Chapter 3 Linear Multivariate Statistical Analysis

As one often encounters datasets with more than a few variables, multivariate statistical techniques are needed to effectively extract the information contained in these datasets. In the environmental sciences, examples of multivariate datasets are ubiquitous --- the air temperatures recorded by all the weather stations around globe, the satellite infrared images composed of numerous small pixels, the gridded output from a general circulation model, … The number of variables or time series from these datasets range from thousands to millions. Without a mastery of multivariate techniques, one is overwhelmed by these gigantic datasets. In this chapter, we review the principal component analysis method and its many variants, and the canonical correlation analysis method. These methods, using standard matrix techniques such as singular value decomposition, are relatively easy to use, but suffer from being linear, a limitation which will be lifted with neural network techniques in later chapters.

3.1 Principal component analysis (PCA)

3.1.1 Geometric approach to PCA

We have a dataset with variables $y_1,...,y_m$. These variables have been sampled *n* times, e.g. the *m* variables could be *m* time series containing *n* observations in time. If m is a large number, we would like to capture the essence of $y_1,...,y_m$ by a smaller set of variables $z_1,...,z_k$ (i.e. k <m; and hopefully k << m for truly large m). This is the objective of principal component analysis (PCA), also called empirical orthogonal function (EOF) analysis in meteorology and oceanography. We first begin with a geometric approach, which is more intuitive than the standard eigenvector approach to PCA.

Let us start with only 2 variables y_1 and y_2, as illustrated in Fig. 4. Clearly the bulk of the variance is along the axis z1. If r_i is the distance between the i^{th} data point and the axis z1, then the optimal z1 is found by minimizing $\sum_{i=1}^{n} r_i^2$. Note that PCA treats all variables equally, whereas regression divides variables into independent and dependent variables, hence the straight line described by z1 is in general different from the regression line.

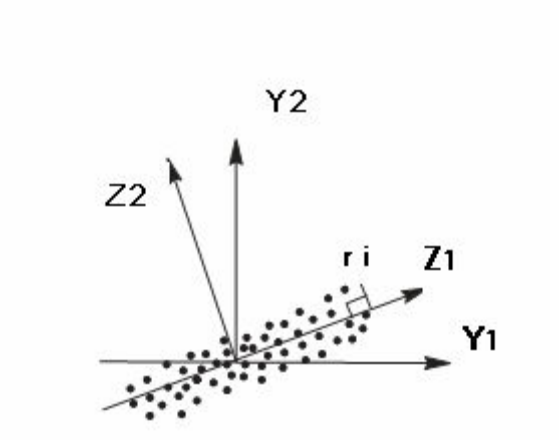


Fig. 4.1

3.1.2 Eigenvector approach to PCA

Taking the above example, a data point is transformed from its old coordinates (y1,y2) to new coordinates (z1,z2) via a rotation of the coordinate system (Fig. 4.2):

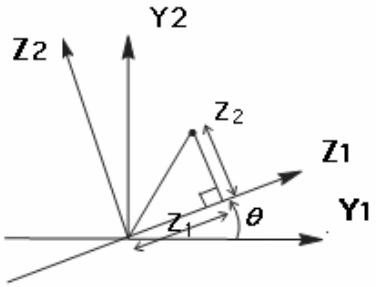
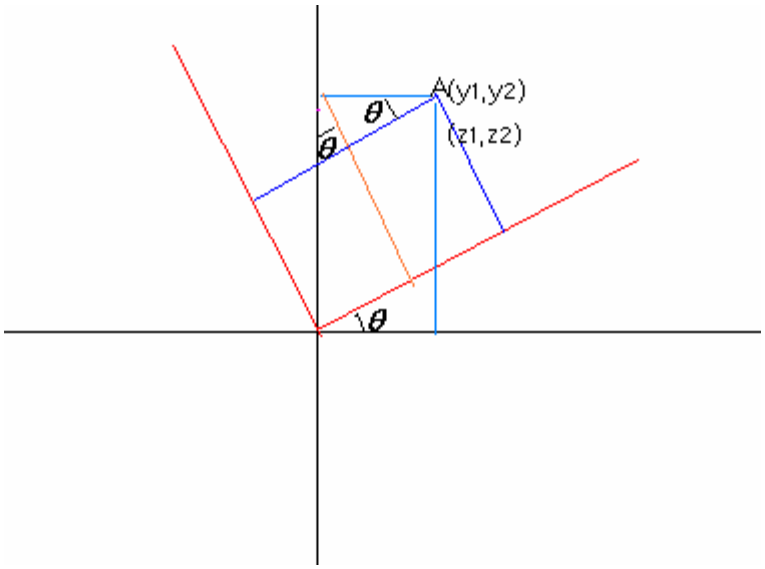


Fig. 4.2

$$z1 = y_1 \cos\theta + y_2 \sin\theta$$
$$z2 = -y_1 \sin\theta + y_2 \cos\theta \qquad (4.1)$$



In the general m-dimensional problem, we want to introduce new coordinates:

$$z_j = \sum_{l=1}^{m} e_{jl} y_l \ , j=1,\ldots,m \qquad (4.2)$$

The objective is to find

$$e_1 = [e_{11},...,e_{1m}]^T \qquad (4.3)$$

which maximizes var(z_1), i.e., find the coordinate transformation such that the variance of the dataset along the direction of the z_1 axis is maximized.

With $z_1 = \sum_{l=1}^{m} e_{1l} y_l = e_1^T y$, $\quad y = [y_1 ... y_m]^T$,

We have

$$\text{var}(z_1) = E[(z_1 - \bar{z}_1)(z_1 - \bar{z}_1)] = E[e_1^T (y - \bar{y})(y - \bar{y})^T e_1], \qquad (4.4)$$

where we have used the vector property $(a^T b)^T = b^T a$. Thus,

$$\text{var}(z_1) = e_1^T E[(y - \bar{y})(y - \bar{y})^T] e_1 = e_1^T C e_1 \qquad (4.5)$$

Where the covariance matrix C is given by

$$\boldsymbol{C} = \boldsymbol{E}[(\boldsymbol{y} - \overline{\boldsymbol{y}})(\boldsymbol{y} - \overline{\boldsymbol{y}})^T] \tag{4.6}$$

Clearly, the larger is $\|\boldsymbol{e}_1\|$, the larger $\text{var}(z_1)$ will be. Hence, we need to place a constraint on $\|\boldsymbol{e}_1\|$ while we try to maximize $\text{var}(z_1)$. Let us impose a normalization constraint $\|\boldsymbol{e}_1\|=1$, i.e.

$$\boldsymbol{e}_1^T \boldsymbol{e}_1 = 1 \tag{4.7}$$

Thus our optimization problem is to find $\boldsymbol{e}_1$ which maximize $\boldsymbol{e}_1^T \boldsymbol{C} \boldsymbol{e}_1$, subject to the constraint

$$\boldsymbol{e}_1^T \boldsymbol{e}_1 - 1 = 0 \tag{4.8}$$

The method of Lagrange multiplier s is commonly used to tackle optimization under constraints. Define the Lagrange function *L* by

$$\boldsymbol{L} = \boldsymbol{e}_1^T \boldsymbol{C} \boldsymbol{e}_1 - \lambda(\boldsymbol{e}_1^T \boldsymbol{e}_1 - 1) \tag{4.9}$$

Where λ is a Lagrange multiplier.

To obtain $\boldsymbol{e}_1$, we ask for

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{e}_1} = 0 \tag{4.10}$$

$$\boldsymbol{C}\boldsymbol{e}_1 - \lambda \boldsymbol{e}_1 = 0 \tag{4.11}$$

Which says that λ is an eigenvalue of the covariance matrix $\boldsymbol{C}$, which $\boldsymbol{e}_1$ the eigenvector. Multiplying this equation by $\boldsymbol{e}_1^T$ on the left, we obtain

$$\lambda = \boldsymbol{e}_1^T \boldsymbol{C} \boldsymbol{e}_1 = \text{var}(z_1) \tag{4.12}$$

Since $\boldsymbol{e}_1^T \boldsymbol{C} \boldsymbol{e}_1$ is maximized, the so are λ and $\text{var}(z_1)$. The new coordinate z_1, called the principal component (PC), is found from (4.2).

