Nonlinear principal component analysis of noisy data

William W. Hsieh
Dept. of Earth and Ocean Sciences, University of British Columbia
Vancouver, BC V6T 1Z4, Canada

Abstract

With very noisy data, having plentiful samples eliminates overfitting in nonlinear regression, but not in nonlinear principal component analysis (NLPCA). To overcome this problem in NLPCA, a new information criterion (IC) is proposed for selecting the best model among multiple models with different complexity and regularization (i.e. weight penalty). This IC gauges the inconsistency I between the nonlinear principal components (u and $\tilde{u}$) for every data point $\mathbf{x}$ and its nearest neighbour $\tilde{\mathbf{x}}$, with $I = 1 - \text{correlation}(u, \tilde{u})$, where I tends to increase with overfitted solutions. Tests were performed using autoassociative neural networks for NLPCA on synthetic and real climate data (tropical Pacific sea surface temperatures and equatorial stratospheric winds), with the IC performing well in model selection and in deciding between an open curve or a closed curve solution.

Keywords: nonlinear principal component analysis, information criterion, model selection, autoassociative neural network, regularization, El Niño, ENSO, quasi-biennial oscillation

Accepted for publication by **Neural Networks**, 2007.

1 Introduction

In principal component analysis (PCA), a given dataset is approximated by a straight line, which minimizes the mean square error (MSE) — pictorially, in a scatterplot of the data, the straight line found by PCA passes through the 'middle' of the dataset. In nonlinear PCA (NLPCA), the straight line in PCA is replaced by a curve. NLPCA can be performed by a variety of methods, e.g. the autoassociative neural network (NN) model (Kramer, 1991; Hsieh, 2004), and the kernel PCA model (Schölkopf et al. 1998). Alternative approaches include the principal curves/surfaces method (Hastie and Stuetzle, 1989; Hastie et al. 2001), and the self-organizing map technique (Kohonen, 1982) which can be regarded as a discrete version of NLPCA or principal curves/surfaces (Cherkassy and Mulier, 1998).

When using nonlinear machine learning methods, the presence of noise in the data can lead to overfitting (i.e. fitting to the noise). However, in the limit of infinite samples, overfitting is not a problem when performing nonlinear regression on noisy data, since it can be shown that the output of a flexible enough nonlinear regression model approximates the conditional mean of the target data (Bishop, 1995, Sec. 6.1.3). While overfitting can also occur in NLPCA (Hsieh, 2001, Christiansen, 2005), the situation is actually far worse than in nonlinear regression, because even in the limit of infinite samples, overfitting is a problem when applying NLPCA to noisy data. As illustrated in Fig. 1, overfitting in NLPCA can arise from the geometry of the data distribution, instead of from the relative scarsity of samples. Here for a Gaussian-distributed data cloud, a nonlinear model with enough flexibility will find the zigzag solution of Fig. 1b as having a smaller MSE than the linear solution in Fig. 1a. Since the distance between the point A and a, its projection on the NLPCA curve, is smaller in Fig. 1b than the corresponding distance in Fig. 1a, it is easy to see that the more zigzags there are in the curve, the smaller is the MSE. However, the two neighbouring points A and B, on opposite sides of an "ambiguity" line (Hastie and Stuetzle, 1989; Malthouse, 1998), are projected far apart on the NLPCA curve in Fig. 1b. Thus simply searching for the solution which gives the smallest MSE is not a sufficient criterion for NLPCA to find the best solution in a highly noisy dataset.

Regularization (e.g. the addition of weight penalty or decay terms in the objective functions in NN models) has been commonly used to control overfitting by limiting model complexity (i.e. the effective number of model parameters) via the size of the weight penalty parameter(s) (Bishop, 1995). A larger weight penalty parameter P tends to give less nonlinear solutions than a smaller P. Typically, to find the appropriate P in nonlinear regression and classification, a number of models are trained with different P values. The models' MSE are validated on independent data not used before in the model training stage, and the model with the lowest validated MSE is selected as the best. Alternatively, Bayesian methods have been developed to automatically estimate the size of the weight penalty parameter in nonlinear regression and classification problems (MacKay, 1992; Foresee and Hagan, 1997).

As Kramer's (1991) NLPCA model tends to extract zigzag solutions, Hsieh (2001) added weight penalty to the NLPCA model, which brought the overfitting under control. Unfortunately, there was no simple way to objectively estimate the appropriate P value needed to avoid overfitting (and underfitting), because with NLPCA, if the overfitting arises from the data geometry (as in Fig. 1b) and not from the relative scarsity of samples, using independent data to validate the MSE from the various models is not a viable method for selecting the appropriate P. Instead, we propose a new "inconsistency" index for detecting the projection of neighbouring points to distant parts of the NLPCA curve, and incorporate the inconsistency index into an information criterion for selecting the best model from a number of models ran with different P values and different model architecture.

In Sec. 2, the use of autoassociative NN models for NLPCA is outlined. In Sec. 3, the inconsistency index and the information criterion are presented, and tested on synthetic data and tropical Pacific sea surface temperature (SST) data. In Sec. 4, the information criterion is used to select either a closed curve or an open curve as the best fit to a dataset, with applications to the equatorial stratospheric wind data. Limitations of the NLPCA method are discussed in Sec. 5. More details of the autoassociative NN models used are given in the Appendix.

