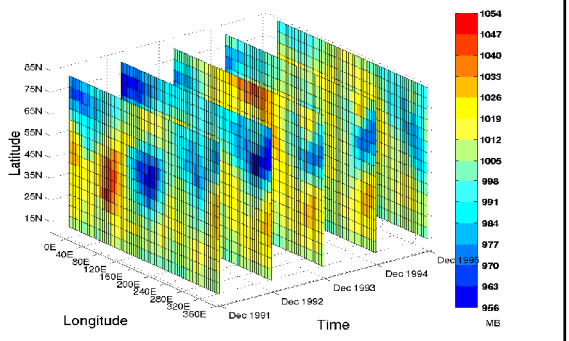
Data analysis methods in weather and climate research

Dr. David B. Stephenson
Department of Meteorology
University of Reading
www.met.rdg.ac.uk/cag

10. Spatio-temporal methods for gridded datasets

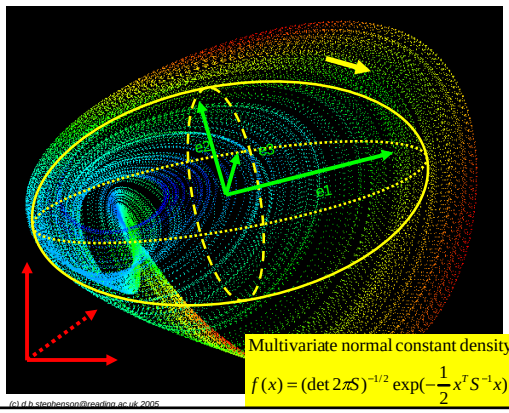
- Multivariate statistics
- EOF/PCA analysis
- Physical meaning
- Other EOF approaches

10. Spatial fields on structured grids



Key idea: treat each spatial map as a p-vector $x_t = (x_{t1}, x_{t2}, ..., x_{tp})^T$

10. Dynamical system $x_{t+1} = M(x_t; t) + \varepsilon_t$



Multivariate normal constant density surfaces

$$f(x) = (\det 2\pi S)^{-1/2} \exp(-\frac{1}{2} x^T S^{-1} x)$$

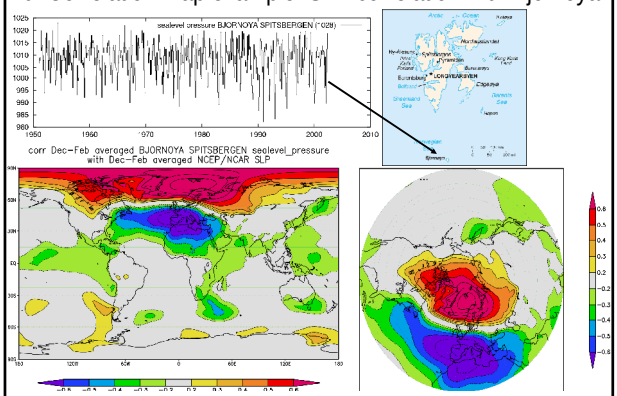
10. Multivariate statistics

- Dynamical evolution of the system described by a p-vector that hops around p-dimensional *state space*
- Store the sample of n vectors in the (n x p) data matrix **X.** Simplest *bivariate* case when p=2 variables.
- Use linear algebra to calculate sample statistics. For example, if x are centered (anomalies) then the (pxp) sample covariance and correlation matrices are given by:

$$S = \frac{1}{n} X^T X \qquad R = D^{-1/2} S D^{-1/2}$$

where $D = diag(S)$

10. Correlation map example: SLP correlation with Bjornoya



Correlation map produced using the Climate Explorer at http://climexp.knmi.nl

10. Spatial dependency: climatic *teleconnections*

A. Ångström (1935) Teleconnections of climatic changes in present time, Geogr. Annal., 17, 243-258

"the weather at a given place is not an isolated phenomenon but is intimately connected with the weather at adjacent places"

Chief causes of teleconnections:
I. Local extension of a given feature
II. Propagation of weather systems
III. Existence of changes of great extension that affects local weather:
1. Energy reaching the Earth
2. Atmospheric circulation
3. Other

FIG. 1. A portrait of Anders K. Ångström sketched by E. Liljas from a photograph taken in 1978.

10. Principal Component Analysis PCA

Problem: to find the linear combination $\tilde{x} = e^T x$ of the p – variables that has maximum variance where e is a unit p – vector ($e^T e = 1$).