Next, we want to find z_2--- our task is to find $\boldsymbol{e}_2$ which maximizes $\text{var}(z_2) = \boldsymbol{e}_2^T \boldsymbol{C} \boldsymbol{e}_2$, subject to the constraint $\boldsymbol{e}_2^T \boldsymbol{e}_2 = 1$, and the constraint that z_2 be uncorrelated with z_1, i.e., the covariance between z_1 and z_2 be zero,

$$\text{cov}(z_1, z_2) = 0. \tag{4.13}$$

As $\boldsymbol{C} = \boldsymbol{C}^T$, we can write

$$\begin{aligned} 0 = \text{cov}(z_1, z_2) = \text{cov}(\boldsymbol{e}_1^T \boldsymbol{y}, \boldsymbol{e}_2^T \boldsymbol{y}) = \boldsymbol{E}[\boldsymbol{e}_1^T (\boldsymbol{y} - \bar{\boldsymbol{y}})(\boldsymbol{y} - \bar{\boldsymbol{y}}) \boldsymbol{e}_2] \\ = \boldsymbol{e}_1^T \boldsymbol{C} \boldsymbol{e}_2 = \boldsymbol{e}_2^T \boldsymbol{C} \boldsymbol{e}_1 = \boldsymbol{e}_2^T \lambda_1 \boldsymbol{e}_1 = \lambda_1 \boldsymbol{e}_2^T \boldsymbol{e}_1 = \lambda_1 \boldsymbol{e}_1^T \boldsymbol{e}_2 \end{aligned} \tag{4.14}$$

The orthogonal condition

$$\boldsymbol{e}_2^T \boldsymbol{e}_1 = 0 \tag{4.15}$$

can be used as a constraint in place of (4.13).

Upon introducing another Lagrange multiplier γ, we want to find an $\boldsymbol{e}_2$ which gives a stationary point of the Lagrange function L,

$$\boldsymbol{L} = \boldsymbol{e}_2^T \boldsymbol{C} \boldsymbol{e}_2 - \lambda(\boldsymbol{e}_2^T \boldsymbol{e}_2 - 1) - \text{r} \boldsymbol{e}_2^T \boldsymbol{e}_1 \tag{4.16}$$

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{e}_1} = 0; \quad \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{e}_2} = 0$$

$$\boldsymbol{C} \boldsymbol{e}_2 - \lambda \boldsymbol{e}_2 = 0 \tag{4.17}$$

Once again λ is an eigenvalue of the covariance matrix $\boldsymbol{C}$, which $\boldsymbol{e}_2$ the eigenvector. As

$$\lambda = \boldsymbol{e}_2^T \boldsymbol{C} \boldsymbol{e}_2 = \text{var}(z_2) \tag{4.18}$$

which is maximized, this $\lambda = \lambda_2$ is as large as possible with $\lambda_2 < \lambda_1$. (The case $\lambda_2 = \lambda_1$ is degenerate and will be discussed later). Hence, λ_2 is the second largest eigenvalue of $\boldsymbol{C}$, with $\lambda_2 = \text{var}(z_2)$. This process can be repeated for $z_3, z_4 \ldots$.

So far, $\boldsymbol{C}$ is the data covariance matrix, but it can also be the data correlation matrix, if one prefers correlation over covariance. In combined PCA, where two or more variables with different units are combined into one large data matrix for PCA --- e.g. finding the PCA modes of the combined sea surface temperature data and the sea level pressure data --- then one needs to normalize the variables so $\boldsymbol{C}$ is the correlation matrix.

So, the general procedure of PCA analysis for dataset $\boldsymbol{y}$ is as below:

(1) calculating covariance matrix (or correlation matrix) $\mathbf{C} = \{C_{ij}\}$

$$C_{i,j} = \sum_{k=1}^{N} (x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j}) \quad (i=1,\ldots M,\ j=1,\ldots M)$$

where M denote the number of variables (or grids), and N, the length of samples.

(2) calculating the eigenvalues and eigenvectors of $\boldsymbol{C}$

$$\boldsymbol{Ce} - \boldsymbol{\lambda e} = 0$$

where $\boldsymbol{e} = \{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_L\}$, $\boldsymbol{\lambda} = \{\lambda_1,\ldots,\lambda_L\}$ L=min{M, N}

(3) Usually $\lambda_1 > \lambda_2 >,\ldots,> \lambda_L$ but sometimes, the outputs are in reverse order. In this case, the first eigenvector corresponding with λ_L.

(3) calculating the PCs, i.e., $\boldsymbol{z} = \boldsymbol{e}^T \boldsymbol{y}$

3.1.3 Real and complex data

In general, for $\boldsymbol{y}$ real,

$$\boldsymbol{C} = \boldsymbol{E}[(\boldsymbol{y} - \overline{\boldsymbol{y}})(\boldsymbol{y} - \overline{\boldsymbol{y}})^T] \qquad (4.19)$$

implies that $\boldsymbol{C}^T = \boldsymbol{C}$, i.e., C is a real, symmetric matrix. A positive semi-defined matrix **A** is defined by the property that for any $\boldsymbol{v} \neq 0$, it follows that $\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \geq 0$. From the definition of $\boldsymbol{C}$ (4.5), it is clear that $\boldsymbol{v}^T \boldsymbol{C} \boldsymbol{v} \geq 0$ is satisfied. Hence $\boldsymbol{C}$ is a real, symmetric, positive semi-definite matrix.

If $\boldsymbol{y}$ is complex, then

$$\boldsymbol{C} = \boldsymbol{E}[(\boldsymbol{y} - \overline{\boldsymbol{y}})(\boldsymbol{y} - \overline{\boldsymbol{y}})^{T*} \quad (4.20)$$

With complex conjugation denoted by the superscript asterisk. As $\boldsymbol{C}^{T*} = \boldsymbol{C}, \boldsymbol{C}$ is a Hermitian matrix. It is also a positive semi-definite matrix.
Theorems on Hermitian matrix, positive semi-definite matrices tell us that: $\boldsymbol{C}$ has real eigenvalues:

$$\lambda_1 > \lambda_2 >,...,> \lambda_L >=0, \ \sum_{j=1}^{L} \lambda_j = \mathrm{var}(\boldsymbol{y}) \quad (4.21)$$

3.1.4 Orthogonality relations

Thus PCA amounts to finding the eigenvectors and eigenvalues of $\boldsymbol{C}$**.** The orthogonal eigenvectors then provide a basis, i.e., the data $\boldsymbol{y}$ can be expanded in terms of the eigenvectors $\boldsymbol{e}_j$:

$$\boldsymbol{y} - \overline{\boldsymbol{y}} = \sum_{j=1}^{m} \boldsymbol{a}_j(\boldsymbol{t})\boldsymbol{e}_j \quad (4.22)$$

where $\boldsymbol{a}_j(\boldsymbol{t})$ are the expansion coefficients. To obtain $\boldsymbol{a}_j(\boldsymbol{t})$, left multiply the above equation by $\boldsymbol{e}_i^{\boldsymbol{T}}$, and use the orthogonal relation of the eigenvectors,

$$\boldsymbol{e}_i^T \boldsymbol{e}_j = \boldsymbol{\delta}_{ij} \quad (4.23)$$

to get

$$\boldsymbol{a}_j(\boldsymbol{t}) = \boldsymbol{e}_j^T (\boldsymbol{y} - \overline{\boldsymbol{y}}), \quad (4.24)$$

i.e., $\boldsymbol{a}_j(\boldsymbol{t})$ is obtained by the projection of the data vector $\boldsymbol{y} - \overline{\boldsymbol{y}}$ onto the eigenvector $\boldsymbol{e}_j$, as the right hand side of this equation is simply a dot product between the two vectors. $\boldsymbol{a}_j(\boldsymbol{t})$ are usually called PCs, and $\boldsymbol{e}_j$ eigenvectors or EOFs.
There are two important properties:

(1) The expansion $\sum_{j=1}^{k} \boldsymbol{a}_j(t)\boldsymbol{e}_j(\boldsymbol{x})$ explains more of the variances of the data than any other linear combination $\sum_{j=1}^{k} \boldsymbol{b}_j(t)\boldsymbol{f}_j(\boldsymbol{x})$. Thus PCA provides the most efficient way to compress data.

