Chapter 5 Extended linear multivariate analysis.

In the last chapter, we have introduced the linear multivariate techniques based on PCA (EOF) for extracting features or recognizing patterns in a dataset or in two datasets. The EOF is capable of using few leading modes to describe the dominant structure that explains the majority of overall variances. However, EOF only depicts stationary modes and fails to explore the propagation features. In this chapter, we will introduce some methods which are able to extract the propagation features.

5.1 Extended EOF (EEOF) analysis

EEOF is similar to EOF in methodology but uses an extended matrix to compute the covariance. The extended matrix is constructed by the raw data matrix plus a series of time-lagged data matrix. Denote by $\boldsymbol{Y}$ a $\boldsymbol{m} \times \boldsymbol{n}$ 2-dimensional data matrix with m for the number of spatial grids and n for the length, i.e.

$$\boldsymbol{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,\boldsymbol{n}} \\ \cdots & \cdots & \cdots \\ y_{m,1} & \cdots & y_{m,n} \end{bmatrix} \tag{5.1}$$

With the matrix $\boldsymbol{Y}$, lagged copies of the $\boldsymbol{Y}$ are stacked to form the augmented data matrix $\boldsymbol{YY}$

$$\boldsymbol{YY} = \begin{vmatrix} \boldsymbol{Y}_0 \\ \boldsymbol{Y}_1 \\ \cdots \\ \boldsymbol{Y}_k \\ \cdots \\ \boldsymbol{Y}_L \end{vmatrix} \tag{5.2}$$

Where

$$\boldsymbol{Y}_{(0)} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,\boldsymbol{n}-\boldsymbol{L}+1} \\ \cdots & \cdots & \cdots \\ y_{m,1} & \cdots & y_{m,n\text{-}L+1} \end{bmatrix}$$

$$\boldsymbol{Y}_{(1)} = \begin{bmatrix} y_{1,2} & \cdots & y_{1,\boldsymbol{n}-\boldsymbol{L}+2} \\ \cdots & \cdots & \cdots \\ y_{m,2} & \cdots & y_{m,n\text{-}L+2} \end{bmatrix}$$

……

$$\boldsymbol{Y}_{(\boldsymbol{k})} = \begin{bmatrix} y_{1,\boldsymbol{k}} & \cdots & y_{1,\boldsymbol{n}-\boldsymbol{L}+\boldsymbol{k}} \\ \cdots & \cdots & \cdots \\ y_{m,k} & \cdots & y_{m,n\text{-}L+k} \end{bmatrix}$$

$$\boldsymbol{Y}_{(\boldsymbol{L})} = \begin{bmatrix} y_{1,\boldsymbol{L}} & \cdots & y_{1,\boldsymbol{n}} \\ \cdots & \cdots & \cdots \\ y_{m,L} & \cdots & y_{m,n} \end{bmatrix}$$

$\boldsymbol{YY}$ is expanded to a $(\boldsymbol{m}\times(\boldsymbol{L}+1),\boldsymbol{n}-\boldsymbol{L}+1)$ 2-D matrix. Each eigenvector of EOF for $\boldsymbol{YY}$ contains m*(L+1) elements, which can characterize the propagation features.

t = 0 t=k t=L

Therefore, EEOF shows not only eigenvectors but also the temporal evolutions of the eigenvectors.

