Application des méthodes de décomposition en composantes principales à la climatologie

Etude d'un exemple: les modes de variabilité de la température de surface de la mer dans l'océan Atlantique tropical et la variabilité atmosphérique associée.

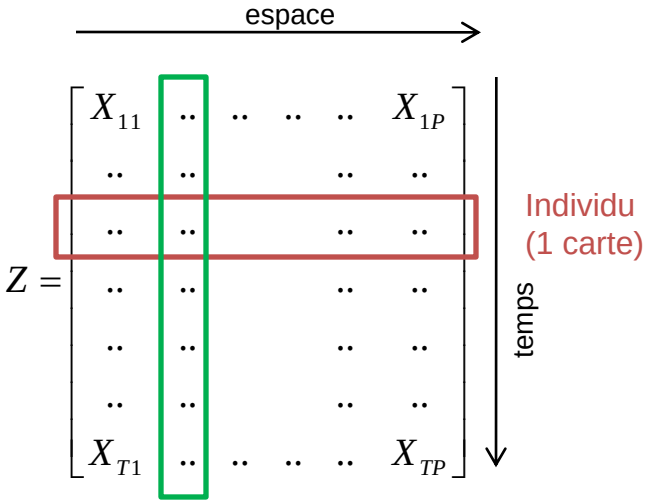
Travaux pratiques sous matlab / octave

DOCUMENT D'INTRODUCTION

Juliette Mignot – LOCEAN / IRD
juliette.mignot@locean-ipsl.upmc.fr

Définition de la matrice des données

On analyse un signal discrétisé Z_i^j où Z (déviation de moyenne temporelle nulle par rapport à $\overline{Z}$) est connu aux points Xj (j∈[1,P]) et aux temps ti (i ∈[1,T]). Typiquement, Z est une succession (série temporelle) de cartes de température à la surface de la mer (SST).



$$Z = \begin{bmatrix} X_{11} & .. & .. & .. & .. & X_{1P} \\ .. & .. & & & .. & .. \\ .. & .. & & & .. & .. \\ .. & .. & & & .. & .. \\ .. & .. & & & .. & .. \\ .. & .. & & & .. & .. \\ X_{T1} & .. & .. & .. & .. & X_{TP} \end{bmatrix}$$

La base canonique de R^P sur laquelle sont représentées les variables initialement est l'ensemble des vecteurs de type $e_i = [0 \quad 1 \quad \ldots \quad 0 \quad 0]$
C'est-à-dire que chaque vecteur correspond à 1 point d'espace pris individuellement.

Idée générale

L'idée de l'ACP (ou ici décomposition en EOF) est de déterminer une nouvelle base de R^P qui représente le mieux les liens (les covariances) entre les différents individus. On peut montrer que cette nouvelle base est obtenue par diagonalisation de la matrice de covariance.

Présentation des données

- Matrice Z: Données de température à la surface de la mer (SST) issues du jeu de réanalyses NCEP/NCAR:
http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis.derived.surfaceflux.html

-Résolution temporelle:
Données mensuelles couvrant la période 1948/01 – 2007/12 (60 années)
T=12x60=720 par défaut. Dans la suite, on travaillera essentiellement
• soit avec les moyennes mensuelles saisonnières, donc T=12 (moyenne sur les 60 années),
• soit avec les moyennes annuelles, donc T=60 (moyenne sur les 12 mois de chaque année)

- Résolution spatiale:
Le jeu de données initial couvre la Terre entière: [88.542N - 88.542S, 0E - 358.125E] soit 192x94 points de grille spatiaux (lonxlat)

On travaille ici avec un sous ensemble couvrant le domaine spatial [40S-40N, 100W-30E] (Atlantique tropical), soit 70x42 points de grille.
Donc P=70x42=2940

Organisation du TP

-Les données sont rangées dans le repertoire DONNEES/

- Un certain nombre de routines matlab/octave sont fournies dans le repertoire TOOLBOX/ . Ces routines sont utiles aux programmes fournis . En principe, vous n'avez rien a modifier dans ce repertoire.