(2) The time series in the set $\{\boldsymbol{a}_j\}$ are uncorrelated. We can write

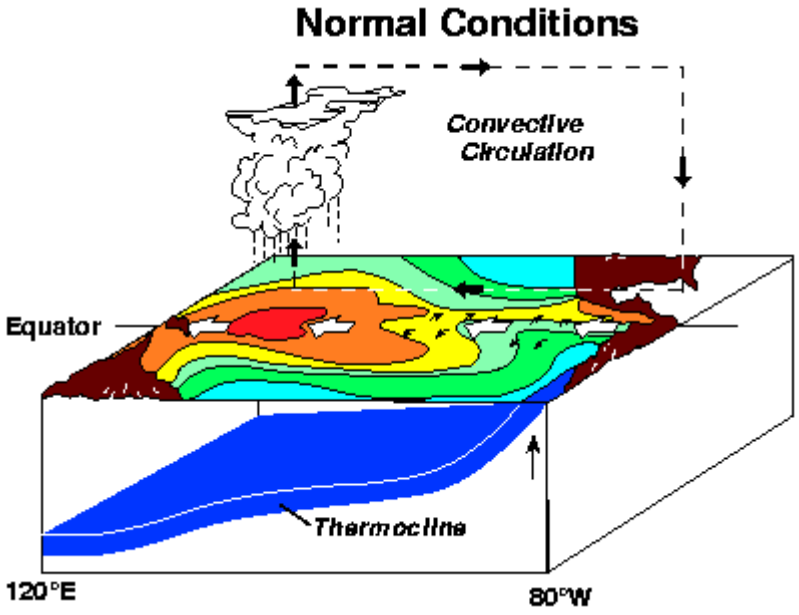
$$\boldsymbol{a}_j(t) = \boldsymbol{e}_j^T(\boldsymbol{y} - \overline{\boldsymbol{y}}) = (\boldsymbol{y} - \overline{\boldsymbol{y}})\boldsymbol{e}_j$$

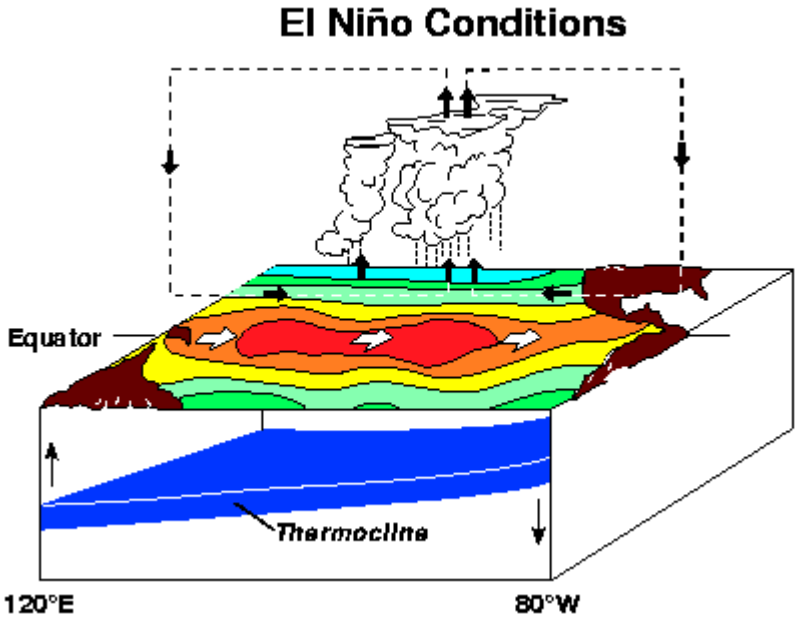
For $\boldsymbol{i} \neq \boldsymbol{j}$,

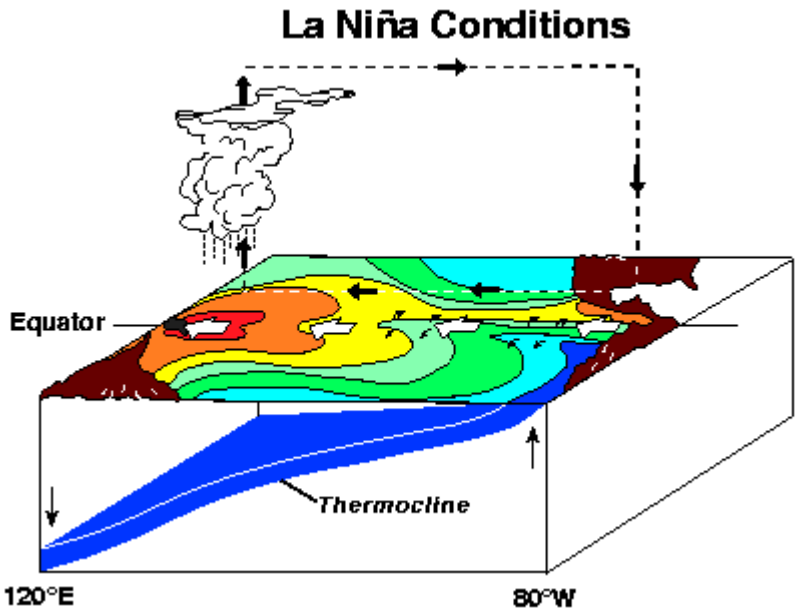
$$\begin{aligned} \mathrm{cov}(\boldsymbol{a}_i, \boldsymbol{a}_j) &= \boldsymbol{E}[\boldsymbol{e}_i^T(\boldsymbol{y} - \overline{\boldsymbol{y}})(\boldsymbol{y} - \overline{\boldsymbol{y}})\boldsymbol{e}_j] = \boldsymbol{e}_i^T \boldsymbol{E}[(\boldsymbol{y} - \overline{\boldsymbol{y}})(\boldsymbol{y} - \overline{\boldsymbol{y}})^T]\boldsymbol{e}_j \\ &= \boldsymbol{e}_i^T \boldsymbol{C}\boldsymbol{e}_j = \boldsymbol{e}_i^T \lambda_j \boldsymbol{e}_j = \lambda_j \boldsymbol{e}_i^T \boldsymbol{e}_j = 0 \end{aligned} \qquad (4.25)$$

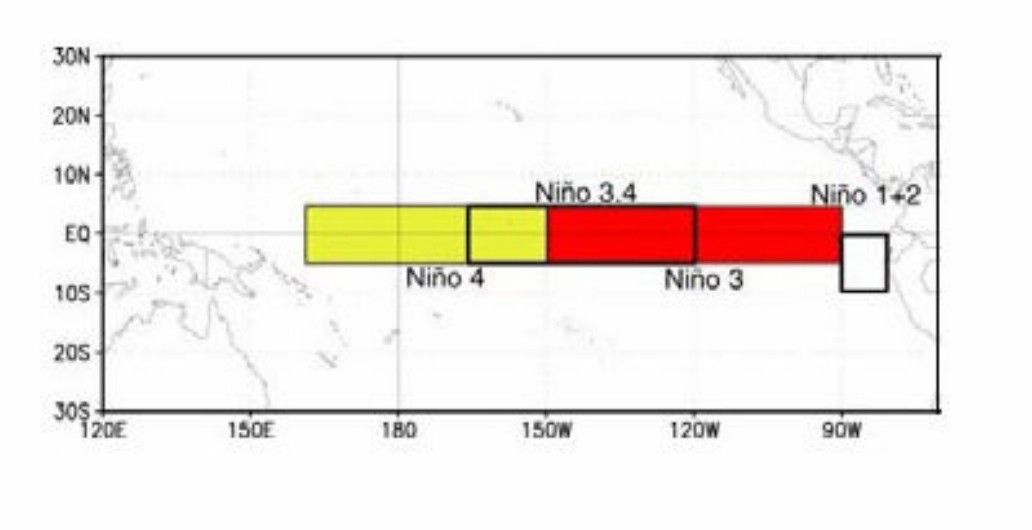
Hence PCA extracts the uncorrelated modes of variability of the data field. Note that no correlation between $(\boldsymbol{a}_i, \boldsymbol{a}_j)$ only means no linear relation between the two, there may still be nonlinear relation between them, which can be extracted by the nonlinear PCA method using neural network.

3.1.5 An Example: PCA of the tropical Pacific climate variability




El Niño Conditions
Equator
Thermocline
120°E
80°W


La Niña Conditions
Equator
Thermocline
120°E
80°W



Let us study the monthly tropical Pacific SST from NOAA (National Oceanic and atmospheric administration) for the period January, 1950 to August 2000. The SST field has 2 spatial dimensions, but can easily be rearranged into the form of $\boldsymbol{y}(t)$ for PCA analysis. Fig.4.3 is the spatial pattern for the first 3 modes (accounting for 51.8%, 10.1% and 7.3% respectively, of the total SST variance). Fig.4.4 are PCs corresponding with the first 3 modes.