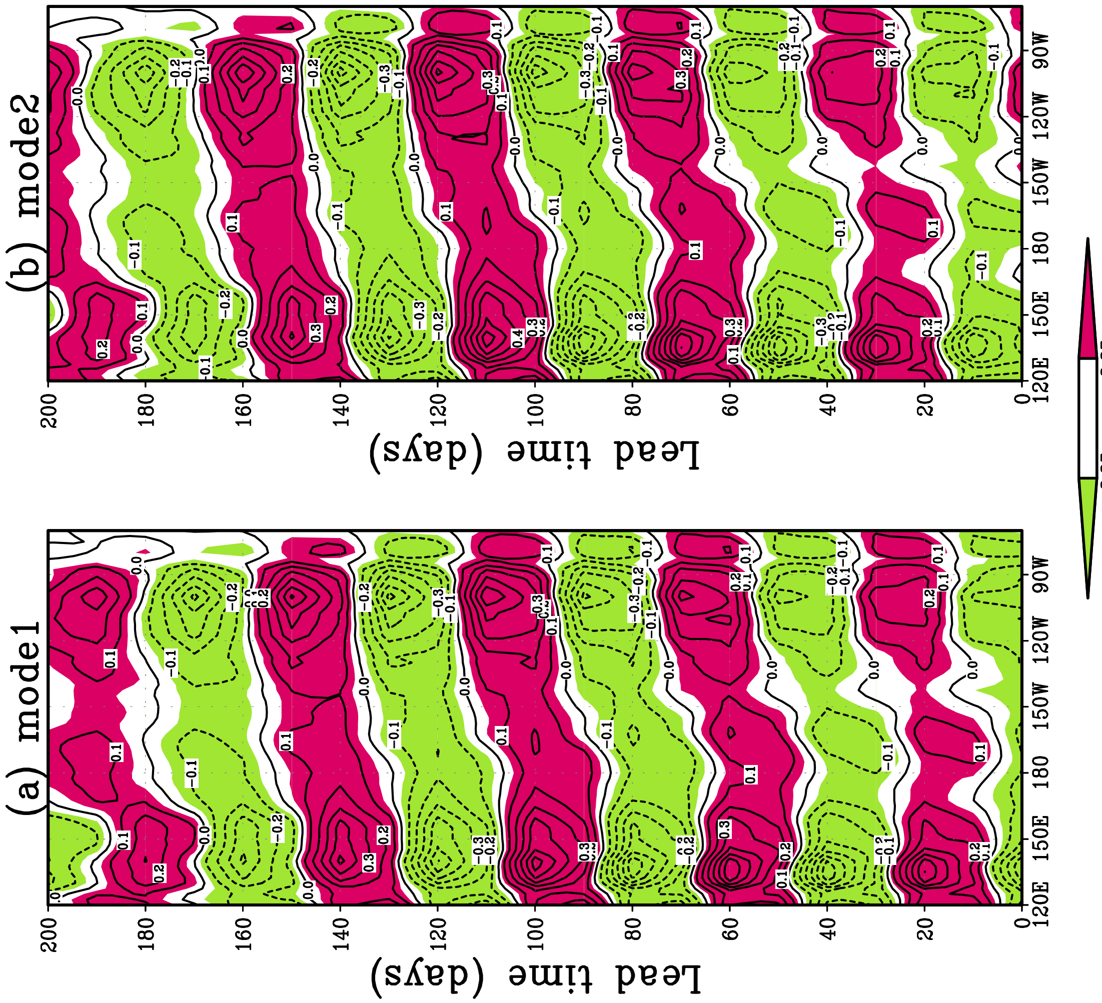

Fig.5.1: The first two EEOF modes for zonal wind as a function of leading times along the equator (Tang et al 2006).

The other example

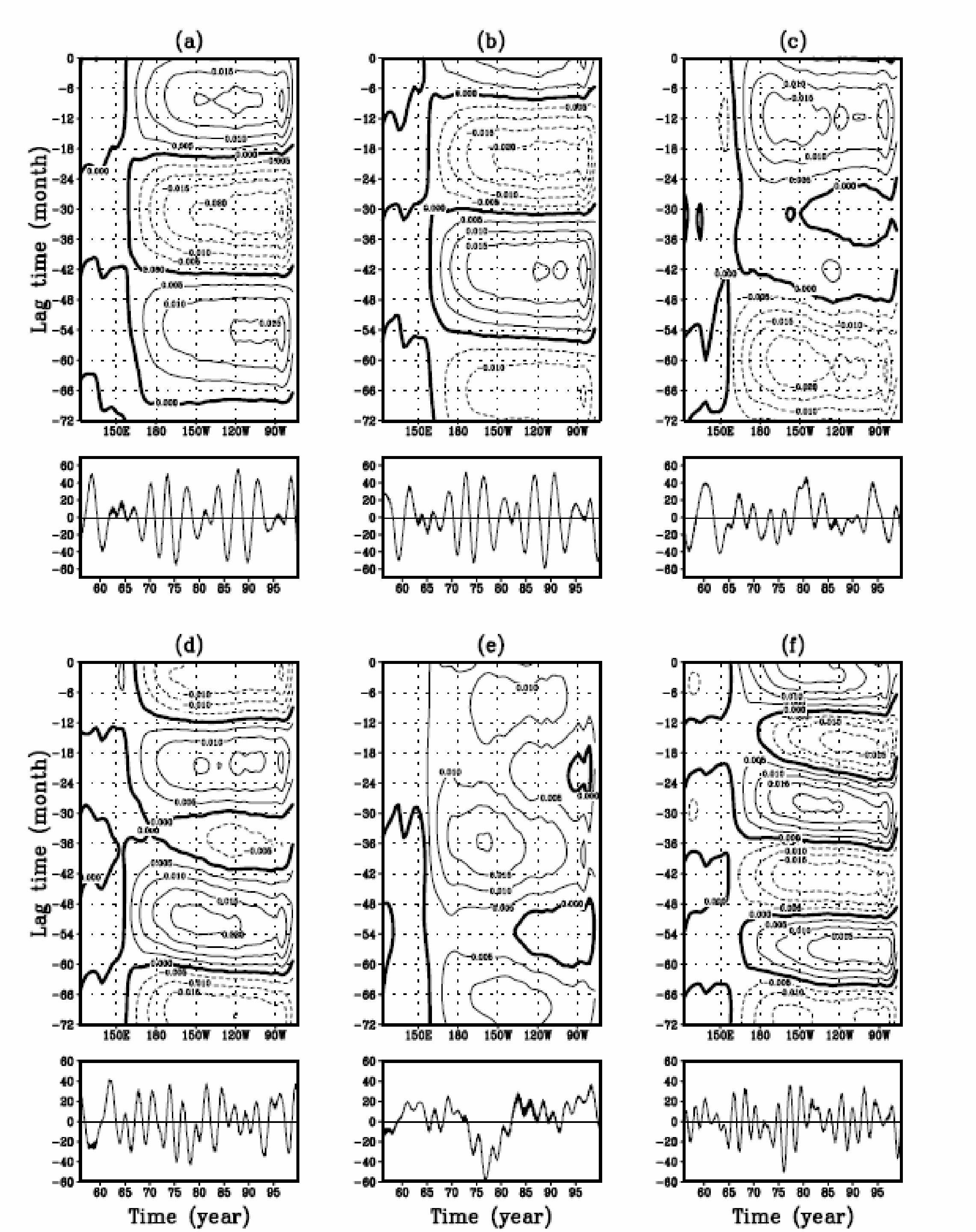
(a)
(b)
(c)
(d)
(e)
(f)
Lag time (month)
Time (year)
150E
180
150W
120W
90W
60
65
70
75
80
85
90
95

