Pollution Sources Detection via Principal Component Analysis and Rotation

Vanessa Kuentz [1]

in collaboration with : Marie Chavent [1] Hervé Guégan[2] Brigitte Patouille [1]

Jérôme Saracco[1,3]

[1]IMB, Université Bordeaux 1, Talence, France
(e-mail: vanessa.kuentz@math.u-bordeaux1.fr)

[2]ARCANE-CENBG, Gradignan, France

[3]GREThA, Université Bordeaux 4, Pessac, France

Outline

1. Factor Analysis
   - Notations
   - Recall about PCA
   - Factor Analysis model
   - Link between FA and PCA
2. Rotation of factors
   - Motivation
   - Criterion
   - Remark on rotation in PCA
3. Application to air pollution sources detection
   - A three steps process
   - Statistical treatment
   - Some confirmation of the results

Issue

- Air pollution = small particles + liquid droplets
- High concentrations of particles = danger to human health, especially fine particles
- Development of air quality control strategies
  $\longrightarrow$ identification of pollution sources
- Receptor modeling, commonly used = Principal Component Analysis
- Part of scientific project initiated by French Ministry of Ecology

- $X = (x_i^j)_{n,p}$ numerical data matrix where n objects are described on $p < n$ variables $x^1, \ldots, x^p$
- $\tilde{X} = (\tilde{x}_i^j)_{n,p}$ standardized data matrix: $\tilde{x}_i^j = \frac{x_i^j - \bar{x}^j}{s^j}$ ($\bar{x}^j$,s^j sample mean and standard deviation of x^j)
- R sample correlation matrix of $x^1, \ldots, x^p$:

$$R = \tilde{X}' M \tilde{X} \text{ where } M = \frac{1}{m} I_n \text{ (with } m = n \text{ or } n-1)$$

We have : $R = Z'Z$ with $\boxed{Z = M^{1/2}\tilde{X}}$

- Replace p variables by $q \leq p$ uncorrelated components
- The **SVD** of Z (of rank $r \leq p$) is $\boxed{Z = U\Lambda^{1/2}V'}$ with :
  - Λ : diagonal matrix (r, r) of nonnull eigenvalues $\lambda_1, \ldots, \lambda_r$ of $Z'Z$ (in decreasing order)
  - U : orthonormal matrix (n, r) of eigenvectors of ZZ' associated with first r eigenvalues
  - V : orthonormal matrix (p, r) of eigenvectors of $Z'Z$ associated with first r eigenvalues
- Decomposition of $\tilde{X}$: $\boxed{\tilde{X} = M^{-1/2}U\Lambda^{1/2}V'}$

- $q = r$:
  In PCA, equation $\tilde{X} = M^{-1/2} U \Lambda^{1/2} V'$ is written :

$$\boxed{\tilde{X} = \Psi V'} \text{ with } \Psi = M^{-1/2} U \Lambda^{1/2}$$

  We have $\Psi = \tilde{X} V$ $(VV' = I_p)$
  Columns of Ψ : **principal components** $\psi^k, k = 1, \ldots, q$
- $q < r$:
  In practice, user retains only first $q < r$ eigenvalues of Λ
  Approximation of $\boxed{\tilde{X} : \hat{\tilde{X}}_q = \Psi_q V_q'}$

(Ψ_q, V_q : matrices Ψ and V reduced to first q columns)

- FA : underlying common factors + dimension reduction
- Model is written :

$$\underset{(p\times 1)}{\tilde{\mathbf{x}}} = \underset{(p\times q)(q\times 1)}{A\mathbf{f}} + \underset{(p\times 1)}{\mathbf{e}}$$

with :
- $\tilde{\mathbf{x}} = (\tilde{x}^1, \ldots, \tilde{x}^p)'$ random vector of $\mathbb{R}^p$
- A : loading (or pattern) matrix
- $\mathbf{f} = (f^1, \ldots, f^q)'$ random vector of q common factors
- $\mathbf{e} = (e^1, \ldots, e^p)'$ random vector of p unique factors

- $\tilde{x}^j$: linear combination of common factors + unique factor

- FA model : various estimation methods
  - Principal factor method
  - Maximum likelihood method
  - Principal component method
  - ...
- Commonly used method : PCA
  - Easy implementation
  - Reasonable results
  - Default method in statistical softwares