2 Autoassociative NN model for NLPCA

To perform NLPCA, the NN model (Fig. 2a) is a standard feed-forward (multi-layer perceptron) NN with 3 'hidden' layers of variables or 'neurons' sandwiched between the input layer $\mathbf{x}$ on the left and the output layer $\mathbf{x}'$ on the right, where the middle hidden layer has only a single "bottleneck" neuron u. As an autoassociative model, the MSE between the output $\mathbf{x}'$ and the input $\mathbf{x}$ is minimized, and data compression is achieved by the bottleneck, yielding the nonlinear principal component (NLPC) u (see the Appendix for details). Model complexity can be increased by increasing m, the number of hidden neurons in layer 1 and in layer 3 of the NN (Fig. 2a).

Using the Bayesian NN code trainbr.m (Foreseee and Hagan, 1997) in the MATLAB Neural Network Toolbox to perform NLPCA failed to prevent the finding of zigzag solutions in Gaussian data clouds, hence a different strategy is needed to choose the weight penalty parameter.

While the NLPCA is capable of finding a continuous open curve solution, there are many phenomena involving waves or quasi-periodic fluctuations, which call for a continuous closed curve solution. Kirby and Miranda (1996) introduced an NLPCA with a circular node at the network bottleneck [henceforth referred to as the NLPCA(cir)], so that NLPCA(cir) is capable of approximating the data by a closed continuous curve. Fig. 2b shows the NLPCA(cir) network, which is identical to the NLPCA of Fig. 2a except at the bottleneck, where there are now two neurons p and q constrained to lie on a unit circle in the p-q plane, so there is effectively only one free angular variable θ, the NLPC (see Appendix). This network has also been used to perform nonlinear singular spectrum analysis (Hsieh and Wu, 2002).

Although NLPCA(cir) is designed for extracting closed curve solutions, it is also capable of extracting an open curve solution. The reason is that if the input data mapped onto the p-q plane covers only a segment of the unit circle instead of the whole circle, then the inverse mapping from the p-q space to the output space will yield a solution resembling an open curve. Hence, NLPCA(cir) may extract either a closed curve or an open curve approximation to a dataset. The new information criterion will be used in Sec. 4 to choose between open and closed curves.

3 Information criterion for model selection

To introduce an inconsistency index for detecting the projection of neighbouring points to distant parts of the NLPCA curve, we first find for each data point $\mathbf{x}$ its nearest neighbour $\tilde{\mathbf{x}}$. The NLPC for $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are u and $\tilde{u}$, respectively. With $C(u, \tilde{u})$ denoting the (Pearson) correlation between all the pairs $(u, \tilde{u})$, the inconsistency index I is defined by

$$I = 1 - C(u, \tilde{u})\,. \tag{1}$$

If for some nearest neighbour pairs, u and $\tilde{u}$ are assigned very different values, $C(u, \tilde{u})$ would have a lower value, leading to a larger I, indicating greater inconsistency in the NLPC mapping. With u and $\tilde{u}$ standardized to having zero mean and unit standard deviation, (1) is equivalent to

$$I = \frac{1}{2} \langle (u - \tilde{u})^2 \rangle \,, \tag{2}$$

where $\langle \cdots \rangle$ denotes averaging over all samples.

In statistics, various criteria, often in the context of linear models, have been developed to select the right amount of model complexity so neither overfitting nor underfitting occurs. These criteria are often called "information criteria" (IC), e.g. the Akaike (1974) IC, the Bayesian IC (Schwarz, 1978), etc. An IC is typically of the form

$$\text{IC} = \text{MSE} + \text{complexity term}, \tag{3}$$

where MSE is evaluated over the training data and the complexity term is larger when a model has more free parameters. The IC is evaluated over a number of models with different free parameters, and the model with the minimum IC is selected as the best. As the presence of the complexity term in the IC penalizes models which use excessive number of free parameters to attain low MSE, choosing the model with the minimum IC would rule out complex models with overfitted solutions.

Due to the presence of multiple minima in the objective function, we randomly divide the data into a training data set and a validation set (containing 85% and 15% of the original data, respectively, in the following examples), and for every given value of P and m, we train the model a number of times from random initial weights, and discard model runs where the MSE evaluated over the validation data is larger than the MSE over the training data. To choose among the model runs which have passed the validation test, a new holistic IC to deal with the type of overfitting arising from the broad data geometry (Fig. 1b) is introduced as

$$\begin{aligned} H &= \text{MSE} + \text{inconsistency term} && (4) \\ &= \text{MSE} - C(u, \tilde{u}) \times \text{MSE} \;=\; \text{MSE} \times I \,, && (5) \end{aligned}$$

where MSE and C are evaluated over all (training and validation) data, inconsistency is penalized, and the model run with the smallest H value is selected as the best. Note that as the inconsistency term only prevents overfitting arising from the broad data geometry, validation data were still needed to prevent "local" overfitting from excessive number of model parameters, since H, unlike (3), does not contain a complexity term.