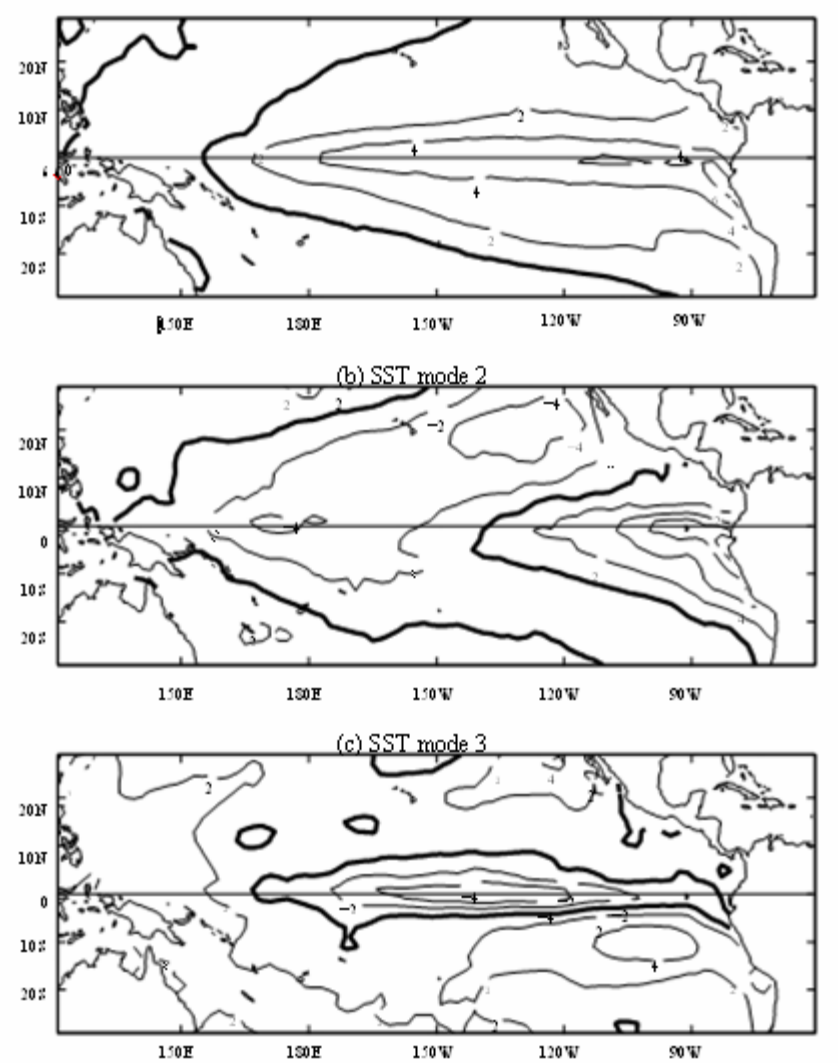

(b) SST mode 2
(c) SST mode 3
20N
10N
0
10S
20S
150E
180E
150W
120W
90W

Fig4.3

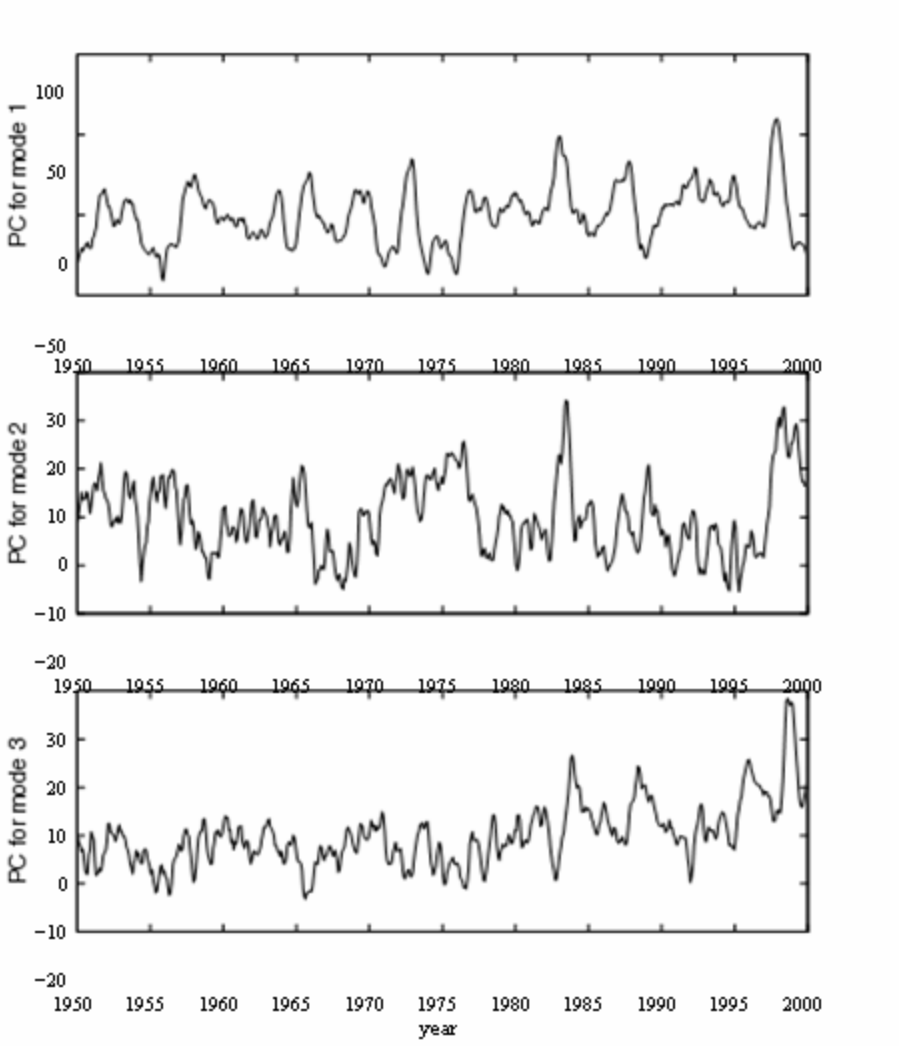


Fig.4.4

Mode1 (Fig. 4.3a) shows the largest SST anomalies occurring in the eastern and central equatorial Pacific. The PC1 (Fig4.4) can be used as an index for El Nino/La Nina.

Mode2 (Fig. 4.3b) has, along the equator, positive anomalies near the east and negative anomalies further west.

Mode3 (Fig. 4.3c) shows the largest anomaly occurring in the central equatorial Pacific, and the PC shows a rising trend after the mid 1970s.

3.1.6 Scaling the PCs and eigenvectors

There are various options for the scaling of the PCs $\{\boldsymbol{a}_j(t)\}$ and the eigenvectors $\{\boldsymbol{e}_j\}$. One can introduce an arbitrary scale factor α,

$$\boldsymbol{a}_j' = \frac{1}{\alpha}\boldsymbol{a}_j \quad \boldsymbol{e}_j' = \alpha \boldsymbol{e}_j \tag{4.26}$$

so that

$$\boldsymbol{y} - \bar{\boldsymbol{y}} = \sum_j \boldsymbol{a}_j' \boldsymbol{e}_j' \tag{4.27}$$

Our choice for the scaling has so far been

$$\boldsymbol{e}_i^T \boldsymbol{e}_j = \boldsymbol{\delta}_{ij} \tag{4.28}$$

which was the choice of Lorenz. The variance of the original data $\boldsymbol{y}$ is then contained in $\{\boldsymbol{a}_j(t)\}$, with

$$\text{var}(\boldsymbol{y}) = \boldsymbol{E}[\sum_{j=1}^{m} \boldsymbol{a}_j^2] \tag{4.29}$$

Another common choice is Hotelling's original choice

$$\boldsymbol{a}_j' = \frac{1}{\sqrt{\lambda_j}}\boldsymbol{a}_j, \qquad \boldsymbol{e}_j' = \sqrt{\lambda_j}\boldsymbol{e}_j \tag{4.30}$$

whence

$$\text{var}(\boldsymbol{y}) = \sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{m} \left\| \boldsymbol{e}_j' \right\|^2$$

From (4.25), $$\text{cov}(\boldsymbol{a}_i', \boldsymbol{a}_j') = \frac{1}{\sqrt{\lambda_i}\sqrt{\lambda_j}} \text{cov}(\boldsymbol{a}_i, \boldsymbol{a}_j) = \frac{\lambda_j}{\sqrt{\lambda_i}\sqrt{\lambda_j}} \boldsymbol{e}_i^T \boldsymbol{e}_j = \boldsymbol{\delta}_{ij} \tag{4.31}$$

The variance of the original data is now contained in $\{\boldsymbol{e}_j(t)\}$ instead. In sum, regardless of the arbitrary scale factor, the PCA eigenvectors are orthogonal and the PCs are uncorrelated.

