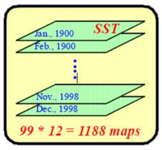
Methods in climatology

II. Multivariate analysis

Multivariate analysis

- **Large datasets**
- **Redundant information**
- **Stochastic character of processes**
- **Signal vs noise**



Main aim:

- to separate climate signal from the background climate variability (noise) and
- to identify physical processes responsible for the generation of the signal

Analysis examples

- **Identification of climate modes (NAO)**
- **Climate zones definition (on different scales)**
- **Statistical downscalling (regional climate vs large-scale atmospheric circulation.**

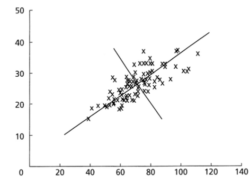
Multivariate analysis

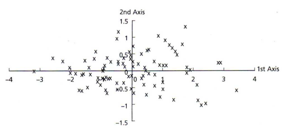
- **Abiltiy to represent spatio-temporal data in a compresed way**

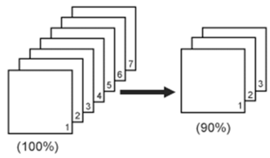
Four main goals of MA in climate research:

- **to recognize the patterns of climate vartiability**
- **to identify physical processes and use them to construct CM**
- **to validate climate models with observations**
- **to use signals for predictions**

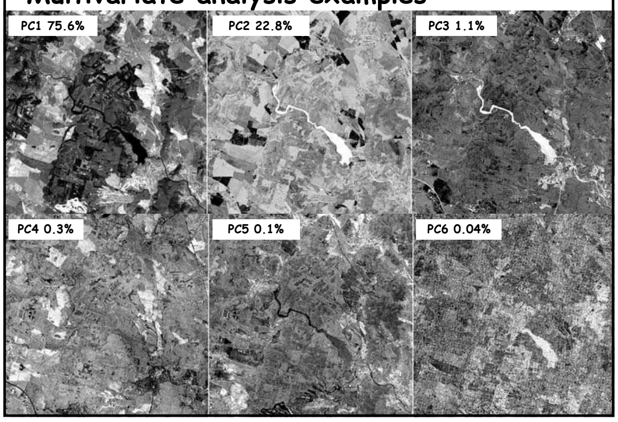
Principal Component Analysis (PCA)







Multivariate analysis examples



Multivariate analysis examples

$$PC_1 = a_1TM_1 + a_2TM_2 + a_3TM_3 + a_4TM_4 + a_5TM_5 + a_6TM_7$$
$$PC_2 = b_1TM_1 + b_2TM_2 + b_3TM_3 + b_4TM_4 + b_5TM_5 + b_6TM_7$$
...

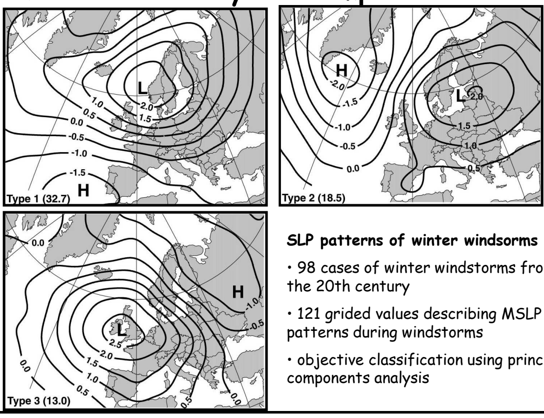
Číslo	Vlastní	Procenta	Kumulov.	Zátěže					
PC	čísla	rozptylu	procenta	TM 1	TM 2	TM 3	TM 4	TM 5	TM 7
1	2262,96	75,62	75,62	0,243	0,181	0,346	0,230	0,728	0,454
2	682,34	22,80	98,42	0,115	0,050	0,229	-0,936	-0,012	0,237
3	33,80	1,13	99,55	0,553	0,323	0,513	0,201	-0,531	-0,064
4	7,79	0,26	99,81	-0,264	-0,141	-0,037	0,168	-0,432	0,833
5	4,54	0,15	99,96	0,712	-0,102	-0,668	-0,034	0,000	0,186
6	1,21	0,04	100,00	-0,212	0,911	-0,343	-0,044	-0,022	0,069