A test problem was set up as follows: For a random number t uniformly distributed in the interval $(-1, 1)$, the signal $\mathbf{x}^{(s)}$ was generated by using a quadratic relation

$$x_1^{(s)} = t, \qquad x_2^{(s)} = \frac{1}{2} t^2. \tag{6}$$

Isotropic Gaussian noise (with variance being one half the average variance of $x_1^{(s)}$ and $x_2^{(s)}$) was then added to the signal $\mathbf{x}^{(s)}$ to give the noisy data $\mathbf{x}$ with 500 samples. NLPCA was performed on the data using the network in Fig. 2a with $m = 4$ and with the weight penalty parameter P at various values (10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0). For each value of P, the model training was done 30 times starting from random initial weights, and model runs where the MSE evaluated over the validation data was larger than the MSE over the training data were deemed ineligible. In the traditional approach, among the eligible runs over the range of P values, the one with the lowest MSE over all (training and validation) data was selected as the best. Fig. 3a shows this solution

where the zigzag curve retrieved by NLPCA is very different from the theoretical parabolic signal (6), demonstrating the pitfall of selecting the lowest MSE run.

In contrast, in Fig. 3b, among the eligible runs over the range of P values, the one with the lowest information criterion H was selected. This solution, which has a much larger weight penalty ($P = 0.1$) than that in Fig. 3a ($P = 10^{-4}$), shows less wiggly behaviour and better agreement with the theoretical parabolic signal.

Even less wiggly solutions can be obtained by changing the error norm used in the objective function from the mean square error to the mean absolute error (MAE), i.e. replacing $\langle \|\mathbf{x}-\mathbf{x}'\|^2 \rangle$ by $\langle \sum_j |x_j - x'_j| \rangle$ in Eq. (A.2). The MAE norm is known to be robust to outliers in the data (Bishop, 1995, p. 210). Fig. 3c is the solution selected based on minimum H with the MAE norm used. While wiggles are eliminated, the solution underestimates the curvature in the parabolic signal. The rest of this paper uses the MSE norm.

The H IC approach was also tested on a real climate dataset, the tropical Pacific SST, where the interannual variability is dominated by the El Niño-Southern Oscillation (ENSO) phenomenon. The monthly SST data on a 2°×2° grid for the period 1948-2005 came from the Extended Reconstructed Sea Surface Temperatures (ERSST version 2) dataset (Smith and Reynolds, 2004) (downloadable from `ftp.ncdc.noaa.gov/pub/data/ersst-v2`). The SST anomalies were obtained by subtracting the climatological seasonal cycle. PCA was performed on the SST anomalies in the tropical Pacific domain of 124°E-70°E, 20°S-20°N. The 7 leading principal components (PC) containing 86.5% of the variance were retained, and served as the inputs for the NLPCA model. Fig. 4 shows the three leading PC time series. In the first PC, the peaks and troughs correspond, respectively, to warm (El Niño) and cool (La Niña) episodes, while PC2 tends to peak during both warm and cool episodes. The SST anomaly spatial patterns from the three leading empirical orthogonal functions (EOF) (i.e. loadings) are shown in Figs. 5 a, b and c. The contribution to the SST anomaly field from a given PCA mode at a particular time is the PC at that time multiplied by the corresponding EOF spatial pattern.

NLPCA was performed over a range of m and P values ($m = 2, \ldots, 6$, and P =10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0). For each combination of m and P, 100 runs starting from random inital weights were made. Among all the runs made over the whole range of m and P values, the one with the lowest H was selected as the best (with $m = 5$, $P = 0$). The NLPCA solution is a curve in the 7-dimensional PC space. Fig. 6 displays this solution in the PC1-PC2-PC3 space, while Fig. 5 d and e show the SST anomaly field when the NLPC u is at its maximum and minimum values, respectively. The u time series has a correlation of 0.983 with PC1, hence strong El Niño episodes occur when u and PC1 are near their maximum, and strong La Niña episodes, when near their minimum. Comparing Fig. 5a and Fig. 5d, we note that during strong El Niño, the NLPCA mode 1 has warmer SST anomalies near the eastern boundary than the PCA mode 1. The strong La Niña pattern in NLPCA mode 1 (Fig. 5e) has the strongest cool anomalies in the central equatorial Pacific, far from the eastern boundary, in contrast to the strong El Niño situation (Fig. 5d), where the strongest warm anomalies are near the eastern boundary. These patterns are somewhat similar to those found in Monahan (2001) and Hsieh (2001).

We next compare the best solutions found for different hidden neuron number m. Since the overall best solution based on minimum H was for $m = 5$, we also showed the best solution found for $m = 2$ and $m = 6$ in the PC1-PC2 plane (Fig. 7). The (normalized) MSE for the 3 solutions in Fig. 7 are 0.898 ($m = 2$), 0.857 ($m = 5$) and 0.826 ($m = 6$), where for easy comparison with the linear mode, the values for the NLPCA solution have been divided by that from the PCA mode 1. For the (normalized) inconsistency index I, the values are 0.896 ($m = 2$), 0.879 ($m = 5$) and 0.946 ($m = 6$), while for the (normalized) H IC, the corresponding values are 0.804, 0.753 and 0.782 respectively. Hence the $m = 6$ solution has a lower MSE than the $m = 5$ solution, but the

increased inconsistency from its wiggly curve (Fig. 7c) led to a larger I and a larger H. Compared to the $m = 2$ solution, the $m = 5$ solution has both lower MSE and lower I.