If $\tilde{\boldsymbol{y}}_l$ is $\boldsymbol{y}_l$ with mean removed and normalized by standard deviation, then one can show that the correlation

$$\boldsymbol{\rho}(\boldsymbol{a}_j'(\boldsymbol{t}), \tilde{\boldsymbol{y}}_l) = \boldsymbol{e}_{jl}' \qquad (4.32)$$

the *l*th element of $\boldsymbol{e}_j'$. Hence the *l*th element of $\boldsymbol{e}_j'$ conveniently provides the correlation between the PC $\boldsymbol{a}_j'$ and the standardized variable $\tilde{\boldsymbol{y}}_l$.

3.1.7 Degeneracy of Eigenvalues

A degenerate case arises when $\lambda_i = \lambda_j$. When two eigenvalues are equal, the eigenvectors are not unique.

A simple example of degeneracy is illustrated by a propagating plane wave,

$$h(x, y, t) = A\cos(ky - \omega t) \qquad (4.33)$$

Which can be expressed in terms of two standing waves:

$$h(x, y, t) = A\cos(ky)\cos(\omega t) + A\sin(ky)\sin(\omega t) \qquad (4.34)$$

If we perform PCA on *h(x,y,t)*, we get two modes with equal eigenvalues. To see this, note that in the x-y plane, cos(ky) and sin(ky) are orthogonal, while $\cos(\omega t)$ and $\sin(\omega t)$ are uncorrelated, so (4.34) satisfies the properties of PCA modes in that the eigenvectors are orthogonal and the PCs are uncorrelated. As (4.34) is a PCA decomposition, with the two modes both having the same amplitude A, hence the eigenvalues $\lambda_1 = \lambda_2$, and the case is degenerate. Thus propagating waves in the data leads to degeneracy in the eigenvalues. If one finds eigenvalues of very similar magnitudes from a PCA analysis, that implies near degeneracy and there may be propagating waves in the data.

3.1.8 A smaller covariance matrix

Let the data matrix be

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \cdots & \cdots & \cdots \\ y_{m1} & \cdots & y_{mn} \end{bmatrix} \qquad (4.35)$$

Where *m* is the number of spatial points and *n* the number of time points.

Assuming the temporal mean has been removed, then covariance matrix

$$\boldsymbol{C} = \frac{1}{\boldsymbol{n}}\boldsymbol{Y}\boldsymbol{Y}^{T} \quad \text{and} \quad \boldsymbol{C}' = \frac{1}{\boldsymbol{n}}\boldsymbol{Y}^{T}\boldsymbol{Y} \tag{4.36}$$

C is a $m \times m$ matrix, and $\boldsymbol{C}'$ is a $n \times n$ matrix.

In most problems, the size of the two matrices are very different. For instance, for global $5^{o} \times 5^{o}$ monthly sea level pressure data collected over 50 years, the total number of spatial grid points is *m*=2592 while the number of time points is *n*=600. Obviously, it will be much cheaper to solve the eigen problem for $\boldsymbol{C}'$ than for C.

The matrix theory says: C and $\boldsymbol{C}'$ have same eigenvalues. The question is now how to get eigenvectors of C from eigenvectors of $\boldsymbol{C}'$?

$$\boldsymbol{C}' = \frac{1}{\boldsymbol{n}}\boldsymbol{Y}^{T}\boldsymbol{Y}$$

$$\frac{1}{\boldsymbol{n}}\boldsymbol{Y}^{T}\boldsymbol{Y}\boldsymbol{v}_{j} = \lambda_{j}\boldsymbol{v}_{j} \tag{4.37}$$

where $\boldsymbol{v}_{j}$ and λ_{j} are eigenvectors and eigenvalues of $\boldsymbol{C}'$.

Multiplying $\boldsymbol{Y}$ on both sides of (4.36), we have

$$\boldsymbol{Y}\frac{1}{\boldsymbol{n}}\boldsymbol{Y}^{T}\boldsymbol{Y}\boldsymbol{v}_{j} = \boldsymbol{Y}\lambda_{j}\boldsymbol{v}_{j} \tag{4.38}$$

$$(\frac{1}{\boldsymbol{n}}\boldsymbol{Y}\boldsymbol{Y}^{T})(\boldsymbol{Y}\boldsymbol{v}_{j}) = \lambda_{j}(\boldsymbol{Y}\boldsymbol{v}_{j}) \tag{4.39}$$

Denoting

$$\boldsymbol{e}_{j} = \boldsymbol{Y}\boldsymbol{v}_{j}\ , \tag{4.40}$$

we have

$$\boldsymbol{C}\boldsymbol{e}_{j} = \lambda_{j}\boldsymbol{e}_{j} \tag{4.41}$$

(4.41) is just the eigen equation for $\boldsymbol{C}$, meaning $\boldsymbol{e}_j$ is an eigenvector for $\boldsymbol{C}$.

In summary, solving the eigen problem for the smaller matrix $\boldsymbol{C}'$ yields the eigenvalues $\{\lambda_j\}$ and eigenvectors $\{\boldsymbol{v}_j\}$. The eigenvectors $\{\boldsymbol{e}_j\}$ for the bigger matrix $\boldsymbol{C}$ are then obtained from (4.40) $\boldsymbol{e}_j = \boldsymbol{Y}\boldsymbol{v}_j$.

3.1.9 Singular value decomposition

Instead of solving the eigen problem of the data covariance matrix $\boldsymbol{C}$, a computationally more efficient way to perform PCA is via Singular Value Decomposition (SVD) of the $m \times n$ data matrix $\boldsymbol{Y}$ given by (4.35). Without loss of generality, we can assume $m \geq n$, then the SVD Theorem says that

$$\boldsymbol{Y} = \boldsymbol{E}\boldsymbol{S}\boldsymbol{F}^T \qquad (4.42)$$

$\boldsymbol{E} : \boldsymbol{m} \times \boldsymbol{m}; \quad \boldsymbol{S} : \boldsymbol{m} \times \boldsymbol{n}; \quad \boldsymbol{F} : \boldsymbol{n} \times \boldsymbol{n}$

E and F are orthonormal matrices, i.e., they satisfy

$$\boldsymbol{E}^T\boldsymbol{E} = \boldsymbol{I}, \qquad \boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{I}, \qquad (4.43)$$

Where I is the identity matrix. The leftmost *n* columns of **E** contain the n left singular vectors, and then columns of **F** the *n* right singular vectors, while the diagonal elements of **S** are the singular values.

The covariance matrix C can be rewritten as

$$\boldsymbol{C} = \frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}^T = \frac{1}{n}\boldsymbol{E}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{E}^T \qquad (4.44)$$

The matrix $$\frac{1}{\boldsymbol{n}}\boldsymbol{S}\boldsymbol{S}^T = \Lambda \qquad (4.45)$$

is diagonal and zero everywhere, except in the upper left $\boldsymbol{n} \times \boldsymbol{n}$ corner, containing $\frac{1}{n}s(i,i)^2$ --- s(i,i) is singular values

Right multiply Eq. (4.44) by $\boldsymbol{E}$ gives

$$\boldsymbol{CE} = \boldsymbol{E}\Lambda \qquad (4.46)$$

(4.46) is a standard eigen equation, with E is the eigenvectors and Λ is the eigenvalues. So, we can use SVD to derive eigenvectors and eigenvalues with the relation: $\lambda_j = \frac{1}{n} s(j,j)^2$.

SVD approach to PCA is at least twice as fast as the eigen approach. So, SVD is in particular useful for large datasets. Matlab program for SVD is svd.

3.1.20 Significance tests

In practice, the higher PCA modes, which basically contain noise, are rejected. How does one decide how many modes to retain? There are some "rules of thumb". One of the simplest approach is to plot the eigenvalues λ_j as a function of the mode number *j*. Hopefully, from the plot, one finds an abrupt transition from large eigenvalues to small eigenvalues around mode number *m*. One can then retain the first *m* modes. Alternatively, the Kaiser test rejects the modes with eigenvalues λ less than the mean value $\bar{\lambda}$.

Computationally more involved is the Monte Carlo test, which involves setting up random data matrices $\boldsymbol{R}_k\,(\boldsymbol{k} = 1,...\boldsymbol{K})$, of the same size as the data matrix $\boldsymbol{Y}$. The random elements are normally distributed, with the variance of the random data matching the variance of the actual data. PCA is performed on each of the random matrices, yielding eigenvalues $\lambda^{(k)}{}_j$. Assume for each *k*, the set of eigenvalues are sorted in descenting order. For each *j*, one examines the distribution of the K values of $\lambda^{(k)}{}_j$, and finds the level $\lambda_{0.05}$, which is exceeded only by 5% of the $\lambda^{(k)}{}_j$ values. The eigenvalues λ_j from $\boldsymbol{Y}$ which failed to rise above this $\lambda_{0.05}$ level are then rejected.