Fig. 5.1.1

The EEOF (SSA) modes 1-6 for the tropical Pacific SSTA shown in (a)-(f), respectively. The contour plots display the space-time eigenvectors (loading patterns), showing the SSTA along the equator as a function of the lag. Solid contours indicate positive anomalies and dashed contours, negative anomalies, with the zero contour indicated by the thick solid curve. In a separate panel beneath each contour plot, the principal component (PC) of each SSA mode is also plotted as a time series, (where each tick mark on the abscissa indicates the start of a year). The time of the PC is synchronized to the lag time of 0 month in the space-time eigenvector.

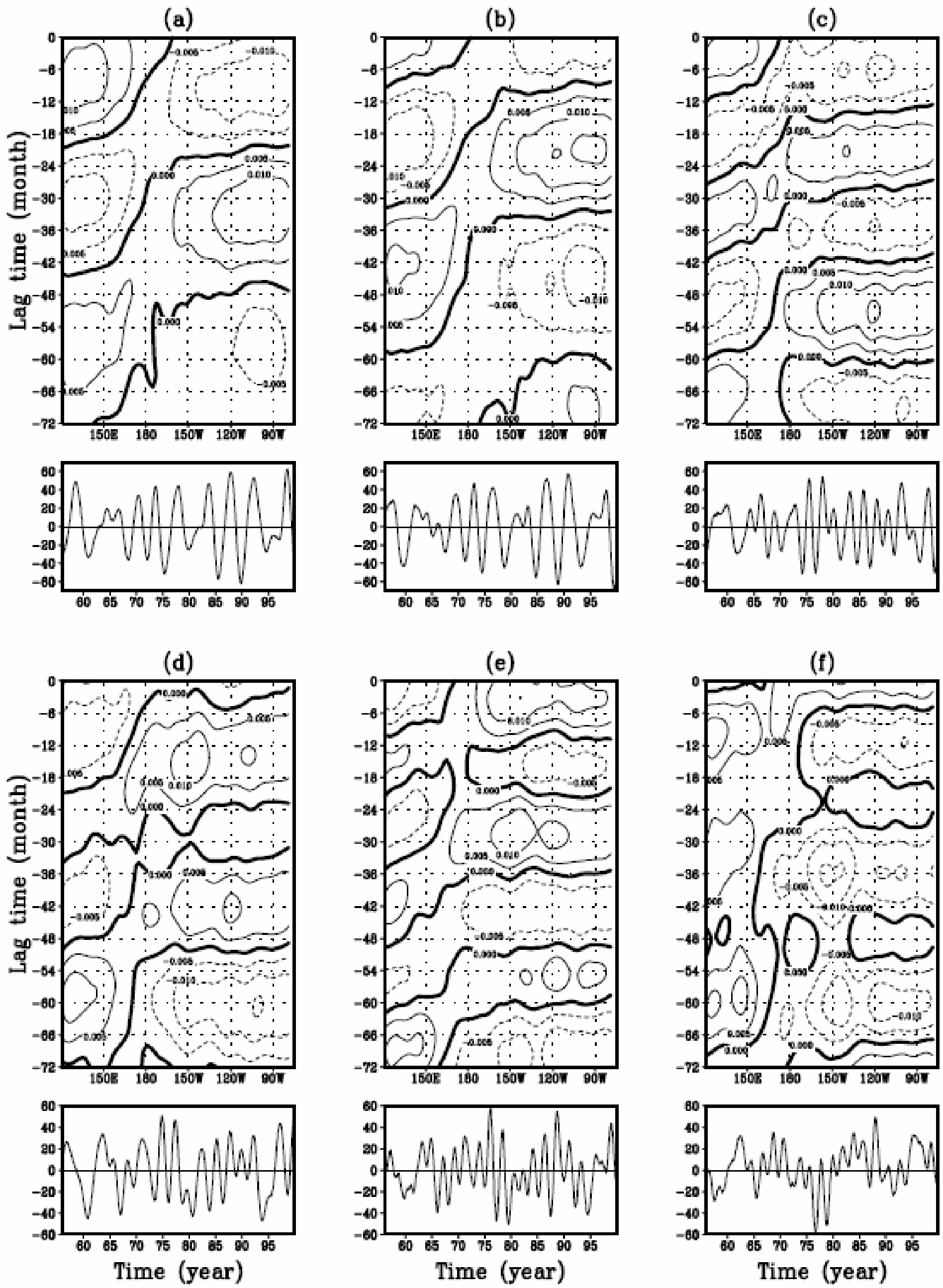

Fig5.1.2
The EEOF(SSA) modes 1-6 for the tropical Pacific SLPA shown in (a)-(f), respectively. The contour plots display the space-time eigenvectors (loading patterns), showing the SLPA along the equator as a function of the lag. The PC of each SSA mode is also plotted as a time series, beneath each contour plot.
(Hiseh and Wu 2002)

5.2 Complex EOF (CEOF)

EEOF accounts for patterns that evolve in time by representing a series of eigenmode as a function of leading time, being able to describe propagating and oscillatory behavior. A generalization of this approach is to model not only the 'state' $\vec{x}_t$ but also an indicator of its tendency $\delta\vec{x}_t$. How to represent the information of $\delta\vec{x}_t$? It has been proven that the