- Les programmes fournis sont rangés dans PROGRAMMES/

- Avant toute chose, une fois sous matlab ou sous octave, lancer le programme PROGRAMMES/tpacpandreg_startup.m

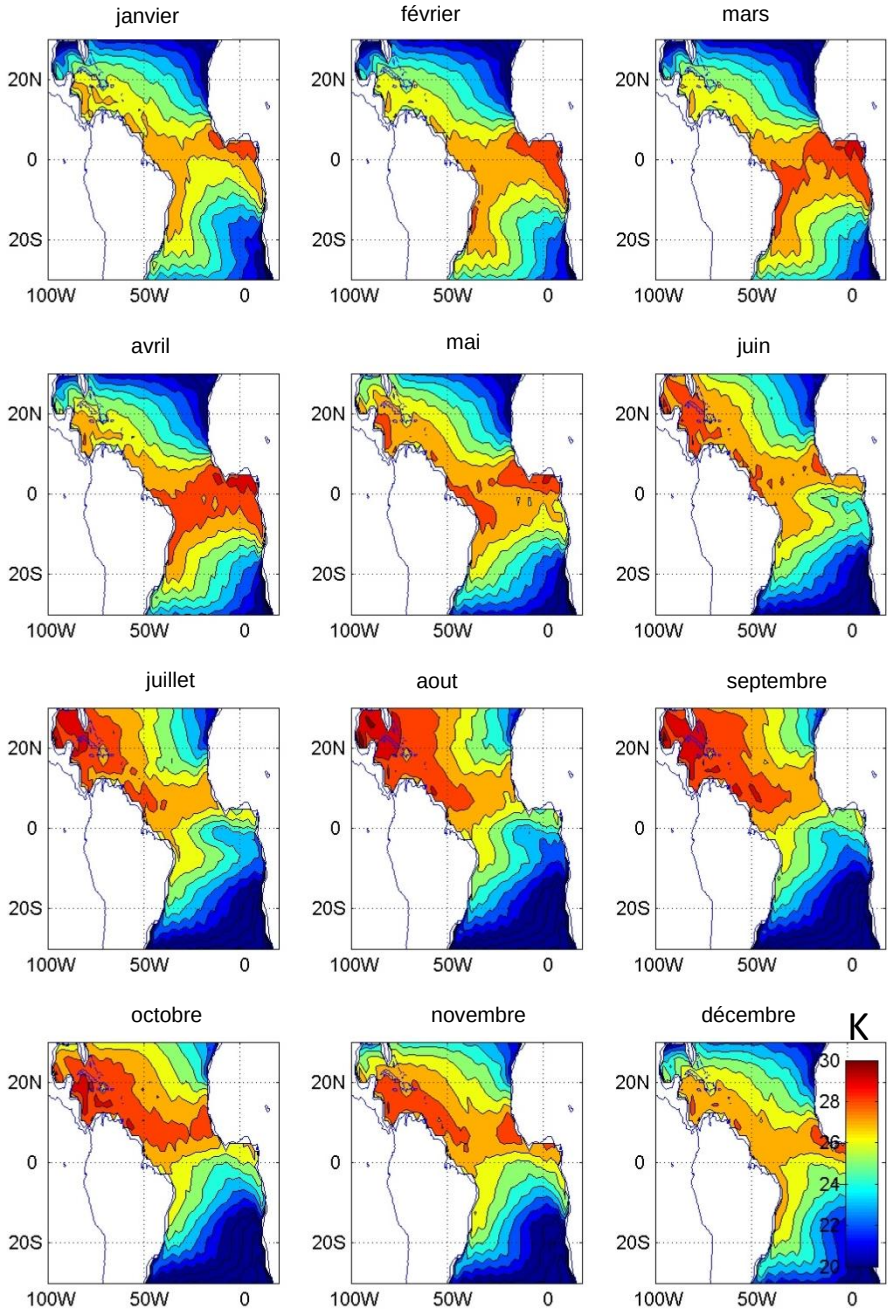
Exemple d'application

Visualisation de la variabilité saisonnière de la SST dans l'Atlantique tropical: cartes mensuelles moyennes

dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

L'information principale qui ressort de ces cartes est la variation saisonnière de la température de surface de la mer dans chaque hémisphère: en été boréal, il fait chaud dans l'hémisphère nord et froid au sud. Et inversement.
L'ACP ou décomposition en EOF est une méthode statistique qui va permettre de retrouver ce signal (qui est visible a l'œil nu ici) de façon objective.