4 Closed curve solutions

In many oscillatory systems, there are two extreme states, e.g. winter and summer. Depending on the variable, the intermediate states, e.g. spring and autumn, may or may not be similar. For instance, in temperature, spring and autumn are similar, but in the colour of forests, the two seasons are very distinct. Thus in the former case, an open curve is most suited to describe the main oscillation in the data, while in the latter, a closed curve is most suited. The NLPCA(cir) model (Fig. 2b) is capable of representing either a closed or an open curve. It is shown below that the IC H not only alleviates overfitting in open curve solution, but also chooses between open and closed curve solutions. The inconsistency index and the IC are now obtained from

$$I = 1 - \frac{1}{2}\left[C(p, \tilde{p}) + C(q, \tilde{q})\right], \quad \text{and} \quad H = \text{MSE} \times I\,, \tag{7}$$

where p and q are from the bottleneck (Fig. 2b), and $\tilde{p}$ and $\tilde{q}$ are the corresponding nearest neighbour values.

For a test problem, we chose a Gaussian data cloud (with 500 samples) in 2-dimensional space, where the standard deviation along the x_1 axis was double that along the x_2 axis. The data set was analyzed by the NLPCA(cir) model with $m = 2, \ldots, 5$ and $P = 10$, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0. From all the runs, the solution selected based on the minimum MSE has $m = 5$ (and $P = 10^{-5}$) (Fig. 8a), while that selected based on minimum H has $m = 3$ (and $P = 10^{-5}$) (Fig. 8b). The minimum MSE solution has (normalized) MSE = 0.370, $I = 9.50$ and $H = 3.52$, whereas the minimum H solution has the corresponding values of 0.994, 0.839 and 0.833, respectively. Thus the IC correctly selected a nonlinear solution (Fig. 8b) which is similar to the linear solution. It also rejected the closed curve solution of Fig. 8a, in favour of the open curve solution of Fig. 8b, despite its much larger MSE.

For a second test problem, we choose an oval signal imbedded in low and high Gaussian noise in 2-dimensional space. Again runs with $m = 2, \ldots, 8$ and P ranging from 10 to 0 were performed, with the H IC selected solutions shown in Fig. 9. Thus in contrast to the Gaussian example (Fig. 8b), the IC selected a closed curve solution even under heavy noise (Fig. 9b).

For an application on real data, we turn to 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, with a mean of about 28 months. Average zonal (i.e. westerly) winds (Naujokat, 1986) 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 (Hamilton and Hsieh, 2002). Since the data were not very noisy (Fig. 10), a rather complex model was tried here, with m ranging from 5 to 9, and $P = 10^{-1}$, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0. The smallest H occurred when $m = 8$ and $P = 10^{-5}$, with the closed curve solution shown in Fig. 10. Thus in this example, by choosing a rather large m and a small P, the H IC justified having considerable model complexity, including the wiggly behaviour seen in the 70 hPa wind (Fig. 10c). The wiggly behaviour can be understood by viewing the phase-pressure contour plot of the zonal wind anomalies (Fig. 11): 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. This example

demonstrates the benefit of having an IC to objectively decide on how smooth or wiggly the fitted curve should be.

5 Summary and Discussion

With noisy data, not having plentiful samples could cause a flexible nonlinear model to overfit. In the limit of infinite samples, overfitting cannot occur in nonlinear regression, but can still occur in NLPCA due to the geometric shape of the data distribution. A new inconsistency index I for detecting the projection of neighbouring points to distant parts of the NLPCA curve has been introduced, and incorporated into a holistic IC H to choose the model with the appropriate weight penalty parameter and the appropriate number of hidden neurons. Tests with synthetic data and real climate data indicated that this IC is effective in model selection, and in deciding between open curve and closed curve solutions. The MATLAB code is downloadable from `http://www.ocgy.ubc.ca/projects/clim.pred/download.html`. Although the NLPCA was performed with autoassociative NN models, this IC for NLPCA model selection should work with other methods for nonlinear PCA.

While this H IC worked well in climate datasets where there is one dominant signal (e.g. ENSO in the tropical Pacific SST; QBO in the stratospheric wind), it remains inadequate for dealing with datsets which contain two or more distinct signals of roughly comparable strength – e.g. in the extratropical N. Hemisphere climate, where there has been considerable controversy on the use of NLPCA (Christiansen, 2005; Monahan and Fyfe, 2007; Christiansen, 2007), there are two signals of comparable magnitude, the Arctic Oscillation and the Pacific-North American teleconnection. The reason is that if there are two comparable signals, the total signal forms a 2-D surface whereas our NLPCA or NLPCA(cir) model will be trying to fit a 1-D curve to this surface, resulting in a hybrid mode with attributes from both signals. While it is possible to have two neurons in the bottleneck layer in the NLPCA network (Fig. 2a), so that a 2-D solution is extracted, there is no simple way to separate the two signals. Clearly more research is needed in developing NLPCA and IC for such complicated noisy datasets.

Acknowledgement: I would like to thank Aiming Wu for his help on the use of the GrADS plotting package, Alex Cannon for useful discussion on robust error norms, and Kevin Hamilton for sending the stratospheric zonal wind data provided by Barbara Naujokat of the Free University of Berlin. This work was supported by Discovery and Strategic grants from the Natural Sciences and Engineering Research Council of Canada, and a project grant from the Canadian Foundation for Climate and Atmospheric Sciences.