Hilbert transform $\vec{x}_t^{\,H}$ is a reasonable measure of $\delta\vec{x}_t$ when variations in $\vec{x}_t^{\,H}$ are confined to a relatively narrow time scale (van Storch and Zwiers 1999). Thus the conventional eigentechniques, EOFs, are applied to the complexified time series $\vec{x}_t + i\vec{x}_t^{\,H}$, called CEOF.

According to the theory of Hilbert transform,

$$\vec{x}_t^{\,H} = \sum_{k=-L}^{L} \vec{x}_t(\mathrm{t}-\mathrm{k})\mathrm{h}(\mathrm{k}) \qquad (5.3)$$

where

$$h(k) = \begin{cases} \dfrac{2}{k\pi}\sin^2(k\pi/2)........if..k \neq 0 \\ 0..............................if...k = 0 \end{cases} \qquad (5.4)$$

Ideally $L = \infty$ in (5.4), but in practice $L = 7 - 15$.
Two important measures are often used in CEOF: amplitude function $\boldsymbol{R}$ and

phase function θ. The former represents the anomalous amplitude of an eigenvector whereas the latter depicts its propagation. Denote by $\boldsymbol{B}$ an eigenvector, $\boldsymbol{R}$ and θ are respectively defined as below

$$\boldsymbol{R} = \boldsymbol{B}\boldsymbol{B}^*$$

$$\theta = \arctan[\frac{\mathrm{Im}(\boldsymbol{B})}{\mathrm{Re}(\boldsymbol{B})}]$$

where $\boldsymbol{B}^*$ is conjugate of $\boldsymbol{B}^*$.

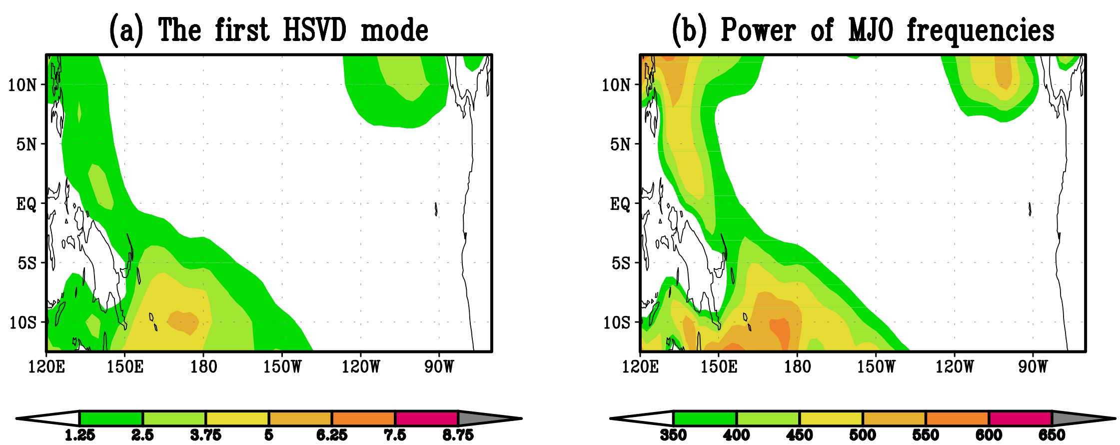

Fig.5.2 (a) Spatial amplitude function of the first CEOF mode, accounting for 31% of total variance; (b) Power of the MJO frequencies (30 to 90 d^{-1}). The units are m^2s^{-2} in both (a) and (b). (Tang 2006) .