Programme matlab carte_SST.m



Matrice de covariance

On définit la matrice de covariance (ou matrice d'inertie) par $C_{il} = \frac{1}{P}\sum_i Z_i^j Z_i^l$

Ou encore $C = {}^tZZ$. *C* est donc une matrice de dimensions *PxP*

La trace de la matrice de covariance est aussi appelée moment d'inertie totale. C'est une mesure de la dispersion du nuage des individus (mesures de SST) par rapport à son centre de gravité. La recherche de directions u_k maximisant l'inertie du nuage (donc tels que $Max({}^tu_k{}^tZZu_k)$) se traduit en algèbre linéaire par une recherche de vecteurs propres de la matrice de covariance C.

Diagonalisation de la matrice de covariance

C est une matrice symétrique de taille PxP définie positive, dont les valeurs propres λ_i sont réelles et positives.
Les vecteurs propres associés constituent une nouvelle base de R^P qui maximise la représentation de la variance/covariance entre les individus. Ils sont appelés les Fonctions Empiriques Orthogonales (EOF) pour *Z*.

Composantes principales

On définit les composantes principales de Z comme la décomposition (normalisée) du signal décrit par Z sur les EOF (vecteurs propres) En:

$$P_n^j = \sum_i Z_i^j E_{n,i} \qquad P_n = {}^{P_n}\!/_{std(P_n)}$$

Les composantes principales sont des vecteurs R^T . Ce sont de « nouvelles variables » dont les composantes donnent les coordonnées du nuage de point (des individus) sur les nouvelles directions principales (vecteurs propres, EOF)

Interprétations

La composante principale Pn est le coefficient de l'EOF En et dépend du temps (indice j) mais pas de l'espace. En est une carte spatiale indépendante du temps.

On a ainsi obtenu une décomposition de Z sous la forme $Z = \sum_n P_n E_n$
Où les EOF En sont orthogonales entre elles et les composantes principales Pn sont orthogonales entre elles.

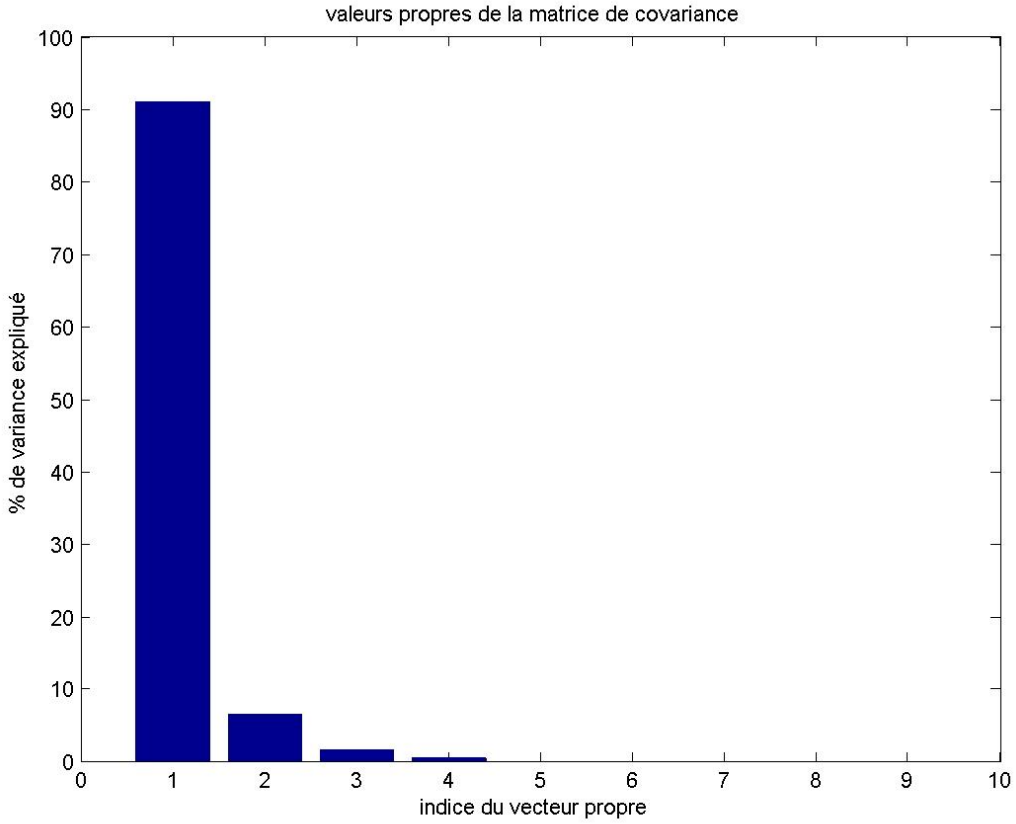
On peut par ailleurs montrer que l'on peut considérer $\lambda_m / \sum_n \lambda_n$ comme la part de la variance associée à l'EOF m.