Appendix

With the input variables forming the 0th layer of the network in Fig. 2a, a neuron $v_j^{(i)}$ at the ith layer ($i = 1, 2, 3, 4$) receives its value from the neurons $\mathbf{v}^{(i-1)}$ in the preceding layer, i.e.

$$v_j^{(i)} = f^{(i)}(\mathbf{w}_j^{(i)} \cdot \mathbf{v}^{(i-1)} + b_j^{(i)}) \,, \tag{A.1}$$

where $\mathbf{w}_j^{(i)}$ is a vector of weight parameters and $b_j^{(i)}$ a bias or offset parameter, and the transfer or activation functions $f^{(1)}$ and $f^{(3)}$ are the hyperbolic tangent functions, while $f^{(2)}$ and $f^{(4)}$ are simply the identity functions. Effectively, a nonlinear function $u = F(\mathbf{x})$ maps from the higher dimension input space to the lower dimension bottleneck space, followed by an inverse transform $\mathbf{x}' = \mathbf{G}(u)$ mapping from the bottleneck space back to the original space, as represented by the outputs. To make the outputs as close to the inputs as possible, the objective function J, basically the MSE, is minimized. More precisely, Hsieh (2004) used

$$J = \langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle + \langle u \rangle^2 + (\langle u^2 \rangle - 1)^2 + P \sum_j \|\mathbf{w}_j^{(1)}\|^2 \,, \tag{A.2}$$

where on the right hand side, the first term is the MSE (with $\langle \cdots \rangle$ denoting a sample or time mean), the second and third terms are for restraining u towards $\langle u \rangle = 0$ and $\langle u^2 \rangle = 1$, and the final term is a weight penalty or regularization term, with P the weight penalty parameter. Hsieh (2001) found that penalizing just the first layer of weights is sufficient to limit the nonlinear modelling capability of the model. By minimizing J, the values of the weight and bias parameters are solved. In this paper, the nonlinear optimization was carried out by the quasi-Newton algorithm fminu.m in the MATLAB Optimization Toolbox. A number of optimization runs was made with random initial values of the weight and bias parameters, and only runs where the MSE evaluated over the validation data was not larger than the MSE over the training data were deemed eligible, with the best solution selected as the one with the smallest value of H, calculated from (5).

To obtain closed curve solutions, we use NLPCA with a circular bottleneck node (Fig. 2b). At the bottleneck, we first calculate the pre-states p_o and q_o by

$$p_o = \mathbf{w}_1^{(2)} \cdot \mathbf{v}^{(1)} + b_1^{(2)} \,, \quad \text{and} \quad q_o = \mathbf{w}_2^{(2)} \cdot \mathbf{v}^{(1)} + b_2^{(2)} \,, \tag{A.3}$$

then with

$$r = (p_o^2 + q_o^2)^{1/2} \,, \tag{A.4}$$

the circular node is defined by

$$p = p_o / r \,, \quad \text{and} \quad q = q_o / r \,, \tag{A.5}$$

which satisfies the unit circle equation $p^2 + q^2 = 1$. Thus, although there are two variables p and q at the bottleneck, there is only one angular degree of freedom (θ) from the circle constraint. For more details, see the review by Hsieh (2004). The model run having the smallest H, as computed from (7), is selected as the best solution.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control, AC-19*, 716-723.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Pr.

Cherkassky, V., & Mulier, F. (1998). *Learning from Data*. New York: Wiley.

Christiansen, B. (2005). The shortcomings of nonlinear principal component analysis in identifying circulation regimes. *J. Clim., 18*, 4814-4823.

Christiansen, B. (2007). Reply to Monahan and Fyfe's comment on "The shortcomings of nonlinear principal component analysis in identifying circulation regimes". *J. Clim., 20*, 378-379. DOI: 10.1175/JCLI4006.1

Foresee, F. D., & Hagan, M. T. (1997). *Gauss-Newton approximation to Bayesian regularization.* Paper presented at the Proceedings of the 1997 International Joint Conference on Neural Networks.

Hamilton, K., & Hsieh, W. W. (2002). Representation of the QBO in the tropical stratospheric wind by nonlinear principal component analysis. *J. Geophys. Res., 107*(D15), DOI: 10.1029/2001JD001250.

Hastie, T., & Stuetzle, W. (1989). Principal curves. *J. Amer. Stat. Assoc., 84*, 502-516.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

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

Hsieh, W. W. (2004). Nonlinear multivariate and time series analysis by neural network methods. *Rev. Geophys., 42*, RG1003, doi:1010.1029/2002RG000112.

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

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.

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Computation, 4*, 415-447.

Malthouse, E. C. (1998). Limitations of nonlinear PCA as performed with generic neural networks. *IEEE Trans. Neural Networks, 9*, 165-173.

Monahan, A. H. (2001). Nonlinear principal component analysis: Tropical Indo-Pacific sea surface temperature and sea level pressure. *J. Clim., 14*, 219-233.

Monahan, A. H., & Fyfe, J. C. (2007). Comment on "The shortcomings of nonlinear principal component analysis in identifying circulation regimes". *J. Clim., 20*, 375-377, DOI: 10.1175/JCLI4002.1.