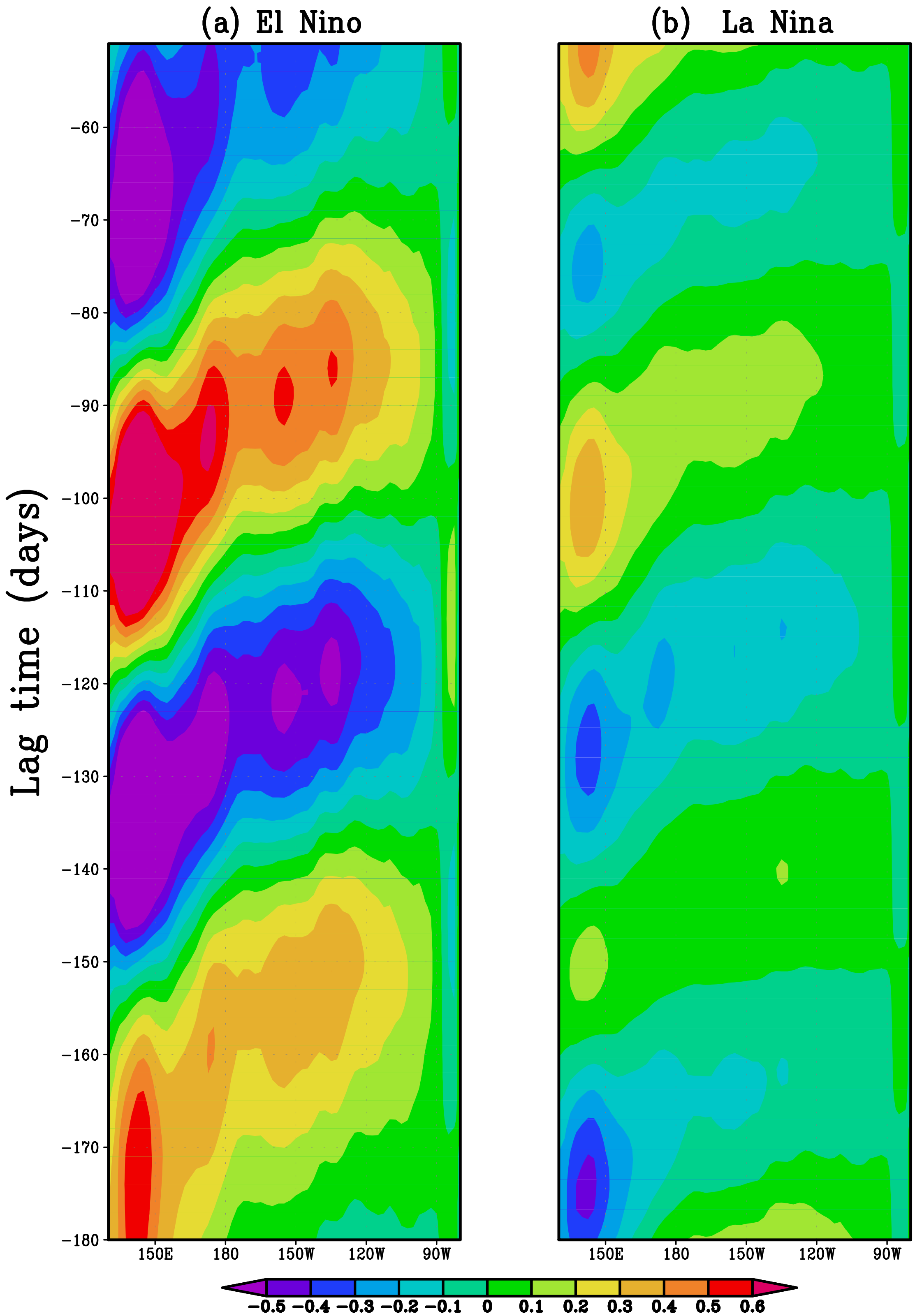
(a) El Nino
(b) La Nina
Lag time (days)
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
150E
180
150W
120W
90W
-0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5 0.6

Fig.5.3 Composite MJO signals, obtained from the recovered field using the first CEOF mode, for the three strongest El Nino (1982/83, 1987/88, 1997/98) and three strongest La Nina events (1983/84, 1988/89, 1998/99) along the equator for the period 50 to 180 d prior to the time of the ENSO peak (Tang 2006).

5.3 Singular spectrum analysis

The PCA (EOF) involves finding eigenvectors containing spatial information. Is it possible to use PCA approach to incorporate time information into eigenvectors?

Given a time series $y_j = y(t_j)$ (j=1,..,n), a new matrix can be contracted

$$\mathbf{Y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{n-L+1} \\ y_2 & y_3 & \cdots & y_{n-L+2} \\ \vdots & \vdots & \vdots & \vdots \\ y_L & y_{L+1} & \cdots & y_n \end{bmatrix}. \tag{5.5}$$

The number of lags L is usually taken to be at most ¼ of the total record length.