Exemple d'application (1)

Décomposition des moyennes **mensuelles saisonnières** en EOF:

dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

Programme matlab
eof_NCEP_sais.m



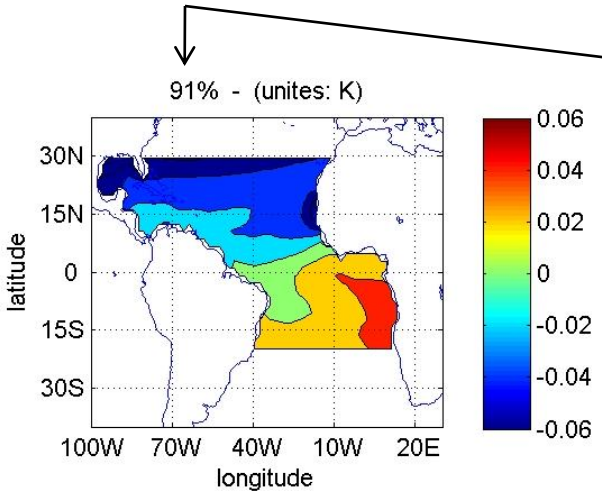
Répartition des valeurs propres: le 1er vecteur propre explique 91% de la variance saisonnière.
Les suivants expliquent une proportion largement plus faible, 6% pour le second et moins de 2% pour les suivants. Au moins dans un premier temps, on limite donc l'analyse au 1er vecteur propre.

Exemple d'application (1)

Décomposition des moyennes **mensuelles saisonnières** en EOF:

dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

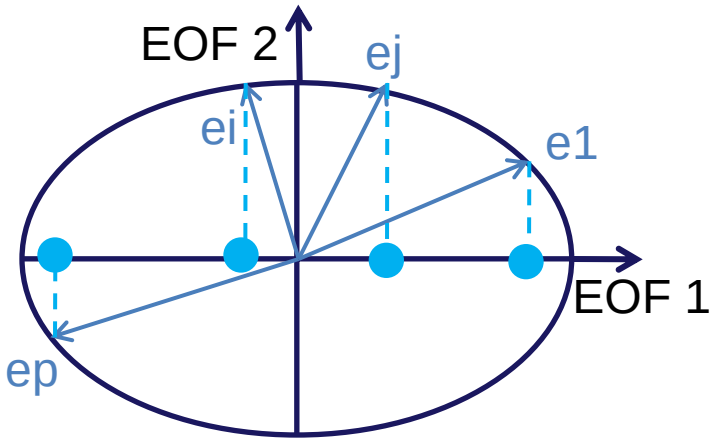
Programme matlab
eof_NCEP_sais.m



Pourcentage de variance expliquée

La **carte** illustre le poids (codé en couleur) de chacun des vecteurs de la base initiale (points d'espace) sur ce vecteur propre EOF1

Les données étant géographiques et nécessitant donc déjà 2 dimensions pour la visualisation, on ne représente qu'un mode à la fois.