Naujokat, B. (1986). An update of the observed quasi-biennial oscillation of the stratospheric winds over the tropics. *J. Atmos. Sci., 43*, 1873-1877.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Smith, T. M., & Reynolds, R. W. (2004). Improved extended reconstruction of SST (1854-1997). *J. Clim., 17*, 2466-2477.

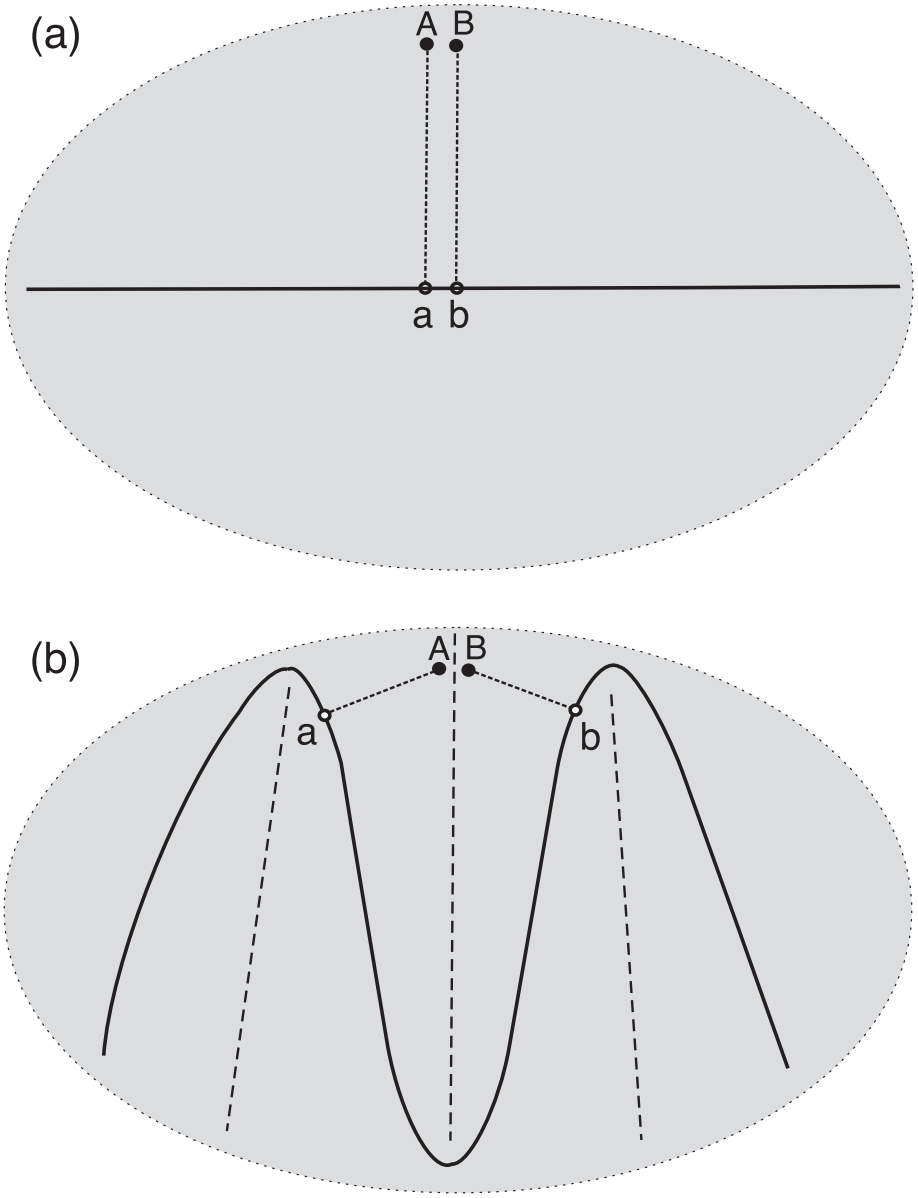

Figure 1: Schematic diagram illustrating how overfitting can occur in NLPCA of noisy data (even in the limit of infinite samples). (a) PCA solution for a Gaussian data cloud (shaded in grey), with two neighbouring points A and B shown projecting to the points a and b on the PCA straight line solution. (b) A zigzag NLPCA solution found by a flexible enough nonlinear model. Dashed lines illustrate "ambiguity" lines where neighbouring points (e.g. A and B) on opposite sides of these lines are projected to a and b, far apart on the NLPCA curve.