Since the Monte Carlo method performs PCA on *K* matrices and *K* is typically about 100, it can be costly for large data matrices.

3.2 Canonical correlation analysis (CCA)

Given a set of variables $\{y_j\}$, PCA finds the linear modes accounting for the maximum amount of variance in the dataset. When there are two sets of variables $\{x_i\}$ and $\{y_j\}$, *canonical correlation analysis* (CCA) , first introduced by ****Hotelling (1936), finds the modes of maximum correlation between $\{x_i\}$ and $\{y_j\}$, rendering CCA a standard tool for discovering linear relations between two fields. CCA is a generalization of the Pearson correlation between two variables x and y to two sets of variables $\{x_i\}$ and $\{y_j\}$. Thus CCA can be viewed as a "doubled-barreled PCA". A variant of the CCA method finds the modes of maximum *covariance* between $\{x_i\}$ and $\{y_j\}$— this variant is called the *maximum covariance analysis* (MCA) by von Storch and Zwiers (1999), and because it uses the SVD matrix technique, it is also simply called the SVD (singular value decomposition) method by other researchers, though this name is confusing as it is used to denote both a matrix technique and a multivariate statistical technique.

In PCA, one finds a linear combination of the y_j variables, i.e. $\mathbf{e}_1^{\mathrm{T}}\mathbf{y}$, which has the largest variance (subject to $\|\mathbf{e}_1\| = 1$). Next, one finds $\mathbf{e}_2^{\mathrm{T}}\mathbf{y}$ with the largest variance, but with $\mathbf{e}_2^{\mathrm{T}}\mathbf{y}$ uncorrelated with $\mathbf{e}_1^{\mathrm{T}}\mathbf{y}$; and similarly for the higher modes.

In CCA, one finds $\mathbf{f}_1$ and $\mathbf{g}_1$, so that the correlation between $\mathbf{f}_1^{\mathrm{T}}\mathbf{x}$ and $\mathbf{g}_1^{\mathrm{T}}\mathbf{y}$ is maximized. Next find $\mathbf{f}_2$ and $\mathbf{g}_2$ so that the correlation between $\mathbf{f}_2^{\mathrm{T}}\mathbf{x}$ and $\mathbf{g}_2^{\mathrm{T}}\mathbf{y}$ is maximized, with $\mathbf{f}_2^{\mathrm{T}}\mathbf{x}$ and $\mathbf{g}_2^{\mathrm{T}}\mathbf{y}$ uncorrelated with both $\mathbf{f}_1^{\mathrm{T}}\mathbf{x}$ and $\mathbf{g}_1^{\mathrm{T}}\mathbf{y}$. And so forth for the higher modes.

3.2.1 CCA theory

Consider two datasets

$$\mathbf{x}(t) = x_{il}, \qquad i = 1, \cdots, n_x, \quad l = 1, \cdots, n_t\,, \tag{4.61}$$

and

$$\mathbf{y}(t) = y_{jl}, \qquad j = 1, \cdots, n_y, \quad l = 1, \cdots, n_t\,. \tag{4.62}$$

i.e., x and y need not have the same spatial dimensions, but need the same time dimension n_t. Assume x and y have zero means. Let

$$u = \mathbf{f}^{\mathrm{T}}\mathbf{x}, \quad v = \mathbf{g}^{\mathrm{T}}\mathbf{y}\,. \tag{4.63}$$

The correlation

$$\rho = \frac{\mathrm{cov}(u,v)}{\sqrt{\mathrm{var}(u)\,\mathrm{var}(v)}} = \frac{\mathrm{cov}(\mathbf{f}^{\mathrm{T}}\mathbf{x}, \mathbf{g}^{\mathrm{T}}\mathbf{y})}{\sqrt{\mathrm{var}(u)\,\mathrm{var}(v)}} = \frac{\mathbf{f}^{\mathrm{T}}\mathrm{cov}(\mathbf{x},\mathbf{y})\mathbf{g}}{\sqrt{\mathrm{var}(\mathbf{f}^{\mathrm{T}}\mathbf{x})\,\mathrm{var}(\mathbf{g}^{\mathrm{T}}\mathbf{y})}}\,, \tag{4.64}$$

Where we have invoked

$$\mathrm{cov}(\mathbf{f}^{\mathrm{T}}\mathbf{x}, \mathbf{g}^{\mathrm{T}}\mathbf{y}) = \mathrm{E}(\mathbf{f}^{\mathrm{T}}\mathbf{x}(\mathbf{g}^{\mathrm{T}}\mathbf{y})^{\mathrm{T}}) = \mathrm{E}(\mathbf{f}^{\mathrm{T}}\mathbf{x}\mathbf{y}^{\mathrm{T}}\mathbf{g}) = \mathbf{f}^{\mathrm{T}}\mathrm{E}[\mathbf{x}\mathbf{y}^{\mathrm{T}}]\mathbf{g}\,. \tag{4.65}$$

We want u and v, the two canonical variates or canonical correlation coordinates, to have maximum correlation between them, i.e., f and g are chosen to maximize ρ. We are of course free to normalize f and g as we like, because if f and g maximize ρ, so will $\alpha\boldsymbol{f}$ and $\beta\boldsymbol{g}$, for any nonzero α and β. We choose the normalization condition:

$$\mathrm{var}(\mathbf{f}^{\mathrm{T}}\mathbf{x}) = 1 = \mathrm{var}(\mathbf{g}^{\mathrm{T}}\mathbf{y})\,. \tag{4.66}$$

Since

$$\mathrm{var}(\mathbf{f}^{\mathrm{T}}\mathbf{x}) = \mathrm{cov}(\mathbf{f}^{\mathrm{T}}\mathbf{x}, \mathbf{f}^{\mathrm{T}}\mathbf{x}) = \mathbf{f}^{\mathrm{T}}\mathrm{cov}(\mathbf{x}, \mathbf{x})\mathbf{f} \equiv \mathbf{f}^{\mathrm{T}}\mathbf{C}_{xx}\mathbf{f} \tag{4.67}$$

and

$$\mathrm{var}(\mathbf{g}^{\mathrm{T}}\mathbf{y}) = \mathbf{g}^{\mathrm{T}}\mathbf{C}_{yy}\mathbf{g} \tag{4.68}$$

So, we have

$$\mathbf{f}^{\mathrm{T}}\mathbf{C}_{xx}\mathbf{f} = 1\,, \quad \mathbf{g}^{\mathrm{T}}\mathbf{C}_{yy}\mathbf{g} = 1\,. \tag{4.69}$$

With (4.66), (4.64) reduces to

$$\rho = \mathbf{f}^{\mathrm{T}}\mathbf{C}_{xy}\mathbf{g}\,, \tag{4.70}$$

Where $\boldsymbol{C}_{xy} = \mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})$

The problem is to maximize (4.70) subject to constraints (4.69). We will again use the method of Lagrange multipliers, where we impose the constrains into the Lagrange function L,

$$L = \mathbf{f}^{\mathrm{T}}\mathbf{C}_{xy}\mathbf{g} + \alpha(\mathbf{f}^{\mathrm{T}}\mathbf{C}_{xx}\mathbf{f} - 1) + \beta(\mathbf{g}^{\mathrm{T}}\mathbf{C}_{yy}\mathbf{g} - 1)\,, \tag{4.71}$$

where α and β are the unknown Lagrange multipliers. To find the stationary points of L, we need

$$\frac{\partial L}{\partial \mathbf{f}} = \mathbf{C}_{xy}\mathbf{g} + 2\alpha\mathbf{C}_{xx}\mathbf{f} = 0\,, \tag{4.72}$$

$$\frac{\partial L}{\partial \mathbf{g}} = \mathbf{C}_{xy}^{\mathrm{T}}\mathbf{f} + 2\beta\mathbf{C}_{yy}\mathbf{g} = 0\,.$$

Hence

$$\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{g} = -2\alpha\mathbf{f}\,, \tag{4.73}$$

$$\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\mathrm{T}}\mathbf{f} = -2\beta\mathbf{g}\,. \tag{4.74}$$

Substituting (4.74) into (4.73) yields

$$\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\mathrm{T}}\,\mathbf{f} \equiv \mathbf{M}_f\mathbf{f} = \lambda\mathbf{f}\,, \tag{4.75}$$

With $\lambda = 4\alpha\beta$. Similarly, substituting (4.73) into (4.74) yields

$$\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\mathrm{T}}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\,\mathbf{g} \equiv \mathbf{M}_g\mathbf{g} = \lambda\mathbf{g}\,. \tag{4.76}$$

Both these equations can be viewed as eigenvalue equations, with $\boldsymbol{M}_f$ and $\boldsymbol{M}_g$ sharing the same non-zero eigenvalues λ. As $\boldsymbol{M}_f$ and $\boldsymbol{M}_g$ are known from the data, f can be found by solving the eigenvalue problem (4.75). $\beta\boldsymbol{g}$ can then be obtained from (4.74). Since β is unknown, the magnitude of g is unknown, and the normalization conditions (4.69) are used to determine the magnitude of g. Alternatively, one can use (4.76) to solve for g first, then obtain f from (4.73), and the normalization condition (4.69). The matrix $\boldsymbol{M}_f$ is of dimension $\boldsymbol{n}_x \times \boldsymbol{n}_x$, while $\boldsymbol{M}_g$ is $\boldsymbol{n}_y \times \boldsymbol{n}_y$, so one usually picks the smaller of the two to solve the eigenvalue problem.