The standard PCA can be performed on Y, resulting in

$$\mathbf{y}^{(l)} = \mathbf{y}(t_l) = \sum_j a_j(t_l)\,\mathbf{e}_j\,, \tag{5.6}$$

This method is known as singular spectrum analysis (SSA), or time-PCA, where a_j is the *j*th principal component (PC), a time series of length n-L+1, and e_j is the *j*th eigenvector (or loading vector) of length L. Together, a_j and e_j, represent the *j*th SSA mode.

SSA has become popular in the field of dynamical system, in particular in the analysis of period and spectrum. As we talked before, the conventional spectrum analysis is based on the Fourier transforms, i.e., the Fourier spectral analysis. The Fourier analysis is generally restricted to sinusoidal shapes; whereas SSA is not restricted to using sinusoidal shaped function. So, SSA can in principal capture an anharmonic wave more efficiently than the Fourier method.

An example:

The *Southern Oscillation Index* (SOI) is defined as the normalized air pressure difference between Tahiti and Darwin. There are several slight variations in the SOI values calculated at various centres; here we use the SOI calculated by the Climate Research Unit at the University of East Anglia, based on the method of ***Ropelewski and Jones (1987). The SOI measures the seesaw oscillations of the sea level air pressure between the eastern and western equatorial Pacific. When the SOI is strongly negative, the eastern equatorial Pacific sea surface temperatures also become warm (i.e. an El Niño episode occurs); when SOI is strongly positive, the central equatorial Pacific becomes cool (i.e. a La Niña episode occurs). The SOI is known to have the main spectral peak at a period of about 4-5 years ***(Troup, 1965). For SSA, the window L needs to be long enough to accommodate this main spectral period, hence $L = 72$ months was chosen for the SSA.

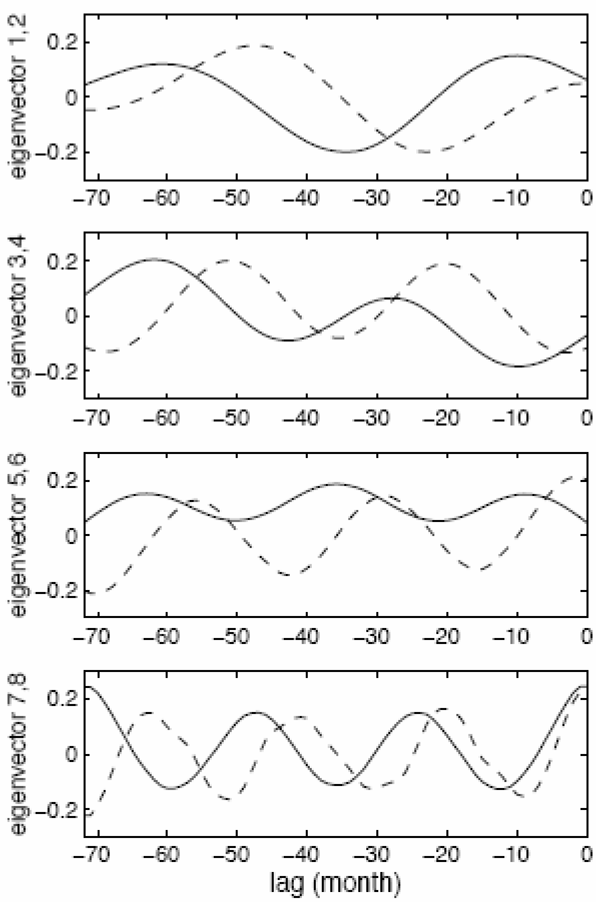

Fig.5.4

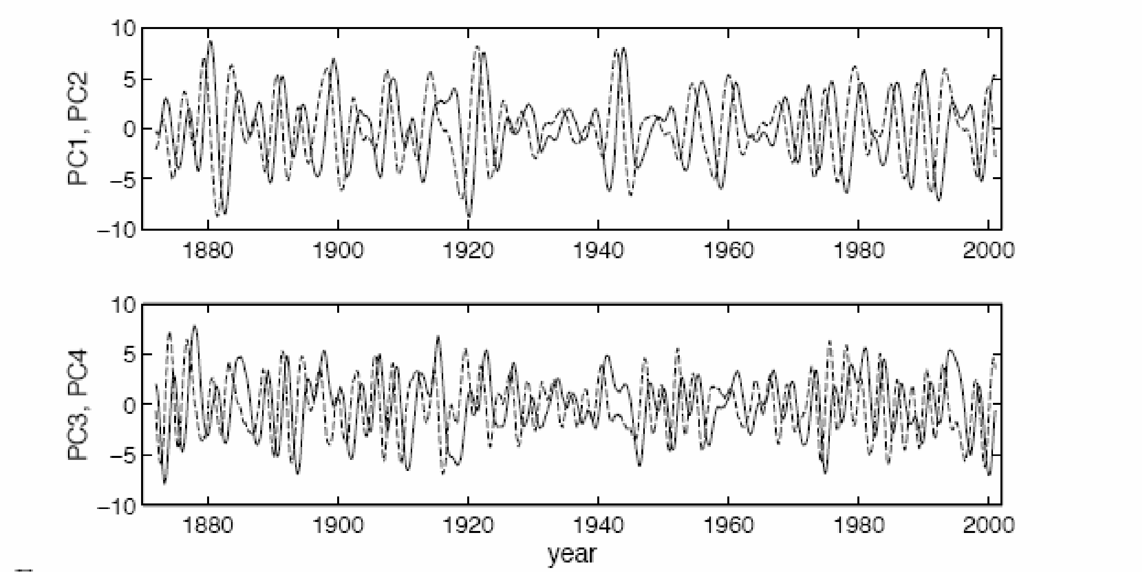

Fig. 5.5 (Hsieh and Wu 2002).