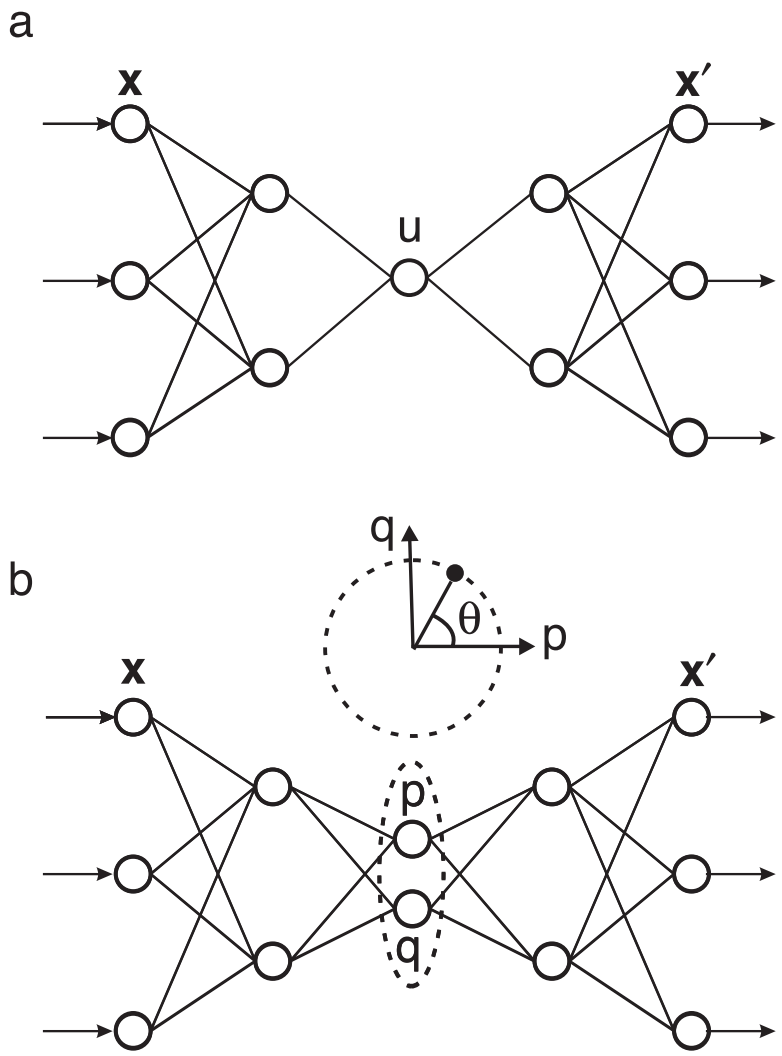

Figure 2: (a) A schematic diagram of the autoassociative feed-forward multi-layer perceptron NN model for performing NLPCA. Between the input layer $\mathbf{x}$ on the left (the 0th layer) and the output layer $\mathbf{x}'$ on the far right (the 4th layer), there are 3 layers of 'hidden' neurons (the 1st, 2nd and 3rd layers). Layer 2 is the 'bottleneck' with a single neuron u giving the nonlinear principal component (NLPC). Layers 1 and 3 each have m hidden neurons. (b) The NN model used for extracting a *closed* curve NLPCA solution. At the bottleneck, there are now two neurons p and q constrained to lie on a unit circle in the p-q plane, giving effectively one free angular variable θ, the NLPC.