eigenvalue = vlastní číslo

eigenvector = vlastní vektor

zátěž = loading

Multivariate analysis examples

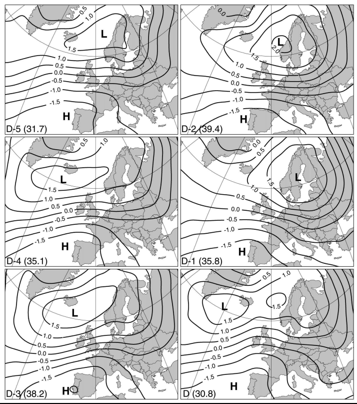


SLP patterns of winter windsorms

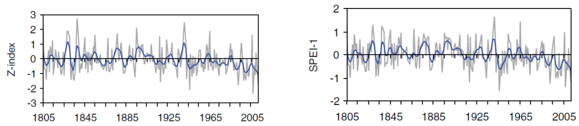
- 98 cases of winter windstorms from the 20th century
- 121 grided values describing MSLP patterns during windstorms
- objective classification using principal components analysis

Multivariate analysis examples

Component scores of the first component of the SLP field on days D-5 to D calculated by the Principal Component Analysis Method (PCA) for 37 floods (1881-2000) of the winter synoptic type on the river Vltava in Prague (brackets - explained variance in %)

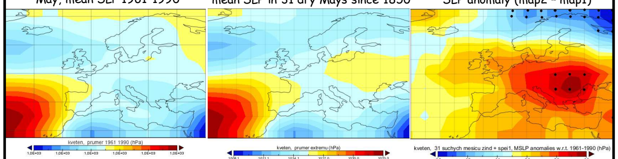


Multivariate analysis examples



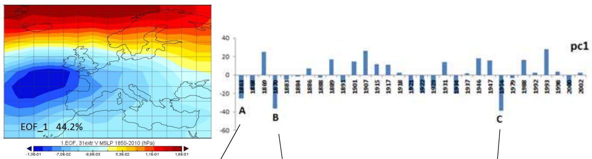
Fluctuations of MAM drought indices (Z-index, SPEI-1) for the Czech Lands in the 1805–2012 period.

May, mean SLP 1961–1990 | mean SLP in 31 dry Mays since 1850 | SLP anomaly (map2 – map1)

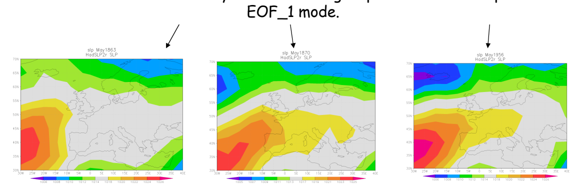


SLP at extremely dry Mays defined in the 1850–2010 period and SLP differences in extremely dry months compared to MSLP of the 1961–1990 reference period; black points mark statistically significant (α=0.05) SLP decrease/incerese w.r.t the 1961–1990 reference period

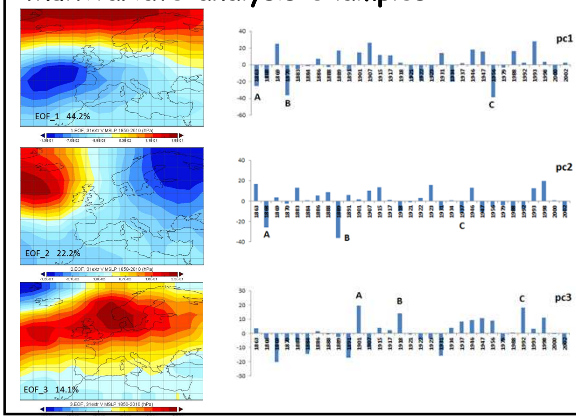
Multivariate analysis examples



Three the most similar monthly means acording to pc1 used for interpretation of EOF_1 mode.