- $q = r$:
  In FA (with PCA estimation), equation $\tilde{X} = M^{-1/2} U \Lambda^{1/2} V'$ is written :

$$\boxed{\tilde{X} = FA'}$$

  with :

  -$F = M^{-1/2} U$: **factor scores** matrix

  -$A = V \Lambda^{1/2}$: **loading** (or pattern) matrix

  We have $F = \tilde{X} V \Lambda^{-1/2}$ $(V'V = I_q)$
  Columns of F : **common factors** $f^k, k = 1, \ldots, q$
- $q < r$:
  In practice, user retains only first $q < r$ eigenvalues of Λ
  Approximation of $\tilde{X}$: $\boxed{\hat{\tilde{X}}_q = F_q A'_q}$
  (F_q, A_q : matrices F and A reduced to first q columns)

We have :

$$\left.\begin{aligned} \text{PCA} : \Psi &= \tilde{X} V \\ \text{FA} : F &= \tilde{X} V \Lambda^{-1/2} \end{aligned}\right\} \Rightarrow \boxed{F = \Psi \Lambda^{-1/2}}$$

$\rightarrow$ factors f^k correspond to standardized principal components :

$$f^k = \frac{\psi^k}{\sqrt{\lambda_k}}, k = 1, \ldots, q$$

- Loading matrix A_q : $\boxed{a_j^k = \text{corr}(x^j, f^k)}$
  $\longrightarrow$ identification of groups of correlated elements
- **Problem :**
  Intermediate values $\longrightarrow$ difficult association variables/factors
- **Objective :**
  - Column of A_q : values close to 0 or 1
  - Row of A_q : only one value close to 1
- **Solution :**
  - Non-uniqueness of FA solution : rotation of factors
  - Orthogonal transformation matrix T ($TT' = T'T = I_q$)
  - $\breve{A}_q = A_q \times T$: correlation of variables to rotated factors $\breve{f}^k$
  - How to determine T?

- Several criteria : most used is **varimax**
- Maximizes empirical variance of squared elements of each column of $\breve{A}_q = (\breve{a}_j^\alpha)_{p,q}$:

$$\sum_{\alpha=1}^{q} \left\{ \frac{\sum_{j=1}^{p} (\breve{a}_j^\alpha)^4}{p} - \left(\frac{\sum_{j=1}^{p} (\breve{a}_j^\alpha)^2}{p} \right)^2 \right\}$$

with $\breve{a}_j^\alpha = a_j t^\alpha$ and u.cs. $t^\alpha (t^\alpha)' = 1, t^l (t^k)' = 0, k \neq l$.

- Possible rotation in PCA as in FA
- Must apply T to the "good" matrices $\longrightarrow$ keep property of non correlation of components
- One must not write :

$$\tilde{X} = \overbrace{M^{-1/2} U \Lambda^{1/2}}^{\Psi_q} \mathbf{TT'} V'$$

but :

$$\tilde{X} = \overbrace{M^{-1/2} U}^{F_q} \mathbf{TT'} \Lambda^{1/2} V'$$

$\longrightarrow$ rotation of standardized principal components = factors of FA

- Part of scientific project of French Ministry of Ecology
- Improve air quality : identification + quantification of pollution sources
- Three steps process :
  - **Collecting** of fine particles in French urban site (Anglet) : $n = 61$ samples (Dec. 2005)
  - **Measurements** of $p = 16$ chemical species : PIXE method (Particle Induced X-ray Emission)
  - **Statistical treatment** of $X = (x_i^j)_{n,p}$: concentration of jth chemical compound in ith sample

- FA model with PCA estimation
- Choice of number of factors : $q = 5$
- Orthogonal rotation of factors with varimax criterion
- Detection of groups of correlated chemical compounds
- Identification of air pollution sources

	f^1	f^2	f^3	f^4	f^5
Al2O3	0.67	-0.66	0.22	-0.19	0.02
SiO2	0.65	-0.67	0.25	-0.20	-0.06
P	0.68	-0.64	0.22	-0.24	0.01
SO4	0.59	0.45	-0.41	-0.22	0.18
Cl	-0.47	-0.28	0.19	0.68	0.34
K	0.89	-0.15	-0.21	0.00	0.28
Ca	0.64	-0.41	0.10	0.40	-0.19
Mn	0.38	0.78	0.18	0.09	-0.25
Fe2O3	0.79	0.32	0.04	0.27	-0.36
Cu	0.80	0.25	-0.10	0.23	-0.36
Zn	0.35	0.66	0.59	-0.07	0.25
Br	0.75	-0.18	-0.13	0.35	0.36
Pb	0.43	0.66	0.52	-0.09	0.30
C-Org	0.60	0.30	-0.61	-0.02	0.22