Solution: $\text{var}(e^T x) = e^T S e$ where $S = X^T X / n$ is the $(p \times p)$ sample covariance matrix is maximised when e is the leading eigenvector of S.
(x is assumed to be centered about time mean).

```
>#  R command
> prcomp(X)
```


10. A brief history of PCA in climate

EOFs = principal axes of the ellipsoids in state space (spatial patterns).
Principal Components = projection of x onto EOFs (time series)

- Fukuoka (1951) – empirical forecasting
- Lorenz (1956) – empirical forecasting
- Obukhov (1960) – sampling on sphere
- Kutzbach (1967) – mixed field EOFs
- Barnett and Preisendorfer (1978) – climate prediction
- Wallace and Gutzler (1981) – teleconnections
- Horel (1981) – rotated EOFs
- Richman (1986) – rotated EOFs
- Barnston and Livezey (1987) – low-freq variability
- … lots more

10. Why do we use PCA?

- Reduce dimensionality (data reduction) because we can't show all the variables – so instead focus on ones that explain the most variance.
- Identify dominant modes (and perhaps the most predictable components)
- Avoid ill-conditioning (collinearity) caused by the dependency between grid point variables
- Useful 1st step before doing other multivariate analysis e.g. CCA, cluster analysis, discriminant analysis etc.
- Factor out spatial from temporal behaviour

10. How to do PCA: Singular Value Decomposition of X

$$X = U\Sigma V^T$$
$$U^T U = UU^T = 1_{n\times n}$$
$$V^T V = VV^T = 1_{p\times p}$$

$$X_{ts} = \sum_{k=1}^{rank} (u\sigma)_{tk} v_{sk} = \sum_{k=1}^{rank} \tilde{X}_{tk} e_{ks}$$

grid - to - PC transformation

($p\times p$ rotation followed by $p\times q$ projection)

$$X_{ts} \rightarrow \tilde{X}_{tk} = X_{tr} V_{rk}$$

```
>#  R command
> svd(X)
```

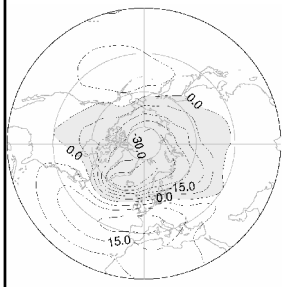

10. Modal expansion

$$X_{ts} = \sum_{k=1}^{q} \tilde{X}_{tk} e_{ks} + \varepsilon_{ts}$$

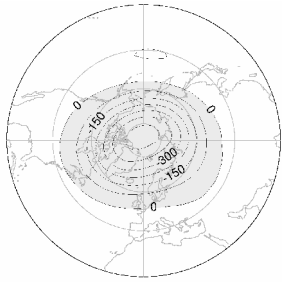
Time-varying spatial field is expressed as:
- a sum of q “modes” plus a “small” residual
- each mode is the product of a time series and a spatial pattern
- metric is total variance ➔ PCA/EOF modes
- not obvious what metric to choose for a dynamical system
  e.g. non-modal growth for non-normal systems

10. How physical is the leading NH EOF?

Arctic Oscillation — **EOF1 at 1000hPa**



Northern Annular Mode — **EOF1 at 10hPa**

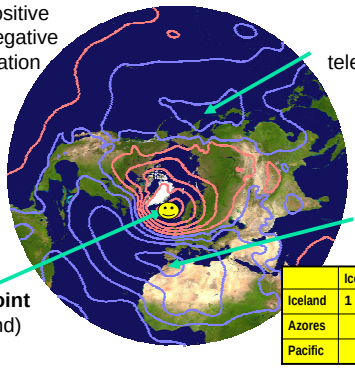


Annular mode patterns are similar from Earth’s surface to 50+km

Thompson and Wallace 1998, 2000, 2001; Baldwin and Dunkerton 2001

10. Correlation map with Iceland SLP

red: positive
blue: negative
correlation



weaker teleconnection (Pacific)

strong teleconnection (Azores)

basepoint (Iceland)

	Iceland	Azores	Pacific
Iceland	1	-0.58	-0.29
Azores		1	-0.11
Pacific			1

10. EOF analysis of Iceland, Azores, Pacific

EOF analysis of SLP at the Iceland, Azores, Pacific centres of action with sample covariance:

	Iceland	Azores	Pacific
Iceland	62.1	-23.9	-14.9
Azores	-23.9	27.0	-3.6
Pacific	-14.9	-3.6	43.8

units of hPa2

gives following EOFs (loading weights):

	Iceland	Azores	Pacific	
EOF1 59%	-0.86	0.38	0.33	Annular mode
EOF2 32%	0.15	-0.43	0.89	Azores-Aleutian
EOF3 9%	0.48	0.82	0.31	All three

10. EOF when Azores and Pacific not correlated

EOF analysis of SLP at the Iceland, Azores, Pacific centres of action with sample covariance:

	Iceland	Azores	Pacific
Iceland	62.1	-23.9	-14.9
Azores	-23.9	27.0	0.0
Pacific	-14.9	0.0	43.8

units of hPa2

gives following EOFs (loading weights):