The variances explained by
SSA: 11.0%, 10.5%, 9.0%, 8.1%
Fourier analysis: 4.0% …..

Fig.5.4 and 5.5: (1) a period of 50 months in first eigenvector
(2) The oscillation displays anharmonic features.

5.4 Principal Oscillation Patterns

For some datasets containing multiple time series, one would like to find a low order linear dynamical system to account for the behavior of the data. The Principal Oscillation Pattern (POP) method, proposed by Hasselmann (1988), is one such technique.

Consider a simple system with two variables y_1 and y_2, obeying the linear dynamical equations

$$\frac{dy_1}{dt} = L_1(y_1, y_2)\,, \qquad \frac{dy_2}{dt} = L_2(y_1, y_2)\,, \tag{5.7}$$

Where L_1 and L_2 are linear functions. The discretized version of the dynamical equations is of the form

$$y_1(t+1) = a_{11}y_1(t) + a_{12}y_2(t)\,,$$
$$y_2(t+1) = a_{21}y_1(t) + a_{22}y_2(t)\,, \qquad (5.8)$$

Where the a_{ij} are parameters. For an m-variable first order linear dynamical system, the discretized governing equations can be expressed as

$$\mathbf{y}(t+1) = \mathbf{A}\mathbf{y}(t)\,, \qquad (5.9)$$

Where y is an m-element column vector, and A is an $m \times m$ matrix. A is a real matrix, but generally not symmetric; hence, its eigenvalues λ and eigenvectors P are in general complex. Taking the complex conjugate of eigenvector equation

$$\boldsymbol{AP} = \Lambda \boldsymbol{P} \qquad (5.10)$$

where P's jth column is simply the jth eigenvector p_j, i.e.,

$$\mathbf{P} = [\mathbf{p}_1|\mathbf{p}_2|\cdots|\mathbf{p}_m]\,, \qquad (5.11)$$

and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_m \end{bmatrix}, \qquad (5.12)$$

then

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_m \end{bmatrix}, \qquad (5.13)$$

Applying $\boldsymbol{P}^{-1}$ to (5.9) yields

$$\mathbf{P}^{-1}\mathbf{y}(t+1) = \mathbf{P}^{-1}\mathbf{A}\mathbf{y}(t) = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\ \mathbf{P}^{-1}\mathbf{y}(t)\,, \tag{5.14}$$

which can be expressed as

$$\mathbf{z}(t+1) = \Lambda\, \mathbf{z}(t)\,, \tag{5.15}$$

where

$$\mathbf{z} = \mathbf{P}^{-1}\mathbf{y}\,, \tag{5.16}$$

By applying $\boldsymbol{P}$ to (5.16), the inverse transform is obtained:

$$\mathbf{y}(t) = \mathbf{P}\,\mathbf{z}(t)\,, \quad \text{i.e.} \tag{5.17}$$

$$\mathbf{y}(t) = \sum_{i=1}^{m} \mathbf{p}_j\, z_j(t)\,. \tag{5.18}$$

The eigenvector

$$\mathbf{p}_j = \mathbf{p}_j^r + i\mathbf{p}_j^i \tag{5.19}$$

is called a Principal Oscillation Pattern (POP) of y(t). The corresponding POP coefficient

$$z_j(t) = z_j^r(t) + iz_j^i(t)\,, \tag{5.20}$$

obeys (5.15), i.e.,

$$z_j(t+1) = \lambda_j\, z_j(t)\,, \tag{5.21}$$

Where

$$\lambda_j = |\lambda_j| e^{i\theta_j} \equiv e^{-1/\tau_j}\, e^{i2\pi/T_j}\,, \tag{5.22}$$

as $|\lambda_j| < 1$ in the real world (von Storch and Zwiers, 1999, pp. 336-337). Hence

$$z_j(t+1) = e^{-1/\tau_j}\, e^{i2\pi/T_j}\, z_j(t)\,, \tag{5.23}$$

Where τ_j is an e-folding decay time scale, and T_j is the oscillatory period, i.e., z_j will evolve in time displaying exponential decay and oscillatory behavior, governed by the parameters τ_j and T_j, respectively. As y(t) is real, (5.17) gives:

$$\mathbf{y}(t) = \mathrm{Re}[\mathbf{P}\mathbf{z}(t)] = \sum_j [\mathbf{p}_j^r z_j^r(t) - \mathbf{p}_j^i z_j^i(t)]\,. \tag{5.24}$$