Table: Loading matrix before rotation A_5

Lots of intermediate values $\longrightarrow$ difficult association variables/factors

	$\breve{f}^1$	$\breve{f}^2$	$\breve{f}^3$	$\breve{f}^4$	$\breve{f}^5$
Al2O3	**0.98**	0.09	-0.04	0.07	-0.04
SiO2	**0.98**	0.01	-0.06	0.10	-0.07
P	**0.97**	0.09	-0.02	0.07	-0.09
SO4	-0.03	**0.77**	0.25	0.18	-0.35
Cl	-0.15	-0.27	-0.14	-0.18	**0.88**
K	0.60	0.72	0.11	0.23	0.03
Ca	0.61	0.09	-0.11	0.56	0.27
Mn	-0.28	0.12	0.60	0.58	-0.24
Fe2O3	0.20	0.28	0.29	**0.85**	-0.11
Cu	0.21	0.36	0.16	**0.82**	-0.15
Zn	-0.03	0.05	**0.98**	0.13	-0.04
Br	0.49	0.62	0.10	0.28	0.39
Pb	0.00	0.16	**0.97**	0.13	-0.05
C-Org	-0.02	**0.89**	0.02	0.22	-0.16

Table: Rotated loading matrix $\breve{A}_5$

Factor	Possible source
Factor1	Soil dust
Factor2	Combustion
Factor3	Industry
Factor4	Vehicle
Factor5	Sea

Table: Pollution sources

Obvious usefulness of rotation : clear associations

ex: Zn + Pb highly correlated to $\breve{f}_3 \longrightarrow$ industrial pollution source

- Rotated factors : columns of $\breve{F}_5 = F_5 \times T = (\breve{f}_i^k)_{n,5}$
- Score $\breve{f}_i^k$ = "relative" contribution of kth source to ith sample
- Confrontation of rotated factors with external parameters :
  - Meteorological data (wind, temperature, ...)
  - Periodicity day/night
  - Other chemical coumpounds