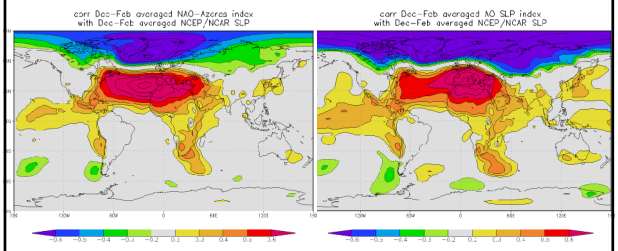
	Iceland	Azores	Pacific	
EOF1 60%	-0.85	0.21	0.48	Annular mode
EOF2 30%	0.39	-0.37	0.84	Azores-Aleutian
EOF3 10%	0.36	0.90	0.24	All three

→ Leading mode is still annular despite no Pac-Atl correlation!

10. SLP correlation maps with NAO/AO indices

Correlation of SLP with NAO

Correlation of SLP with AO



→ Not much evidence of correlation in Pacific subtropics

10. And what about these phenomena?

Prof Brian Hoskins FRS

Britney Spears

Correlation of r=0.56 (n=150x110=16500 pixels)
Highly significant (p<0.001)

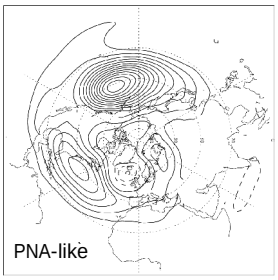
Why is the correlation so high?

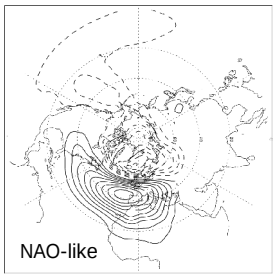
Example kindly prepared by Matt Sapiano

10. Dynamical consistency of leading EOFs

EOF1 of 850hPa sfn 28%

EOF2 of 850hPa sfn 16%





→ EOF analysis of different fields gives different leading modes!

M.H.P. Ambaum, B.J. Hoskins, and D.B. Stephenson,
North Atlantic Oscillation or Arctic Oscillation?
J. Climate, 14, (2001) plus the Corrigendum in 2002

10. EOF analysis good and bad points

Good points:

- Maximise something that is simple and important (domain total variance!)
- Easy to do using SVD for large data sets
- Not overly sensitive to outliers or distributional assumptions
- Produce uncorrelated PCs

Bad points:

- Does not exploit physical information to simplify the EOFs
- Depends on the choice of domain
- Takes no account of local spatial or temporal dependency
- Not based on a probability model (descriptive technique)
- Resulting PCs are linear functions of the non-linearly evolving variables

10. Some variants of EOF analysis

- Extended EOF
  - Augmented in time (Multi-channel SSA)
  - Multiple variables
  - Multiple vertical levels (3-d EOF)
- Complex EOF
- Simplified EOF
  - Rotated PCA
  - Simplified PCA
- Non-linear (y=f(x) with max variance)
- Other …

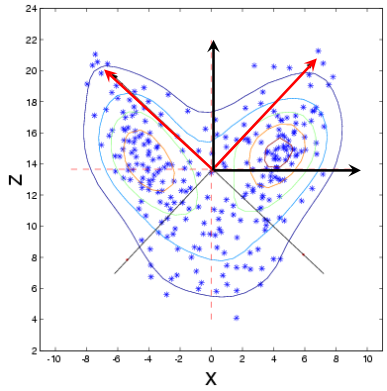
10. Rotated EOFs

- Horel (1981), Richman (1986), Jolliffe (1987,1995)
- Aim: **simplify** the loading weights by performing either an orthogonal or an oblique rotation of the EOFs
- Many possible simplicity conditions
  Example: Varimax

$$E \to B = ET$$

$$\max \sum_{k=1}^{q} \left[\sum_{j=1}^{p} B_{jk}^4 - \frac{1}{p} \left(\sum_{j=1}^{p} B_{jk}^2 \right) \right]$$

10. Rotated EOFs for the Lorenz system

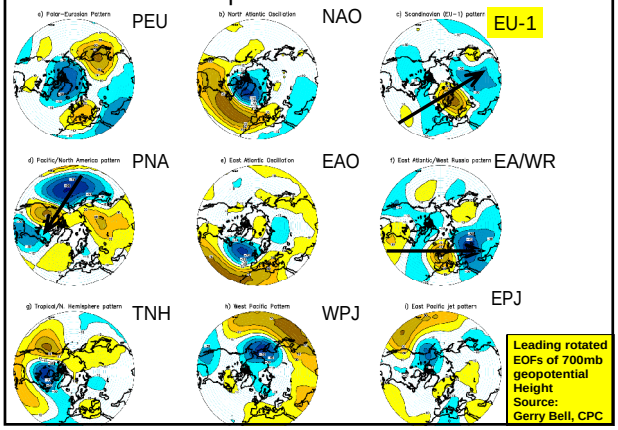


EOFs and rotated EOFs for the 3-variable Lorenz system

Kindly prepared by Abdel Hannachi

Stephenson et al. "On the existence of multiple regimes", Quart. J. Roy. Met. Soc., 2004.