As t progresses, the sign of z_j^r and z_j^i oscillate, resulting in an evolving $P_j^r z_j^r - P_j^i z_j^r$ pattern:

$$z_j^r \quad : \quad \cdots \rightarrow + \rightarrow 0 \rightarrow - \rightarrow 0 \rightarrow + \rightarrow \cdots \tag{5.25}$$

$$z_j^i \quad : \quad \cdots \rightarrow 0 \rightarrow + \rightarrow 0 \rightarrow - \rightarrow 0 \rightarrow \cdots \tag{5.26}$$

$$\mathbf{p}_j^r z_j^r - \mathbf{p}_j^i z_j^i \quad : \quad \cdots \rightarrow \mathbf{p}_j^r \rightarrow -\mathbf{p}_j^i \rightarrow -\mathbf{p}_j^r \rightarrow \mathbf{p}_j^i \rightarrow \mathbf{p}_j^r \rightarrow \cdots\,. \tag{5.27}$$

Unlike PCA, POP allows the representation of propagating waves using only one POP mode. For instance, Fig. 5.6 shows a POP pattern representing wave propagating from right to left, while Fig. 5.7 represents a pair of eddies rotating around the centre of the figure.

For real data, noise ε must be added to the dynamical system,

$$\mathbf{y}(t+1) = \mathbf{A}\mathbf{y}(t) + \epsilon\,. \tag{5.28}$$

From $E[Eq(5.28)\boldsymbol{y}^T(\boldsymbol{t})]$ and the fact that $E[\varepsilon] = 0,$ one can estimate A from

$$\mathbf{A} = \mathrm{E}[\mathbf{y}(t+1)\,\mathbf{y}^{\mathrm{T}}(t)]\,\{\mathrm{E}[\mathbf{y}(t)\,\mathbf{y}^{\mathrm{T}}(t)]\}^{-1}\,. \tag{5.29}$$

From A, one can computer the eigenvectors $\boldsymbol{p}_j$, eigenvalues λ_j, and z_j

from (5.16).

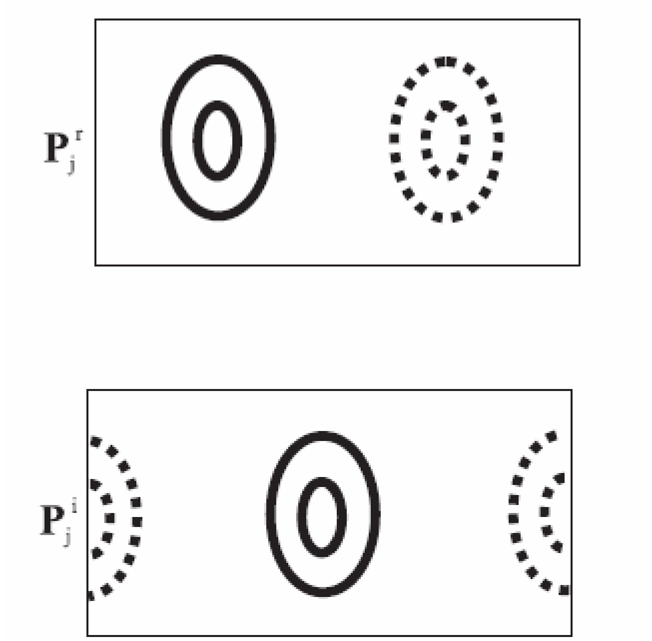

Fig.5.6

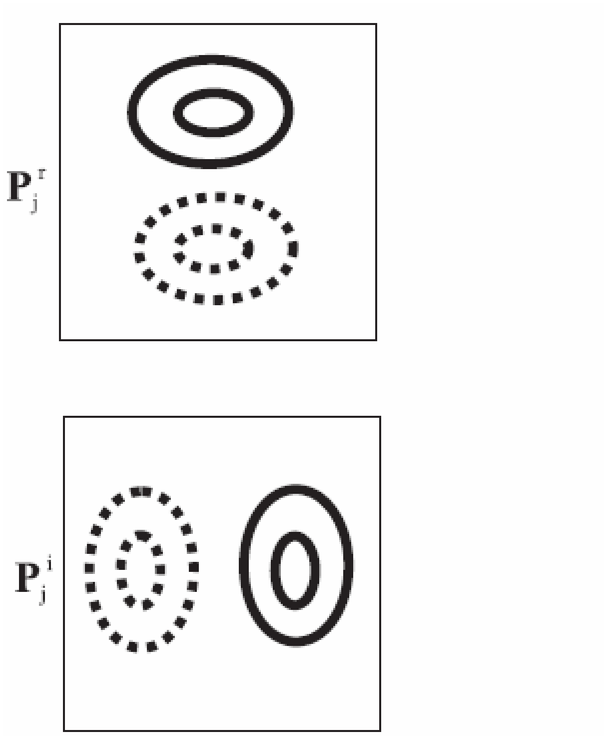

Fig 5.7

An example of POP analysis: