the columns of the matrix $\breve{F}_5$. The coefficient $\breve{f}_i^{k}$ of the matrix $\breve{F}_5$ represents a "relative" contribution of the source k to the sample i. Fig. 1 gives for instance the evolution of the relative contribution of the source associated with $\breve{f}_5^{\;4}$. The night samples have been distinguished from the day ones, which enables to notice that the contribution of this source is stronger during the day than at night. It is then a confirmation that this source corresponds to vehicle pollution.
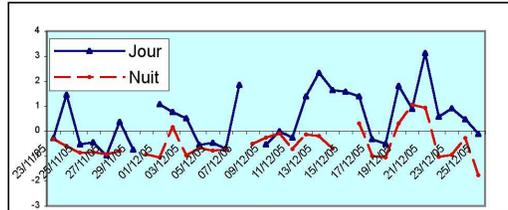
Fig. 1. Evolution of the Factor4 associated to cars pollution

In the same way, Fig. 2 gives the evolution of the relative contribution of the source associated with $\breve{f}_5^{\;2}$. We notice an increase in the contribution of this source at the middle of the sampling period, which corresponds to a decrease in the temperature measured on the sampling site (see Fig. 3). This is a confirmation that this source corresponds to combustion and heatings pollution.

To conclude we can say that the identification of the sources by PCA is only a first step of a more difficult work which consists in quantifying the
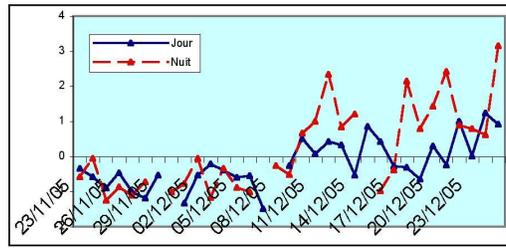
Fig. 2. Evolution of the Factor2 associated to heatings pollution
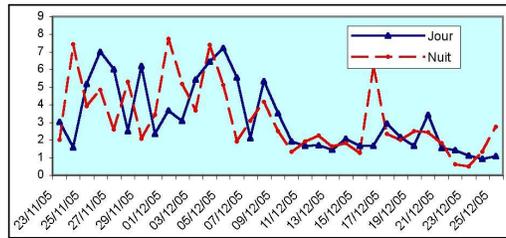

Fig. 3. Evolution of temperatures

sources. Although, it is important to discover the sources, the real problem is to define, in percentage of total fine dust mass, the quantity of each source.

References

[Hopke, 1991]P.K. Hopke. *Receptor Modeling for Air Quality Management.* Elsevier, Amsterdam, 1991.
[Jolliffe, 2002]I.T. Jolliffe. *Principal Component Analysis.* Springer Verlag, New York, 2002.