From (4.70),

$$\rho^2 = \mathbf{f}^{\mathrm{T}}\mathbf{C}_{xy}\mathbf{g}\,\mathbf{g}^{\mathrm{T}}\mathbf{C}_{xy}^{\mathrm{T}}\mathbf{f} = 4\alpha\beta\,(\mathbf{f}^{\mathrm{T}}\mathbf{C}_{xx}\mathbf{f})\,(\mathbf{g}^{\mathrm{T}}\mathbf{C}_{yy}\mathbf{g})\,, \tag{4.77}$$

Where (4.72) has been invoked. From (4.69) , (4.77) reduces to

$$\rho^2 = \lambda. \tag{4.78}$$

The eigenvalue problems (4.75) and (4.76) yield *n* number of λ_s, with n=min(n_x, n_y). Assuming the λ_s to be all distinct and nonzero, we have for each λ_j (j=1,…,n), a pair of eigenvectors, f_j and g_j, and a pair of canonical variates, u_j and v_j, with correlation $\rho_j = \sqrt{\lambda_j}$ between the two.

Let us write the forward mapping from the variables x(t) and y(t) to the canonical variates $u(t) = [u_1(t),...,u_n(t)]^T$ and $v(t) = [v_1(t),...,v_n(t)]^T$ as

$$\mathbf{u} = [\mathbf{f}_1^{\mathrm{T}}\mathbf{x}, \cdots, \mathbf{f}_n^{\mathrm{T}}\mathbf{x}]^{\mathrm{T}} = \mathcal{F}^{\mathrm{T}}\mathbf{x}, \quad \mathbf{v} = \mathcal{G}^{\mathrm{T}}\mathbf{y} \tag{4.79}$$

Next, we need to find the inverse mapping from $u(t) = [u_1(t),...,u_n(t)]^T$ and $v(t) = [v_1(t),...,v_n(t)]^T$ to the original variables x and y. Let

$$\mathbf{x} = \mathbf{F}\mathbf{u}, \quad \mathbf{y} = \mathbf{G}\mathbf{v}. \tag{4.80}$$

We note that

$$\mathrm{cov}(\mathbf{x}, \mathbf{u}) = \mathrm{cov}(\mathbf{x}, \mathcal{F}^{\mathrm{T}}\mathbf{x}) = \mathrm{E}[\mathbf{x}(\mathcal{F}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}] = \mathrm{E}[\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\mathcal{F}] = \mathbf{C}_{xx}\mathcal{F}, \tag{4.81}$$

and

$$\mathrm{cov}(\mathbf{x}, \mathbf{u}) = \mathrm{cov}(\mathbf{F}\,\mathbf{u}, \mathbf{u}) = \mathbf{F}\,\mathrm{cov}(\mathbf{u}, \mathbf{u}) = \mathbf{F}. \tag{4.82}$$

Eqs (4.81) and (4.82) imply

$$\mathbf{F} = \mathbf{C}_{xx}\mathcal{F}. \tag{4.83}$$

Similarly

$$\mathbf{G} = \mathbf{C}_{yy}\mathcal{G}. \tag{4.84}$$

Hence the inverse mappings F and G (from the canonical variates to x and y) can be calculated from the forward mapping $\mathcal{F}^T$ and $\mathcal{G}^T$. The matrix **F** is composed of column vectors $\mathbf{F}_j$, and **G**, of column vectors $\mathbf{G}_j$. $\mathbf{F}_j$ and $\mathbf{G}_j$ are the *canonical correlation patterns* associated with u_j and v_j, the canonical variates. In general, orthogonality of vectors within a set is not satisfied by any of the four sets $\{\mathbf{F}_j\}$, $\{\mathbf{G}_j\}$, $\{\mathbf{f}_j\}$ and $\{\mathbf{g}_j\}$, and

$$\mathrm{cov}(u_i, u_j) = \mathrm{cov}(v_i, v_j) = \mathrm{cov}(u_i, v_j) = 0\,.$$

Fig.4.7 schematically illustrates the canonical correlation patterns.

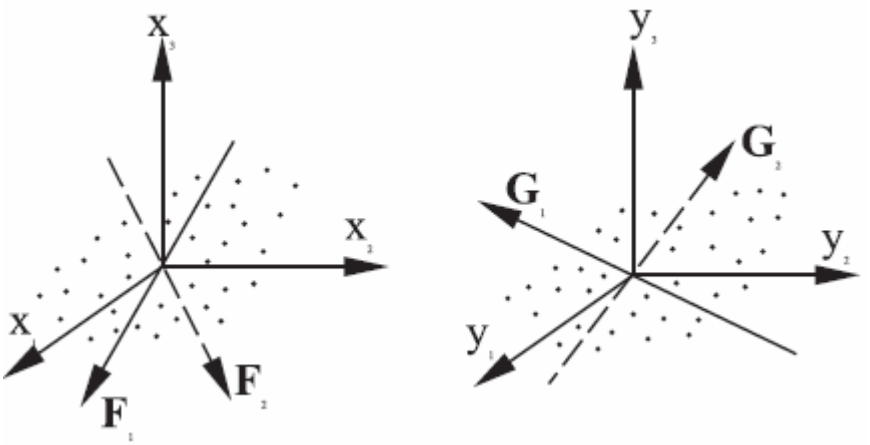


Fig4.7

3.2.2 Pre-filter with PCA

When x and y contain many variables, it is common to use PCA to pre-filter the data to reduce the dimensions of the datasets, i.e. apply PCA to x and y separately, extract the leading PCs, then apply CCA to the leading PCs of x and y.

Using Hotelling's choice of scaling for the PCAs (Eq 4.30), we express the PCA expansions as

$$\mathbf{x} = \sum_j a'_j \mathbf{e}'_j, \quad \mathbf{y} = \sum_j a''_j \mathbf{e}''_j\,. \tag{4.85}$$

CCA is then applied to

$$\tilde{\mathbf{x}} = [a'_1, \cdots, a'_{m_x}]^{\mathrm{T}}, \quad \tilde{\mathbf{y}} = [a''_1, \cdots, a''_{m_y}]^{\mathrm{T}}\,, \tag{4.86}$$

where only the first m_x and m_y modes are used. Another reason for using the PCA pre-filtering is that when the number of variables is not small relative to the number of samples, the CCA method may become unstable, as the many higher modes may by chance attain high correlation although they account for negligible variance.