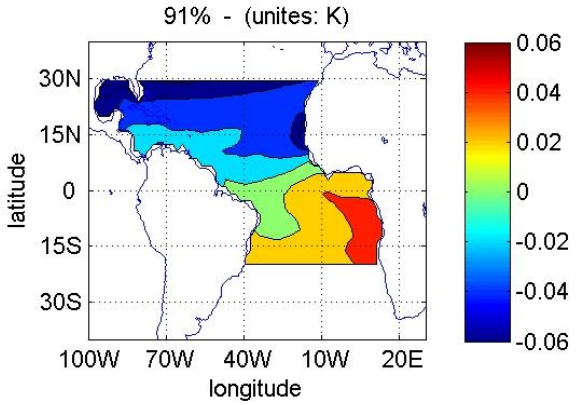


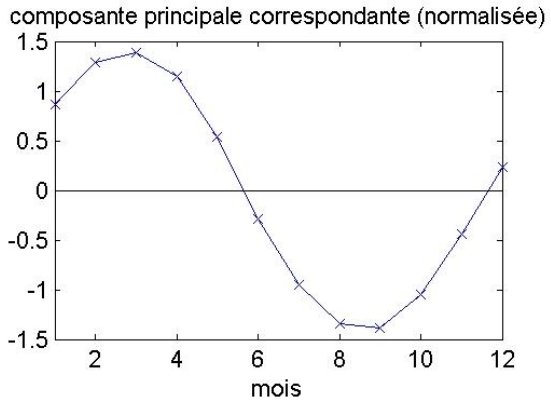
Exemple d'application (1)

Décomposition des moyennes **mensuelles saisonnières** en EOF:

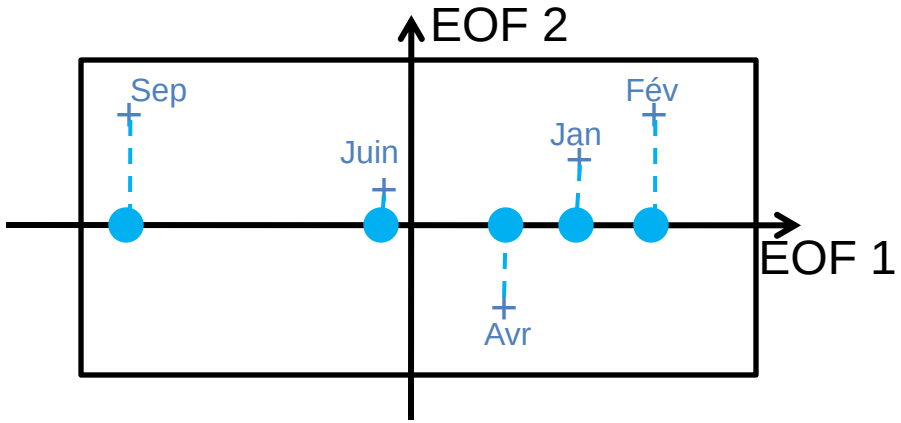
dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

Programme matlab
eof_NCEP_sais.m





La **série temporelle** correspond à la projection des individus de la matrice Z initiale sur le vecteur propre EOF 1

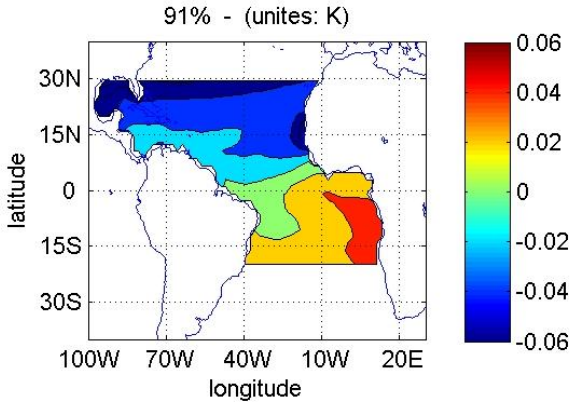


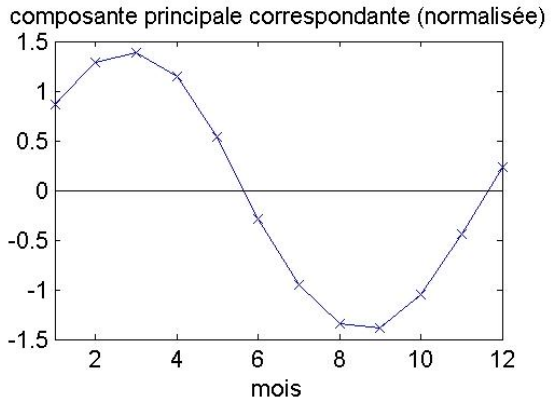
Exemple d'application (1)

Décomposition des moyennes **mensuelles saisonnières** en EOF:

dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

Programme matlab eof_NCEP_sais.m





91% de la variabilité saisonnière moyenne de la SST est expliquée par un refroidissement relatif dans l'hémisphère nord associé à un réchauffement au sud
La série temporelle illustre la modulation saisonnière de ce mode (sur les 12 mois de l'année): ce mode est en phase positive durant l'hiver boréal et négative (il fait + chaud au nord et + froid au sud) durant l'hiver austral