Chap.11 Nonlinear principal component analysis [Book, Chap. 10]

We have seen machine learning methods nonlinearly generalizing the linear regression method. Now we will examine ways to nonlinearly generalize principal component analysis (PCA). Fig. below illustrates the difference between (a) linear regression, (b) PCA, (c) nonlinear regression, and (d) nonlinear PCA.

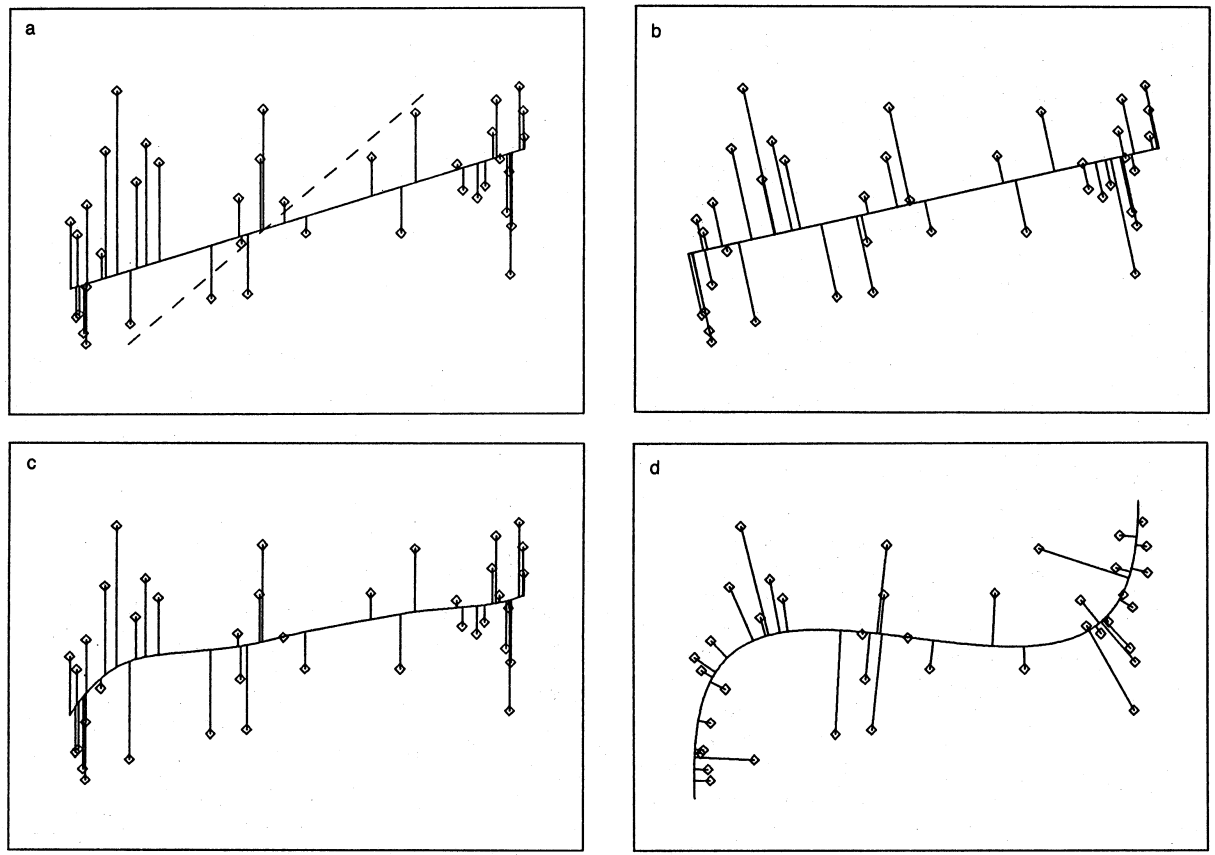

Figure : (a) The linear regression line minimizes the mean squared error (MSE) in the response variable. The dashed line illustrates the dramatically different result when the role of the predictor variable and the response variable are reversed. (b) PCA minimizes the MSE in all variables. (c) Nonlinear regression methods produce a curve minimizing

the MSE in the response variable. (d) Nonlinear PCA methods use a curve which minimize the MSE of all variables. In both (c) and (d), the smoothness of the curve can be varied by the method. [Reproduced from Hastie and Stuetzle (1989)].

Nonlinear PCA can be performed by a variety of methods, e.g. the auto-associative NN model using multi-layer perceptrons (MLP) (Kramer, 1991; Hsieh, 2001, 2007), and the kernel PCA model (Schölkopf et al., 1998).

Nonlinear PCA belongs to the class of nonlinear dimensionality reduction techniques, which also includes principal curves (Hastie and Stuetzle, 1989), locally linear embedding (LLE) (Roweis and Saul, 2000) and isomap (Tenenbaum et al., 2000).

Self-organizing map (SOM) (Kohonen, 1982) can also be regarded as a *discrete* version of NLPCA.

11.1 Auto-associative neural networks for nonlinear PCA [Book, Sect.10.1]

Open curves [Book, Sect.10.1.1]

Kramer (1991) proposed a neural-network based nonlinear PCA (NLPCA) model where the straight line solution in PCA is replaced by a continuous open curve for approximating the data.

The fundamental difference between NLPCA and PCA is that PCA only allows a linear mapping ($u = \mathbf{e} \cdot \mathbf{x}$) between $\mathbf{x}$ and the PC u, while NLPCA allows a nonlinear mapping.

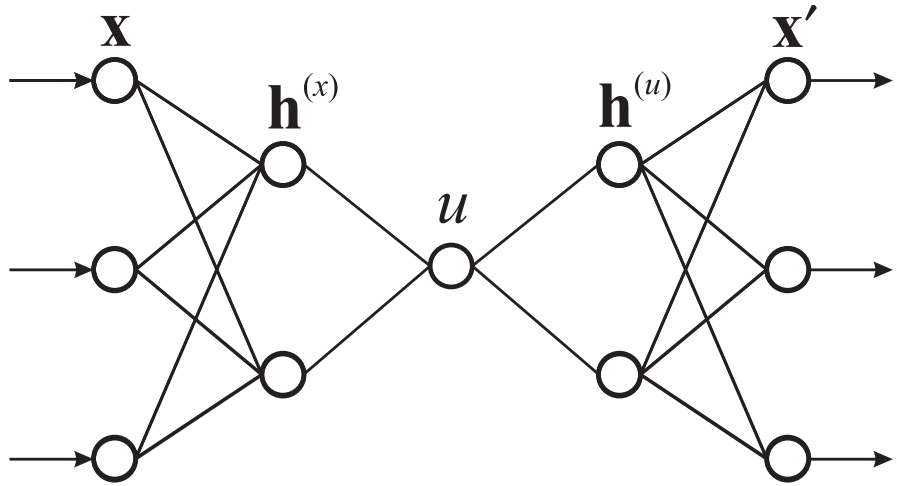

Figure : A schematic diagram of the NN model for calculating the NLPCA. There are 3 layers of hidden neurons sandwiched between the input layer $\mathbf{x}$ on the left and the output layer $\mathbf{x}'$ on the right. Next to the input layer is the encoding layer, followed by the 'bottleneck' layer (with a single neuron u), which is then followed by the decoding layer. A nonlinear function maps from the higher dimension input space to the 1-dimension bottleneck space, followed by an inverse transform mapping from the bottleneck space back to the original space represented by the outputs, which are to be as close to the inputs as possible by minimizing the objective function $J = \langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle$, where $\langle \ldots \rangle$ denotes calculating

the average over all the data points. Data compression is achieved by the bottleneck, with the bottleneck neuron giving u, the nonlinear principal component (NLPC).

One can view the NLPCA network as composed of two standard 2-layer MLP NNs placed one after the other. The first 2-layer network maps from the inputs $\mathbf{x}$ through a hidden layer to the bottleneck layer, here with only one neuron u, i.e. a nonlinear mapping $u = f(\mathbf{x})$.

The next 2-layer MLP NN inversely maps from the nonlinear PC (NLPC) u back to the original higher dimensional $\mathbf{x}$-space, with the objective that the outputs $\mathbf{x}' = \mathbf{g}(u)$ be as close as possible to the inputs $\mathbf{x}$, where $\mathbf{g}(u)$ nonlinearly generates a curve in the $\mathbf{x}$-space, hence a 1-dimensional approximation of the original data.

Because the target data for the output neurons $\mathbf{x}'$ are simply the input data $\mathbf{x}$, such networks are called auto-associative NNs.

Squeezing the input information through a bottleneck layer (here with only one neuron) accomplishes the dimensional reduction.

NLPCA of sea surface temperature anomalies [Book, Sect. 10.1.2]

The tropical Pacific SST anomaly (SSTA) data (1950-1999) (i.e. the SST data with the climatological seasonal cycle removed) were pre-filtered by PCA, with only the 3 leading modes retained (Hsieh, 2001). PCA modes 1, 2 and 3 accounted for 51.4%, 10.1% and 7.2%, respectively, of the variance in the SSTA data.

Due to the large number of spatially gridded variables, NLPCA could not be applied directly to the SSTA time series, as this would lead to a huge NN with the number of model parameters vastly exceeding the number of observations. Instead, the first 3 PCs (PC1, PC2 and PC3) were used as the input $\mathbf{x}$ for the NLPCA network.

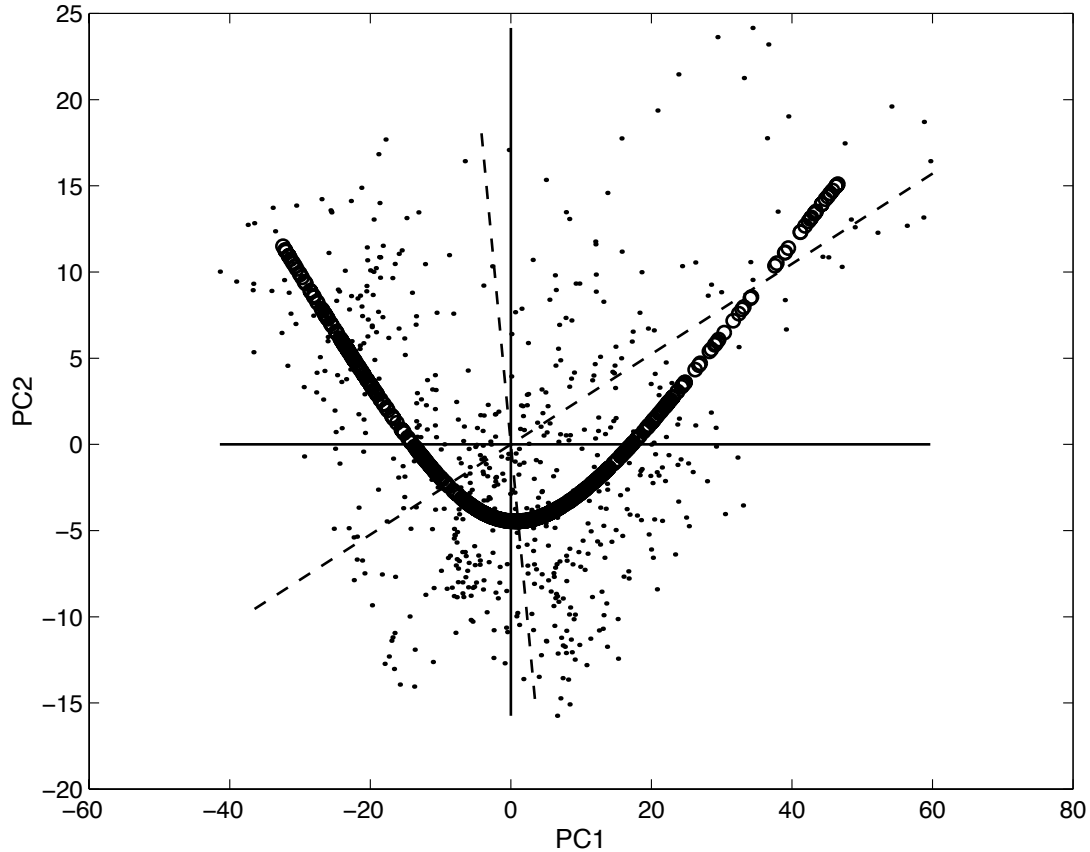

Figure : Scatter plot of the SST anomaly (SSTA) data (shown as dots) in the PC1-PC2 plane, with the El Niño states lying in the upper right corner, and the La Niña states in the upper left corner. The PC2 axis is stretched relative to the PC1 axis for better visualization. The first mode NLPCA approximation to the data is shown by the (overlapping) small

circles, which traced out a U-shaped curve. The first PCA eigenvector is oriented along the horizontal line, and the second PCA, by the vertical line. The varimax method rotates the two PCA eigenvectors in a counterclockwise direction, as the rotated PCA (RPCA) eigenvectors are oriented along the dashed lines.

In terms of variance explained, the first NLPCA mode explained 56.6% of the variance, versus 51.4% by the first PCA mode, and 47.2% by the first RPCA mode.

With the NLPCA, for a given value of the NLPC u, one can map from u to the 3 PCs. This is done by assigning the value u to the bottleneck neuron and mapping forward using the second half of the network. Each of the 3 PCs can be multiplied by its associated PCA

(spatial) eigenvector, and the three added together to yield the spatial pattern for that particular value of u.

Unlike PCA which gives the same spatial anomaly pattern except for changes in the amplitude as the PC varies, the NLPCA spatial pattern generally varies continuously as the NLPC changes.

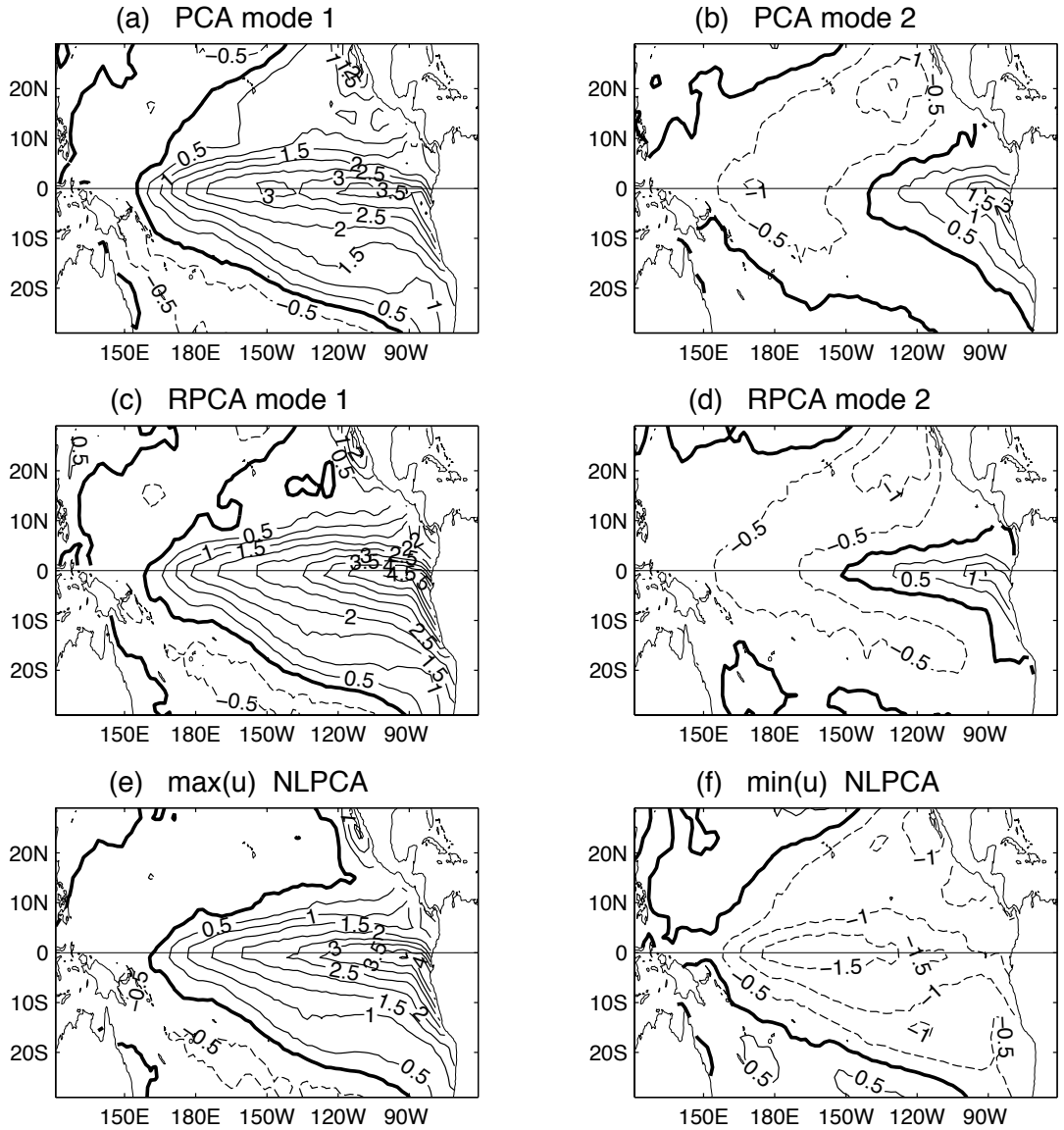
(a) PCA mode 1
(b) PCA mode 2
(c) RPCA mode 1
(d) RPCA mode 2
(e) max(u) NLPCA
(f) min(u) NLPCA
20N
10N
0
10S
20S
150E 180E 150W 120W 90W

Figure : The SSTA patterns (in °C) of the PCA, RPCA and the NLPCA. The first and second PCA spatial modes are shown in (a) and (b) respectively, (both with their corresponding PCs at maximum value). The first and second varimax RPCA spatial modes are shown in (c) and (d) respectively, (both with their corresponding RPCs at maximum value). The anomaly pattern as the NLPC u of the first NLPCA mode varies from (e) maximum (strong El Niño) to (f) its minimum (strong La Niña). With a contour interval of 0.5°C , the positive contours are shown as solid curves, negative contours, dashed curves, and the zero contour, a thick curve.

Clearly the asymmetry between El Niño and La Niña, i.e. the cool anomalies during La Niña episodes (Fig. (f)) are observed to centre much further west of the warm anomalies during El Niño (Fig. (e)), is well captured by the first NLPCA mode.

With a linear approach, it is generally impossible to have a solution simultaneously (a) explaining maximum global variance of the dataset and (b) approaching local data clusters, hence the dichotomy between PCA and RPCA, with PCA aiming for (a) and RPCA for (b).

With the more flexible NLPCA method, both objectives (a) and (b) may be attained together, thus the nonlinearity in NLPCA unifies the PCA and RPCA approaches.

The tropical Pacific SST example illustrates that with a complicated oscillation like the El Niño-La Niña phenomenon, using a linear method such as PCA results in the nonlinear mode being scattered into several linear modes (in fact, all 3 leading PCA modes are related to this phenomenon).

Closed curves [Book, Sect. 10.1.4]

Many phenomena involving waves or quasi-periodic fluctuations call for a continuous closed curve solution for NLPCA.

Kirby and Miranda (1996) introduced an NLPCA with a circular node at the network bottleneck [henceforth referred to as the NLPCA(cir)], so that the nonlinear principal component (NLPC) as represented by the circular node is an angular variable θ, and the

NLPCA(cir) is capable of approximating the data by a closed continuous curve.

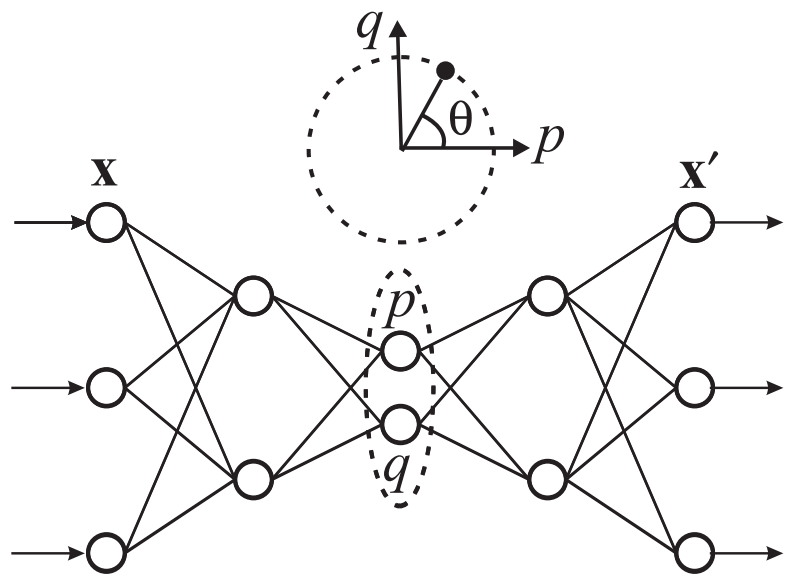

Figure : Schematic diagram of the NN model used for NLPCA with a circular node at the bottleneck (NLPCA(cir)). Instead of having one bottleneck neuron u, there are two neurons p and q constrained to lie on a unit circle in the p-q plane, so there is only one free angular variable θ, the NLPC. This network is for extracting a closed curve solution.

For an application of NLPCA(cir), consider the Quasi-Biennial Oscillation (QBO), which dominates over the annual cycle or other variations in the equatorial stratosphere, with the period of oscillation varying roughly between 22 and 32 months.

Average zonal (i.e. the westerly component of the) winds at 70, 50, 40, 30, 20, 15 and 10 hPa (i.e. from about 20 km to 30 km altitude) during 1956-2006 were studied.

After the 51-year means were removed, the zonal wind anomalies U at 7 vertical levels in the stratosphere became the 7 inputs to the NLPCA(cir) network.

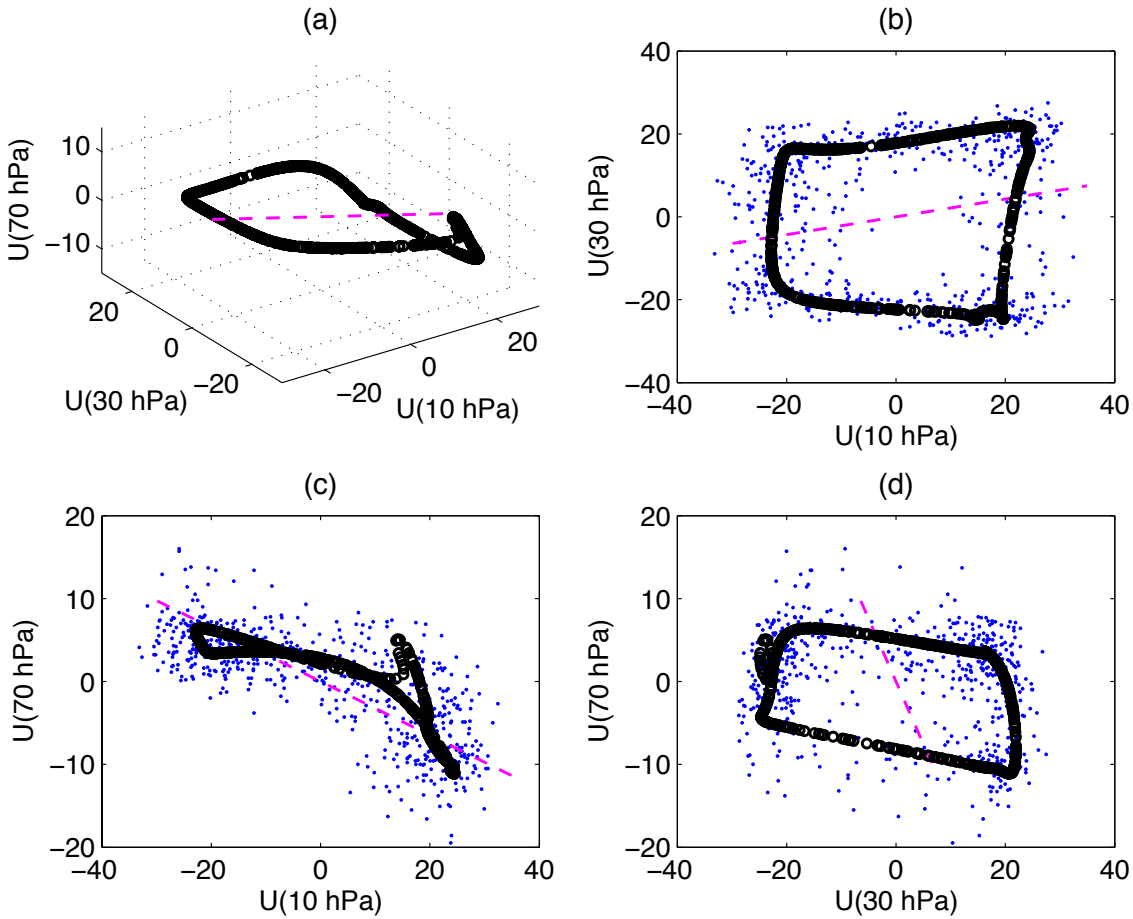

Figure : The NLPCA(cir) mode 1 solution for the equatorial stratospheric zonal wind anomalies. For comparison, the PCA mode 1 solution is shown by the dashed line. Only 3 out of 7 dimensions are shown, namely

the zonal velocity anomaly U at the top, middle and bottom levels (10, 30 and 70 hPa). Panel (a) gives a 3-D view, while (b)-(d) give 2-D views.

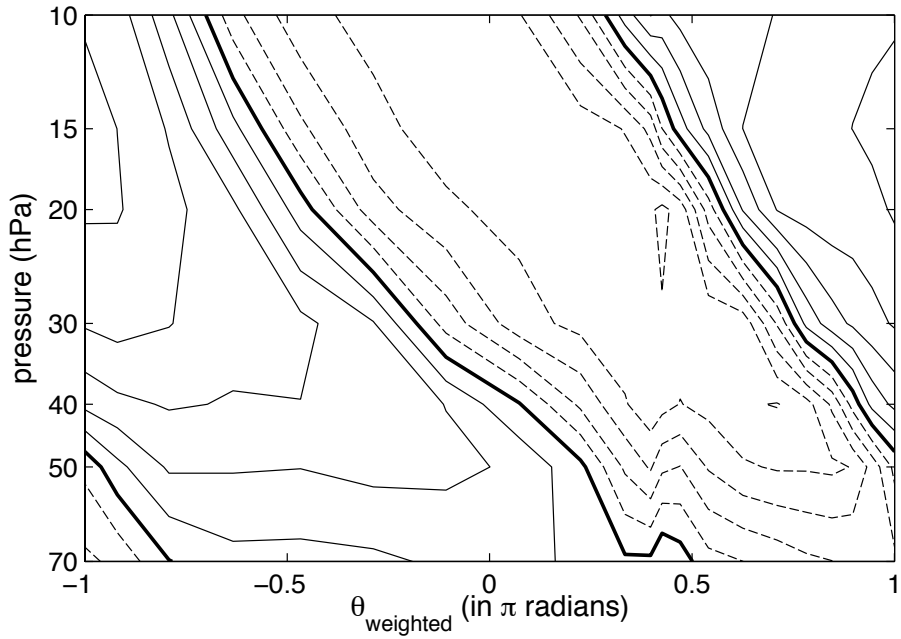

Figure : Contour plot of the NLPCA(cir) mode 1 zonal wind anomalies as a function of pressure and phase θ_{weighted}, where θ_{weighted} is θ

weighted by the histogram distribution of θ (see Hamilton and Hsieh, 2002). Thus θ_{weighted} is more representative of actual time during a cycle than θ. Contour interval is 5 ms^{-1}, with westerly winds indicated by solid lines, easterlies by dashed lines, and zero contours by thick lines. As the easterly wind anomaly descends with time (i.e. as phase increases), wavy behaviour is seen in the 40, 50 and 70 hPa levels at θ_{weighted} around 0.4-0.5.

Nonlinear singular spectrum analysis (NLSSA) has also been developed based on the NLPCA(cir) model (Hsieh and Wu, 2002) (Book, Sect.10.6).

Also nonlinear canonical correlation analysis (NLCCA) by NN (Cannon and Hsieh, 2008) (Book, Chap.11)

Functions in Matlab:

NLPCA and NLCCA codes at:

`www.ocgy.ubc.ca/projects/clim.pred/download.html`

References:

Cannon, A. J. and Hsieh, W. W. (2008). Robust nonlinear canonical correlation analysis: application to seasonal climate forecasting. *Nonlinear Processes in Geophysics*, 12:221–232.

Hamilton, K. and Hsieh, W. W. (2002). Representation of the QBO in the tropical stratospheric wind by nonlinear principal component analysis. *Journal of Geophysical Research*, 107(D15). 4232, DOI: 10.1029/2001JD001250.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84:502–516.

Hsieh, W. W. (2001). Nonlinear principal component analysis by neural networks. *Tellus*, 53A:599–615.

Hsieh, W. W. (2007). Nonlinear principal component analysis of noisy data. *Neural Networks*, 20:434–443.

Hsieh, W. W. and Wu, A. (2002). Nonlinear multichannel singular spectrum analysis of the tropical Pacific climate variability using a neural network approach. *Journal of Geophysical Research*, 107(C7). DOI: 10.1029/2001JC000957.

Kirby, M. J. and Miranda, R. (1996). Circular nodes in neural networks. *Neural Computation*, 8:390–402.

Kohonen, T. (1982). Self-organzing formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37:233–243.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326.

Schölkopf, B., Smola, A., and Muller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.