With Hotelling's scaling

$$\mathrm{cov}(a'_j, a'_k) = \delta_{jk}, \quad \mathrm{cov}(a''_j, a''_k) = \delta_{jk}\,, \tag{4.87}$$

leading to

$$\mathbf{C}_{\tilde{x}\tilde{x}} = \mathbf{C}_{\tilde{y}\tilde{y}} = \mathbf{I}\,. \tag{4.88}$$

Eqs. (4.75) and (4.76) simplify to

$$\mathbf{C}_{\tilde{x}\tilde{y}}\mathbf{C}^{\mathrm{T}}_{\tilde{x}\tilde{y}}\,\mathbf{f} \equiv \mathbf{M}_f\mathbf{f} = \lambda\mathbf{f}\,, \tag{4.89}$$

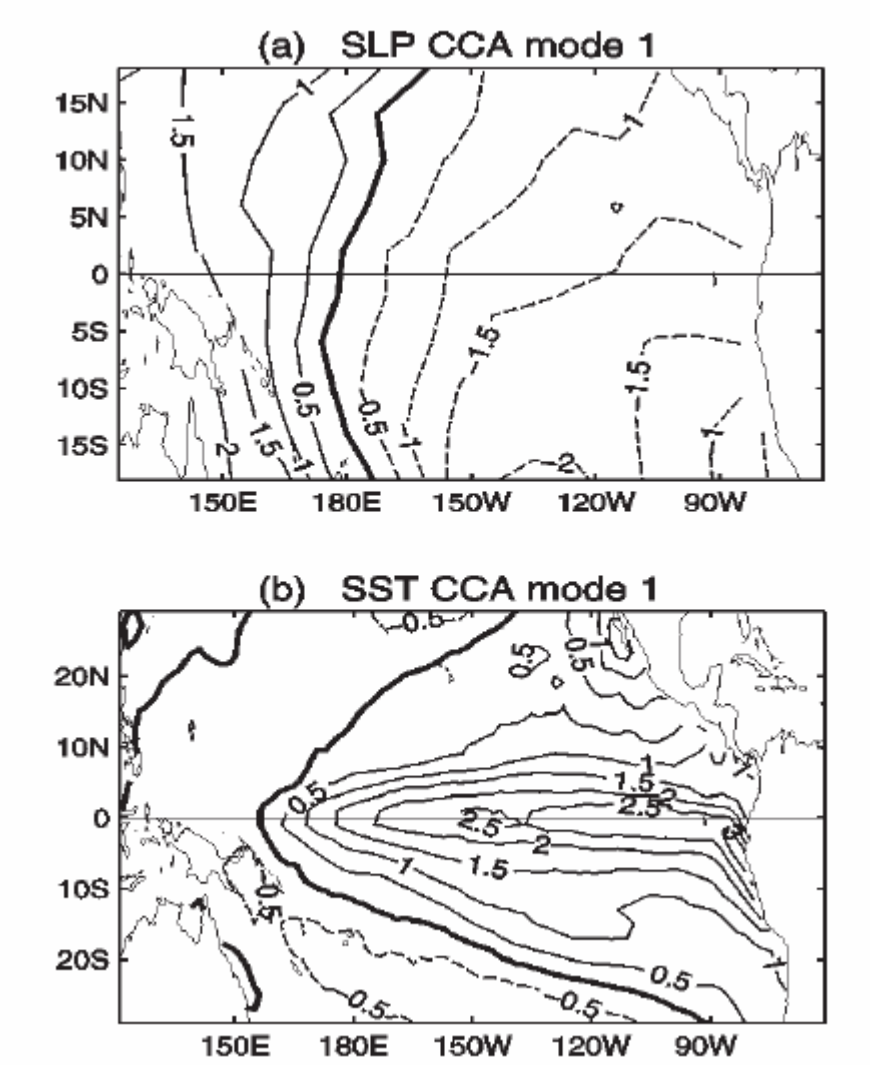
$$\mathbf{C}^{\mathrm{T}}_{\tilde{x}\tilde{y}}\mathbf{C}_{\tilde{x}\tilde{y}}\,\mathbf{g} \equiv \mathbf{M}_g\mathbf{g} = \lambda\mathbf{g}\,. \tag{4.90}$$

As $\boldsymbol{M}_f$ and $\boldsymbol{M}_g$ are non-negative definite symmetric matrices, the eigenvectors $\{\mathbf{f}_j\}$ $\{\mathbf{g}_j\}$ are now sets of orthogonal vectors. Eqs (4.83) and (4.84) simplify to

$$\mathbf{F} = \mathcal{F}, \quad \mathbf{G} = \mathcal{G}\,. \tag{4.91}$$

Hence $\{\mathbf{F}_j\}$ and $\{\mathbf{G}_j\}$ are also two sets of orthogonal vectors, and are identical to $\{\mathbf{f}_j\}$ and $\{\mathbf{g}_j\}$, respectively. Because of these nice propoerties, pre-filtering by PCA (with the Hotelling scaling) is recommended when $\mathbf{x}$ and $\mathbf{y}$ have many variables (relative to the number of samples). However, the orthogonality only holds in the reduced dimensional spaces, $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. If transformed into the original space $\mathbf{x}$ and $\mathbf{y}$, $\{\mathbf{F}_j\}$ and $\{\mathbf{G}_j\}$ are in general not two sets of orthogonal vectors.

Fig4.8 shows the mode 1 CCA of the tropical Pacific sea level pressure (SLP) field and the SST field, showing clearly the Southern Oscillation pattern in the SLP and the El Niño/La Niña pattern in the SST. The canonical variates u and v (not shown) fluctuate with time, both attaining high values during El Niño, low values during La Niña, and neutral values around zero during normal conditions.



3.2.3 Singular value decomposition (SVD)

Instead of maximizing the correlation as in CCA, one can maximize the covariance between two datasets. This alternative method is often called the singular value decomposition (SVD).

SVD is identical to CCA except that it maximize the covariance instead of the correlation. As mentioned before, CCA can be somewhat unstable (when the number of variables is not small relative to the number of samples) in that modes with high correlation may account for negligible variance, hence the recommended pre-filtering of data by PCA before applying CCA. SVD, by using covariance instead of correlation, does not have the unstable nature of the CCA, and does not need the pre-filtering by PCA.

In SVD, one simply performs SVD on the data covariance matrix $\boldsymbol{C}_{xy}$

$$\mathbf{C}_{xy} = \mathbf{USV}^{\mathrm{T}}, \tag{4.92}$$

Where the matrix U contains the left singular vectors $\boldsymbol{f}_i$, V the right singular vectors $\boldsymbol{g}_i$, and S the singular values. Maximum covariance between $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ is attained (Bretherton et al. 1992)

$$u_i = \mathbf{f}_i^{\mathrm{T}}\mathbf{x}, \quad v_i = \mathbf{g}_i^{\mathrm{T}}\mathbf{y}. \tag{4.93}$$

The inverse transform is given by

$$\mathbf{x} = \sum_i u_i \mathbf{f}_i, \quad \mathbf{y} = \sum_i v_i \mathbf{g}_i. \tag{4.94}$$

For most application, SVD yields rather similar results to the CCA (with CCA pre-filtering).

The matrix technique SVD can also be used to solve the CCA problem. Similar to (4.63), we seek

$$u = \mathbf{f}^{\mathrm{T}}\mathbf{x}, \quad v = \mathbf{g}^{\mathrm{T}}\mathbf{y}. \tag{4.95}$$

such that

$$\mathrm{cov}(\boldsymbol{u},\boldsymbol{v}) = \mathrm{cov}(\boldsymbol{f}^T\boldsymbol{x}, \boldsymbol{g}^T\boldsymbol{y}) = \boldsymbol{f}^T \mathrm{cov}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{g} \rightarrow \text{maximum} \tag{4.96}$$

subject to

$$\boldsymbol{f}^T\boldsymbol{f} = 1 \quad \text{and} \quad \boldsymbol{g}^T\boldsymbol{g} = 1 \tag{4.97}$$

The solution is obtained by using Lagrange multiplier

$$\boldsymbol{L} = \boldsymbol{f}^T \mathrm{cov}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{g} + \alpha(\boldsymbol{f}^T\boldsymbol{f} - 1) + \beta(\boldsymbol{g}^T\boldsymbol{g} - 1)$$

Similar to CCA (4.72)-(4.76), we have

$$\boldsymbol{C}_{xy}\boldsymbol{g} + 2\alpha\boldsymbol{f} = 0$$

$$\boldsymbol{C}_{xy}^{T}\boldsymbol{f} + 2\beta\boldsymbol{g} = 0 \tag{4.98}$$

$$\begin{aligned}\boldsymbol{C}_{xy}\boldsymbol{g} &= \mathrm{s}_x\boldsymbol{f}\\ \boldsymbol{C}_{xy}^{T}\boldsymbol{f} &= \mathrm{s}_y\boldsymbol{g}\end{aligned} \tag{4.99}$$

(4.99) implies that the solution g and can be solved by a singular value decomposition. The same solution is obtained by substituting the two equations into each other to obtain

$$\begin{aligned}\boldsymbol{C}_{xy}^{T}\boldsymbol{C}_{xy}\boldsymbol{g} &= 4\alpha\beta\boldsymbol{g} = \lambda\boldsymbol{g}\\ \boldsymbol{C}_{xy}\boldsymbol{C}_{xy}^{T}\boldsymbol{f} &= 4\alpha\beta\boldsymbol{f} = \lambda\boldsymbol{f}\end{aligned} \tag{4.100}$$

(4.50) is very similar to (4.75) or (4.76). So, the solution can be obtained by eigen equations.

So, there are two approaches to perform SVD: (1) simply perform SVD on $\boldsymbol{C}_{xy} = \mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})$; (2) solving eigen equations (4.100)