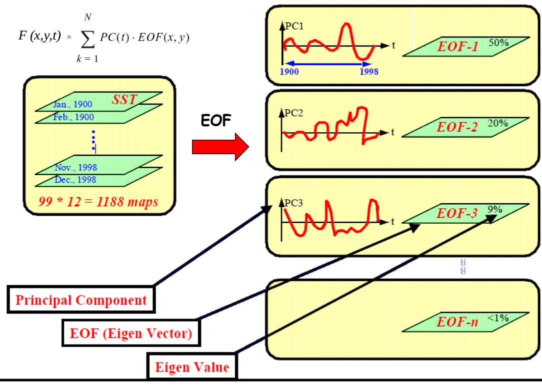


Multivariate analysis examples



Patterns in Space: EOF analysis (Empirical Orthogonal Functions)

$$F(x,y,t) = \sum_{k=1}^{N} PC(t) \cdot EOF(x,y)$$



Patterns in Space: EOF analysis (Empirical Orthogonal Functions)

- DATA: Instantaneous samples (maps) of geophysical fields (air temperature) defined in a number of points (stations or grid-points) recorded over period of time
- EOF (PCA – Principal Component Analysis) – technique for compressing the variability in the data set
- Introduced by Edward Lorenz in 1956
- Widely applied in climatology and oceanography
- Goal: compact description of the spatial and temporal variability of data series in terms of orthogonal functions – statistical „modes“
- Most of variability is in the first few orthogonal functions whose patterns MAY BE be linked to possible dynamical mechanisms

EOF – data preparation

- *A set of N maps at times $t = 1 \ldots N$*
- Each map contains measurements of the field ψ at locations *$m = 1 \ldots M$*
- We have M time series $\psi_m(t)$, each of length N
- We assume that N > M (number of time steps is larger than the number of locations
- Annual (seasonal) cycle is necessary to remove BEFORE EOF analysis – subtract climatological cycle from the field $\psi_m(t)$.

EOF – data preparation

- Data standardization:

$$F_m(t) = \frac{\psi_m(t) - \mu_m}{\sigma_m}$$

where μ_m is the record mean:

$$\mu_m = \frac{1}{N} \sum_{t=1}^{N} \psi_m(t)$$

and σ_m is the record standard deviation:

$$\sigma_m = \left[\frac{1}{N-1} \sum_{t=1}^{N} \psi_m^2(t) \right]^{1/2}$$

EOF – data preparation

We construct M x N data matrix F with M rows (locations m) nad the N columns (times t):

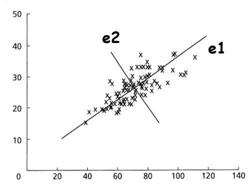
$$\begin{array}{c} \text{Time} \longrightarrow \\ \mathbf{F} = \begin{bmatrix} F_1(1) & F_1(2) & \ldots & F_1(N) \\ F_2(1) & F_2(2) & \ldots & F_2(N) \\ \ldots & \ldots & \ldots & \ldots \\ F_M(1) & F_M(2) & \ldots & F_M(N) \end{bmatrix} \Big\downarrow \text{Location} \end{array}$$

Two approaches for EOFs computing

- *Covariance matrix decomposition to eigenvalues and eigenvectors (rozklad kovariančni matice na vlastní čísla a vlastní vektory)*
- *Singular Value Decomposition of the data matrix (singulární rozklad matice)*

Vlastní čísla a vlastní vektory matice

- Existují pro čtvercové matice, které neobsahují lineárně závislé proměnné
- Vlastní čísla informují o variabilitě vyčerpané vytvářenými faktorovými osami
- Vlastní číslo představuje rozptyl „nové“ proměnné definované v souřadném systém vlastních vektorů
- Vlastní vektory definují směr nových faktorových os v prostoru původních proměnných
- Vlastní vektory jsou navzájem ortogonální – tj. nezávislé – tedy každý nese unikátní informaci
- Vlastní vektory mohou být různým způsobem standardizovány a jejich interpretace se liší podle použité standardizace



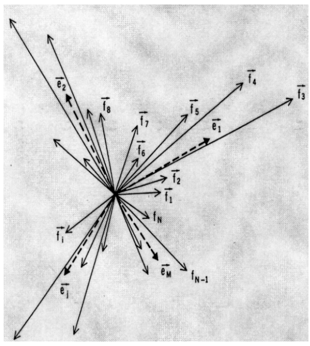
Geometric interpretation of eigenvalues and eigenvectors

An n × n matrix **A** multiplied by n × 1 vector **x** results in another n × 1 vector **y**=**Ax**. Thus **A** can be considered as a transformation matrix.

In general, a matrix acts on a vector by changing both its magnitude and its direction. However, a matrix may act on certain vectors by changing only their magnitude, and leaving their direction unchanged (or possibly reversing it). These vectors are the **eigenvectors** of the matrix.

A matrix acts on an eigenvector by multiplying its magnitude by a factor, which is positive if its direction is unchanged and negative if its direction is reversed. This factor is the **eigenvalue** associated with that eigenvector.

Geometric interpretation of eigenvalues and eigenvectors



Possible configuration of the data vectors f_n (*n = 1 … N* denote the time steps) and the empirical orthogonal vectors e_m, m = 1 … M. (from Peixoto and Oort, 1992)

EOF – The Covariance Matrix Approach

Data matrix **F** is used to derive spatial covariance matrix $\mathbf{R_{FF}}$ of the field $F_m(t)$ by multiplying **F** by its transpose $\mathbf{F}^\dagger$:

$$\mathbf{R_{FF}} = \mathbf{F} * \mathbf{F}^\dagger \tag{2.5}$$

Expanding the product of matrices:

$$\mathbf{R_{FF}} = \begin{bmatrix} \langle F_1F_1\rangle & \langle F_1F_2\rangle & \dots & \langle F_1F_M\rangle \\ \langle F_2F_1\rangle & \langle F_2F_2\rangle & \dots & \langle F_2F_M\rangle \\ \dots & \dots & \dots & \dots \\ \langle F_MF_1\rangle & \langle F_MF_2\rangle & \dots & \langle F_MF_M\rangle \end{bmatrix} \tag{2.6}$$

where $\langle F_iF_j\rangle$ is the covariance between time series F_i and F_j (F at locations i and j) defined as:

$$\langle F_iF_j\rangle \;=\; \langle F_jF_i\rangle \;=\; \frac{1}{N-1}\sum_{t=1}^{N} F_i(t)\, F_j(t) \tag{2.7}$$

i,j = 1 .. M

$\mathbf{R_{FF}}$ is square (M x M) and symetric

EOF – The Covariance Matrix Approach

We solve the eigenproblem:

$$\mathbf{R_{FF}} * \mathbf{E} = \mathbf{E} * \mathbf{\Lambda} \tag{2.8}$$

That is, we decompose $\mathbf{R_{FF}}$ into matrices $\mathbf{\Lambda}$ and $\mathbf{E}$. Here $\mathbf{\Lambda}$ is the $M \times M$ diagonal matrix containing the eigenvalues λ_k of $\mathbf{R_{FF}}$:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_M \end{bmatrix} \tag{2.9}$$

- the eigenvalues are sorted in decreasing order
- all eigenvalues are greater or equal to zero
- typically only first K eigenvalues are non-zero, K ≤ min(N,M)
- thus only K EOF modes can be determined

EOF – The Covariance Matrix Approach

We solve the eigenproblem:

$$\mathbf{R_{FF}} * \mathbf{E} = \mathbf{E} * \mathbf{\Lambda} \tag{2.8}$$

The square matrix $\mathbf{E}$ has dimension $M \times M$. Its column vectors E^k are the eigenvectors of $\mathbf{R_{FF}}$ corresponding to eigenvalues λ_k:

$$\mathbf{E} = \begin{bmatrix} E_1^1 & E_1^2 & \dots & E_1^M \\ E_2^1 & E_2^2 & \dots & E_2^M \\ \dots & \dots & \dots & \dots \\ E_M^1 & E_M^2 & \dots & E_M^M \end{bmatrix} \tag{2.10}$$

$$\begin{array}{cccc} \downarrow & \downarrow & & \downarrow \\ E^1 & E^2 & & E^M \longrightarrow \text{Eigenvectors } E^k \end{array}$$

- each non-zero eigenvalue λ_k in matrix **Λ** is associated with a column eigenvector E^k in matrix **E**.
- only K eigenvectors are used in decomposition
- K are modes of EOF decomposition

EOF – The Covariance Matrix Approach

The eigenvector matrix E has the property that $\mathbf{E} * \mathbf{E}^\dagger = \mathbf{E}^\dagger * \mathbf{E} = \mathbf{I}$, where **I** is Identity matrix.

This means that the eigenvectors are uncorrelated over space – they are orthogonal to one another

Each eigenvector Ek represents the spatial EOF pattern of mode k

The spatial EOF patterns – Loadings

EOF – The Covariance Matrix Approach

The time evolution of the kth EOF (that is, how pattern E^k evolves with time) is given by the time series $A^k(t)$, which is obtained by projecting the original data series $F_m(t)$ onto eigenvector E^k and summing over all locations m:

$$A^k(t) = \sum_{m=1}^{M} E_m^k \, F_m(t) \qquad (2.11)$$

where $m = 1 \ldots M$ counts the locations, $t = 1 \ldots N$ counts the time steps and $k = 1 \ldots K$ counts the EOF modes. In matrix notation, matrix $\mathbf{A}$ is obtained by multiplying matrices $\mathbf{E}^\dagger$ and $\mathbf{F}$:

$$\mathbf{A} = \mathbf{E}^\dagger * \mathbf{F} \qquad (2.12)$$

where $\mathbf{E}^\dagger$ is K x M, **F** is M x N, **A** is K x N

Rows in A are time series of length N – **Principal Components (Time coefficients, Scores)**

EOF – The Covariance Matrix Approach

Each eigenvalue λ_k is proportional to the percentage of the variance of the field F that is accounted for by the mode k:

$$\%\ \text{Variance Mode k} = \frac{\lambda_k}{\sum_{i=1}^{K} \lambda_i} * 100 \qquad (2.13)$$

The original field F can be totally reconstructed by multiplying each EOF pattern E^k by its corresponding principal component A^k and adding the products over all K modes:

$$F_m(t) = \sum_{k=1}^{K} E_m^k \, A^k(t) \qquad (2.14)$$

In matrix notation:

$$\mathbf{F} = \mathbf{E} * \mathbf{A} \qquad (2.15)$$

where **F** is M x N, **E** is M x K, **A** is K x N

EOF – The Covariance Matrix Approach

The goal of the EOF decomposition is reconstruction of **compressed** and **less noisy** version of the original field F

This is done by truncating the decomposition ni 2.14 eq. using only first H modes with H < K

The H first modes account for the largest fraction of the field variance:

$$\hat{F}_m(t) = \sum_{k=1}^{H} E_m^k \, A^k(t) \qquad (2.16)$$

This leads to a significant reduction of the amount of data while retaining most of the variance of the field F.

The choice of H may be rather subjective

The first or the few first EOF modes **sometimes** represent meaningful physical processes

The Singular Value Decomposition Approach

- one-step method to compute all components of eigenvalue problem
- Results are computationally more stable and robust

SVD is performed directly on the data matrix **F** with M rows (spatial points) and N columns (samples in time)

SVD is based on the concept that any **M x N** matrix can be written as the product of three matrices:

$$\mathbf{F} = \mathbf{U} * \mathbf{\Gamma} * \mathbf{V}^\dagger \qquad (2.19)$$

U is M x M matrix

$\mathbf{V}^\dagger$ is transpose of the N x N matrix **V**

Γ is M x N matrix with zero elements outside the diagonal and positive or zero elements on the diagonal

Scalars γ_k on the diagonal are called singular values. They are placed in decreasing order and they are proportional to eigenvalues λ_k $\lambda_k = \gamma_k^2$

There is a maximum of K ≤ min(N,M) non-zero singular values which defines the maximum number of EOF modes that we can determine.

The Singular Value Decomposition Approach

$$\mathbf{F} = \mathbf{U} * \mathbf{\Gamma} * \mathbf{V}^\dagger \qquad (2.19)$$

The columns in **U** matrix are orthogonal and are called *left singular vectors* of **F**

They are identical to the eigenvectors **E** and they are the EOF patterns associated with each singular value. There is only K useful *left singular vectors*

The rows in $\mathbf{V}^\dagger$ matrix are orthogonal and are called *right singular vectors* of **F**

They are proportional to the principal components A obtained from equations 2.11 and 2.12 and the constant of proportionality are the singular values γ_k such that:

$$\mathbf{A} = \mathbf{\Gamma} * \mathbf{V}^\dagger \qquad (2.20)$$

$$A^k(t) = \gamma_k \, V^{\dagger k}(t) \qquad (2.21)$$

Matrix **A** contains the principal coefficients of data matrix **F** and effective size of **A** is K x N

The Singular Value Decomposition Approach

Using equation (2.19) we can reconstruct field F adding all K modes of the decomposition:

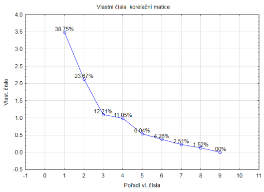
$$F_m(t) = \sum_{k=1}^{K} U_m^k \, \gamma_k \, V^{\dagger k}(t) \qquad (2.22)$$

Note the similarity between (2.22) and (2.14) where: U = E and $A = \gamma V^\dagger$

Number of important modes

- **Data compression is more important than physical interpretation**
- **We try to separate signal from noise**
- **Several methods:**
  - Scree plot
  - Guttman criterion
  - Eigenvalues > 1
  - Modes with eigenvalue that is higher than mean of all eigenvlaues
  - Modes that explain more that 70 – 90 % of total variability

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	3.487151	38.75	38.75	||||||||
2	2.130173	23.67	62.41	|||||
3	1.098958	12.21	74.63	|||
4	0.994483	11.05	85.68	|||
5	0.543218	6.04	91.71	||
6	0.383428	4.26	95.97	|
7	0.225754	2.51	98.48	|
8	0.136790	1.52	100.00	|
9	0.000046	0.00	100.00	|



Rotated EOFs

- physical interpretation is more important than data compression
- Due to orthogonality
- Orthogonal or obligue rotation
- Rules of **„simple structure“**

Eigenvalues

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	2.728683	45.48	45.48	||||||||||
2	1.128792	18.81	64.29	||||
3	0.615291	10.25	74.55	|||
4	0.602809	10.05	84.59	|||
5	0.522514	8.71	93.30	||
6	0.401910	6.70	100.00	||

Eigenvalues after Varimax Rotation

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	1.596863	56.94	56.94	||||||||||||
2	1.207981	43.08	100.02	|||||||||
3	0.050820	1.81	101.83	|
4	0.011910	0.42	102.26	|
5	-0.008657	-0.31	101.95	|
6	-0.054642	-1.95	100.00	|

Factor Loadings

Variables	Factor1	Factor2
Gaelic	-0.660803	-0.444475
English	-0.688465	-0.289771
History	-0.516356	-0.639552
Arithmetic	-0.735620	0.417018
Algebra	-0.741868	0.372759
Geometry	-0.678168	0.354100

Factor Loadings after Varimax Rotation

Variables	Factor1	Factor2
Gaelic	-0.233132	-0.659253
English	-0.322810	-0.552071
History	-0.084713	-0.589192
Arithmetic	-0.765986	-0.170657
Algebra	-0.718105	-0.214689
Geometry	-0.573340	-0.214994

Notes on EOFs interpretation

- Some EOFs not necessarily correspond to real physical behavior of dynamical modes
- A clue to the interpretation of EOF modes may be found in the principal component.
- Their temporal variability may be similar to some known processes
- The physical interpretation is limited due to spatial orthogonality of the EOF patterns
- Real world processes do not have orthogonal patterns or may not be represented with uncorrelated indices
- Traditional EOFs can detect standing oscillations, however signal may be propagating in space

Notes on EOFs interpretation

- The EOF patterns depend on the size of the study area
- Variable with uniform distribution of variance and with the spatial scale comparable (or larger) to spatial domain produce monopole EOF 1 (the same sign in all points)
- The need to be orthogonal to the first EOF creates a second EOF with dipole pattern
- Thus the size of the domain should be greater than the typical spatial scale of field analyzed

Units of presentation

- Units of field F are carried by the PCs while the EOFs are dimensionless
- It is common to re-normalize results (e.g. EOFs carry units of F and PCs have variance of 1)
- Re-normalization is SW-specific – see e.g. Climate explorer application
- EOFs can be presented as a correlation maps – correlations between principal component and the values of the field F at each location.

EOF analysis example

Main mode of SST vatiability in Central Pacific