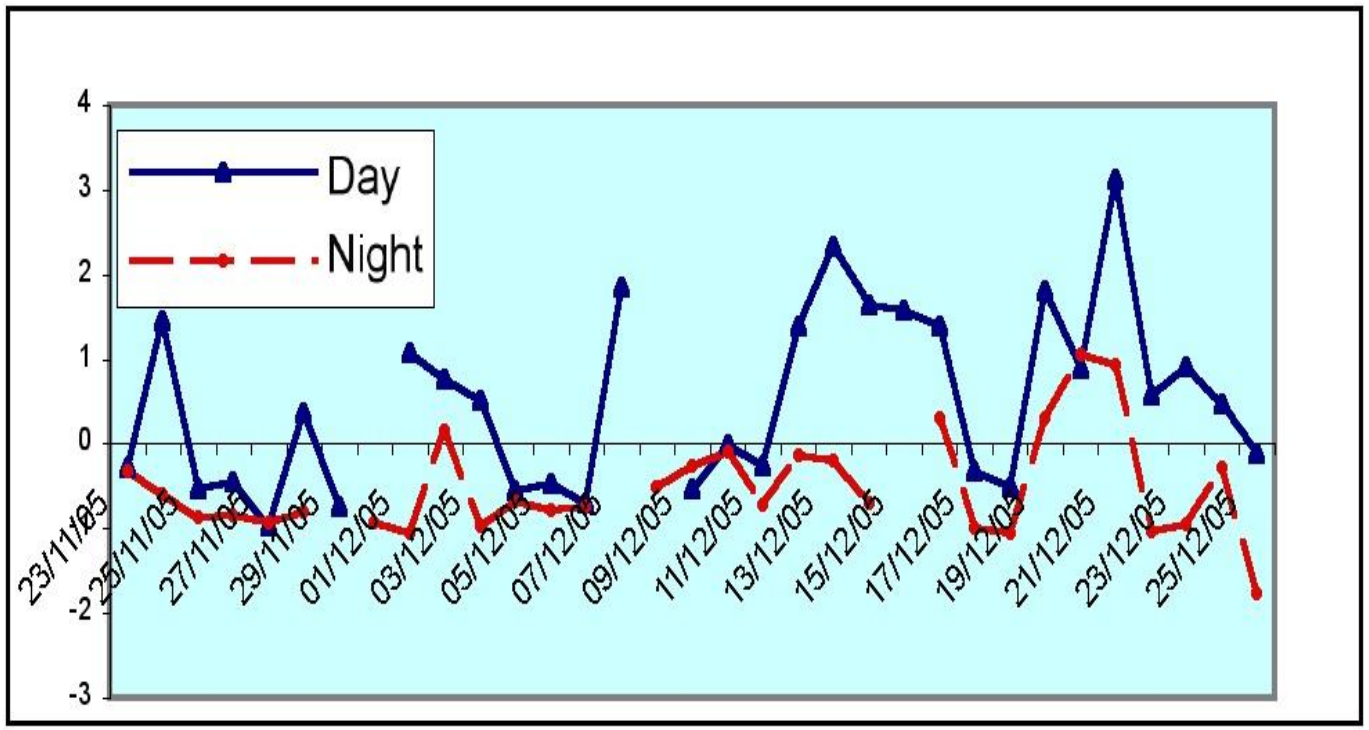


Figure: Evolution of vehicle pollution source

Stronger contribution during day than night → confirmation of cars pollution source

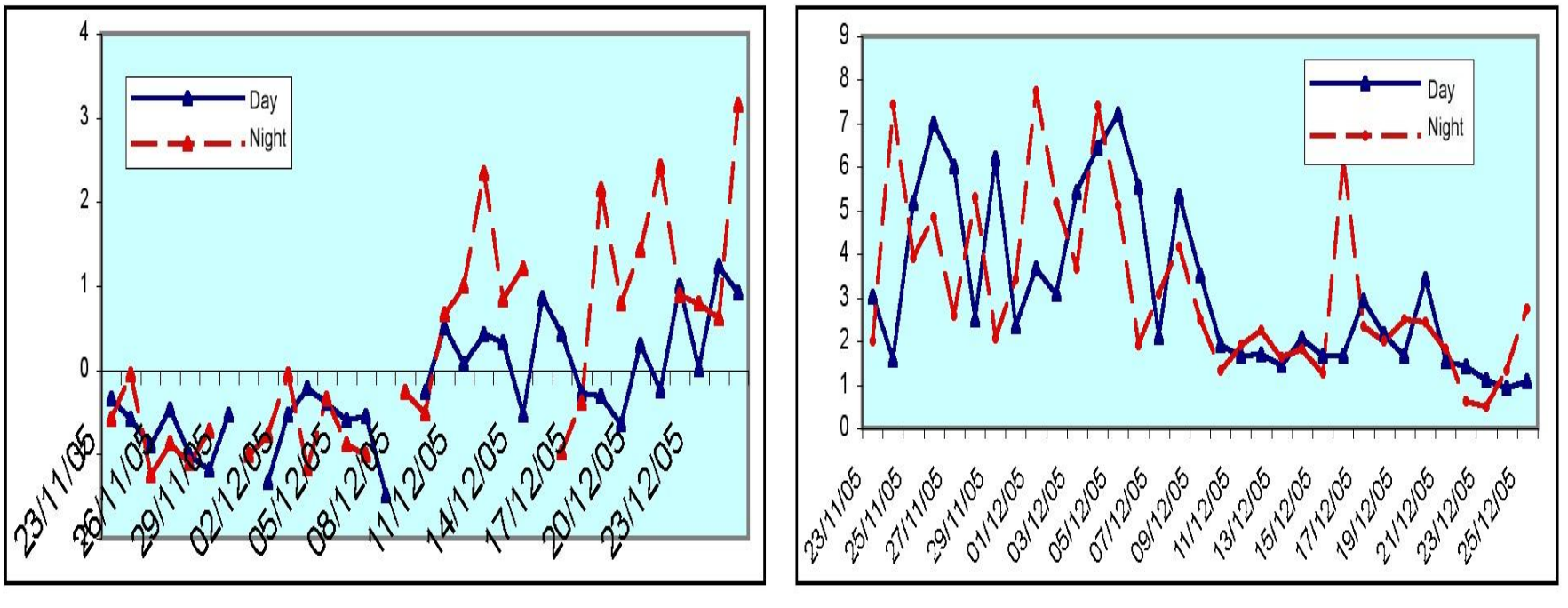
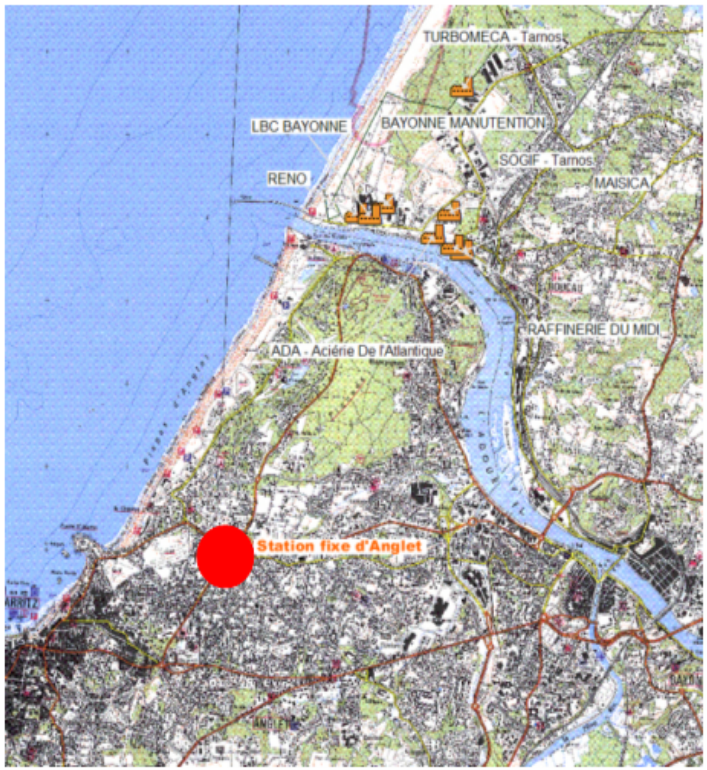


Figure: Evolution of heating pollution and temperatures

Middle of period :

Increase in contribution
Decrease in temperature
} → confirmation of heating pollution



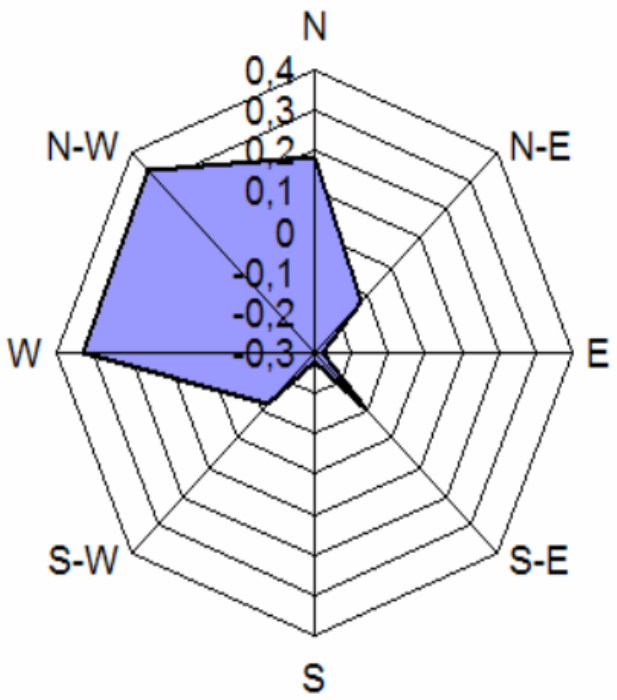


Figure: Map of sampling site and correlations wind/sea pollution

Strong correlation to N-W wind
Location toward Atlantic Ocean
} → validation of sea pollution

Conclusion

- FA (with PCA estimation) followed by rotation : identification of five sources of particulate emission
- No prior knowledge : number, composition ?
  $\longrightarrow$ more complex sampling site
- Sources quantification : percentage of total fine dust mass of each source (JDS, Angers, 2007)