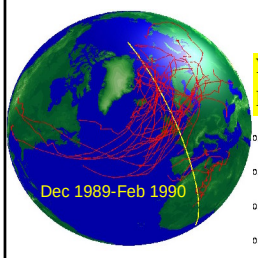
10. Northern Hemisphere wintertime rotated EOFs



Leading rotated EOFs of 700mb geopotential Height
Source: Gerry Bell, CPC

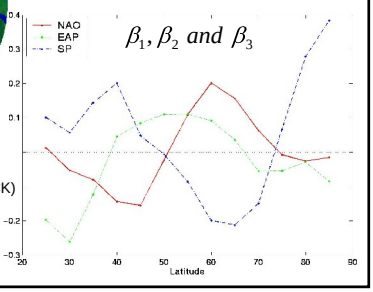
10. Example of use of rotated PCs

courtesy of Pascal Mailier



$$Y \mid X \sim Poisson(\mu)$$

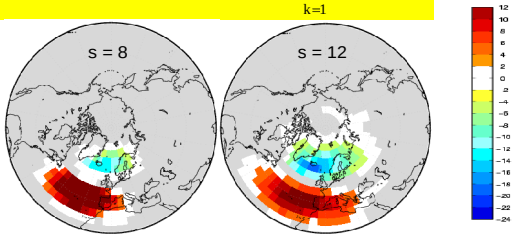
$$\log(\mu) = \beta_0 + \beta_1 X_{NAO} + \beta_2 X_{EAP} + \beta_3 X_{SCA}$$



- Oct-Mar 1950-2003
- NCAR/NCEP 6h 850mb vorticity
- Objective eastward tracks (TRACK)
- 4350 storms cross Greenwich

10. Simplified EOFs using SCoTLASS

$$\max\ e^T S e \ \text{ with } \ e^T e = 1 \ \text{ and } \ \sum_{k=1}^{p} |e_k| \le s$$



Analysis by Abdel Hannachi of NCAR/NCEP (DJF) SLP 1948-2000 p=1080 n=159
grey shading means EXACTLY zero!
projection gradient approach 11hours of MATLAB on workstation for 1 EOF!

See I.T. Jolliffe, Principal Component Analysis, 2nd edition, Springer 2002

Where is EOF analysis going next?

- Not likely to see less use of EOFs! And interpretation will remain controversial.
- More holistic use of EOFs as a basis set rather than over-interpretation of individual EOFs.
- EOF variants that can incorporate prior knowledge about spatial smoothness, locality, and other physical constraints. (Functional Data Analysis FDA)

10. Factor analysis for modes?

Aim of PCA:
Transform variables to diagonalise covariance matrix

$$X = \tilde{X}E$$
$$\Rightarrow S_X = E S_{\tilde{X}} E^T$$

Aim of Factor Analysis:
To find smaller number of factors that can explain covariance structure

$$X = \Lambda F + D$$
$$\Rightarrow S_X = \Lambda S_F \Lambda^T + S_D$$

→ Shouldn't we use factor analysis not PCA to isolate modes?

Summary

- Treat gridded spatial fields as vectors of p variables
- Use multivariate statistics to analyse the fields:
  - Exploratory Data Analysis (e.g. correlation maps)
  - Multivariate regression
  - Principal Component Analysis (Empirical Orthogonal Functions)
  - Factor Analysis
  - Cluster Analysis – for finding "regimes" in state space
  - Discriminant analysis – for partitioning state space
- High dimensionality and spatial and temporal dependency make this particularly challenging (and interesting!)
- Interpretation of resulting structures not easy!

Probability Problem Solution

Two types of simple event: G=girl B=boy
Event space: GG GB BG BB

P(GG|(GB or BG or GG))
=P(GG)/P(GB or BG or GG)
=0.25/3*0.25=1/3

Three types of simple event: B=boy R=girl rare name C=girl not rare name
P(B)=1/2 P(R)=p/2 p(C)=(1-p)/2
Event space: RR RC RB CC CR CB BR BC BB

P(RR or RC or CR|(RR or RC or CR or RB or BR))
=(2-p)/(4-p) (but note that there are many girl's names so p is often small!)