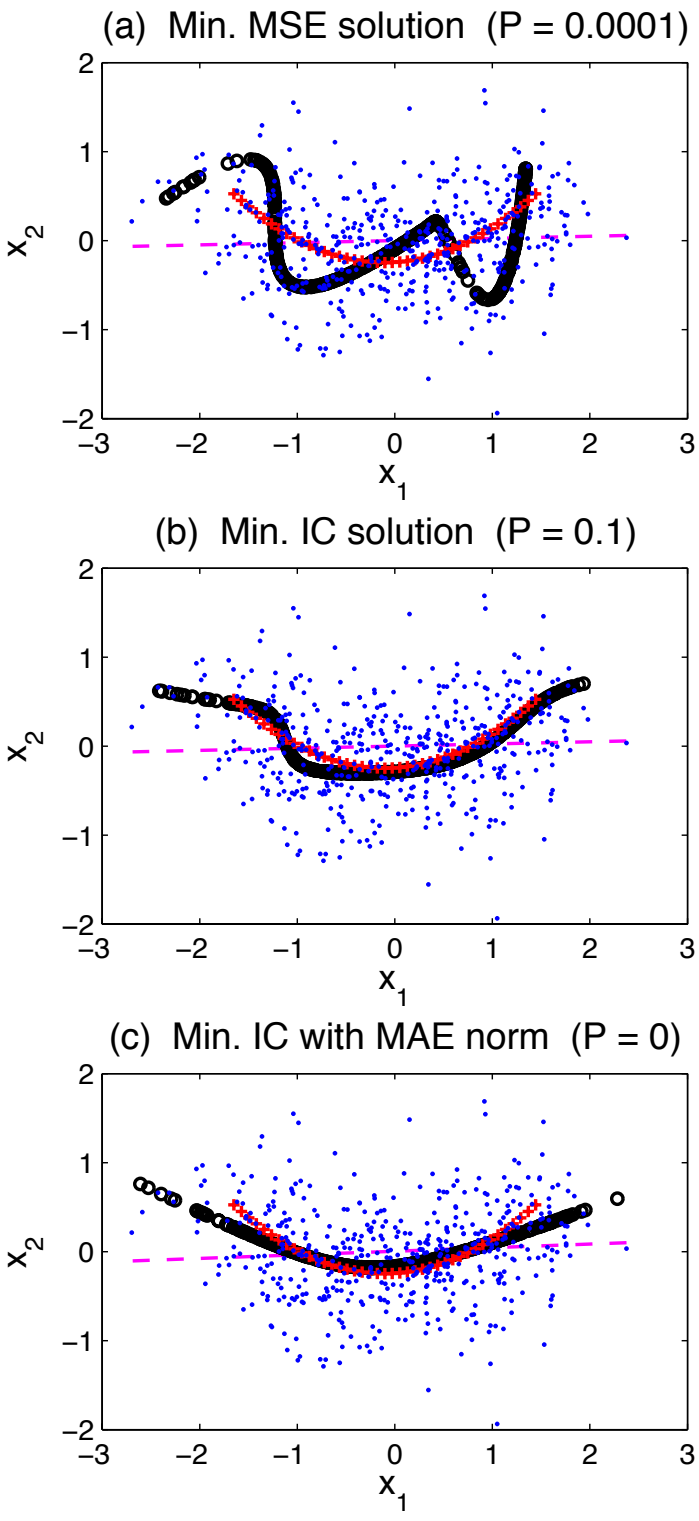

Figure 3: The NLPCA solution (shown as densely overlapping black circles) for the synthetic dataset (dots), with the parabolic signal curve indicated by "+" and the linear PCA solution by the dashed line. The solution was selected from the multiple runs over a range of P values based on (a) minimum MSE, (b) minimum IC H, and (c) minimum IC together with the MAE norm.

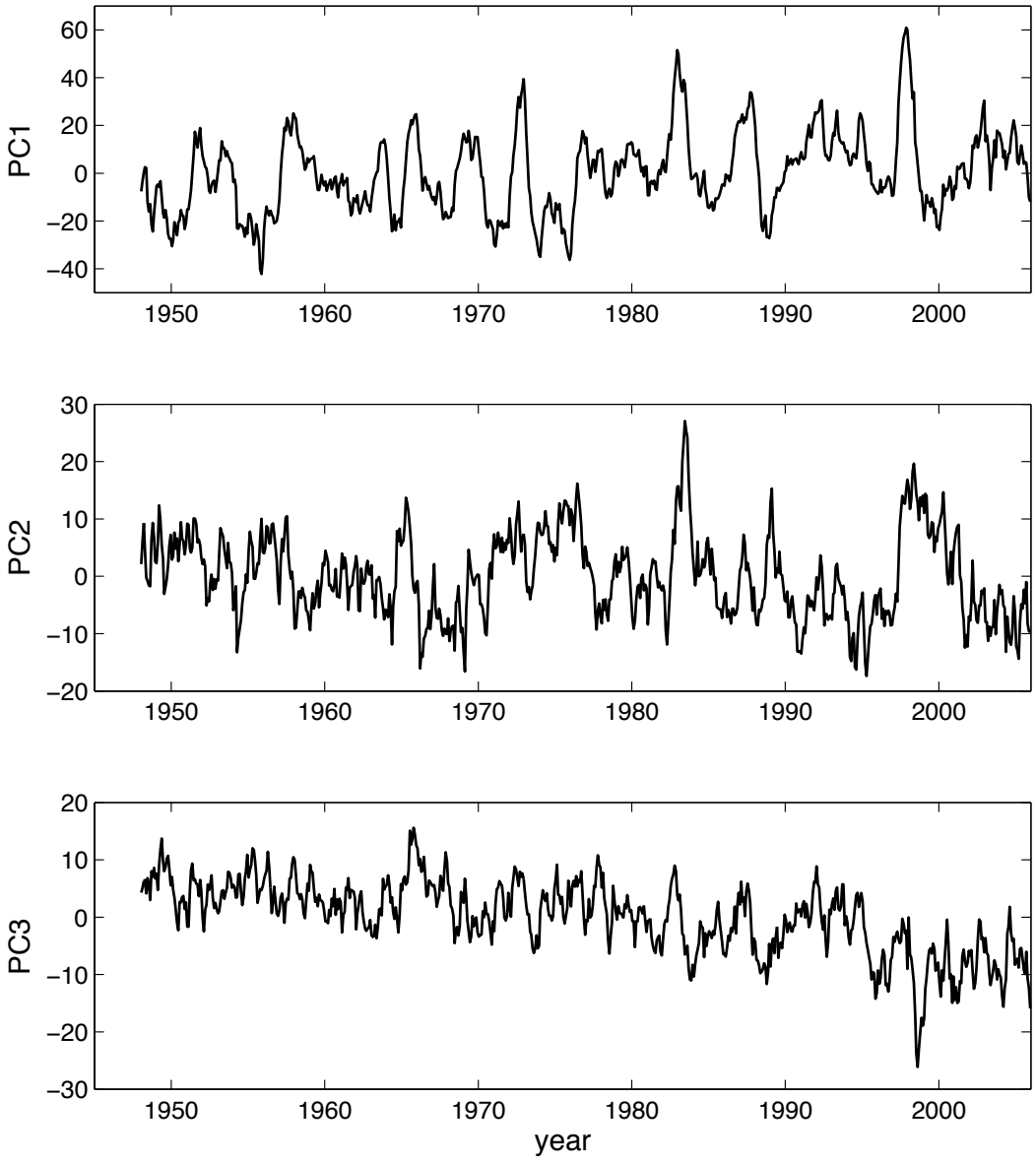

Figure 4: The three leading PC time series (accounting respectively for 56.3%, 10.6% and 7.9% of the variance) of the tropical Pacific SST anomaly data. The tick mark for a year marks the January of that year.

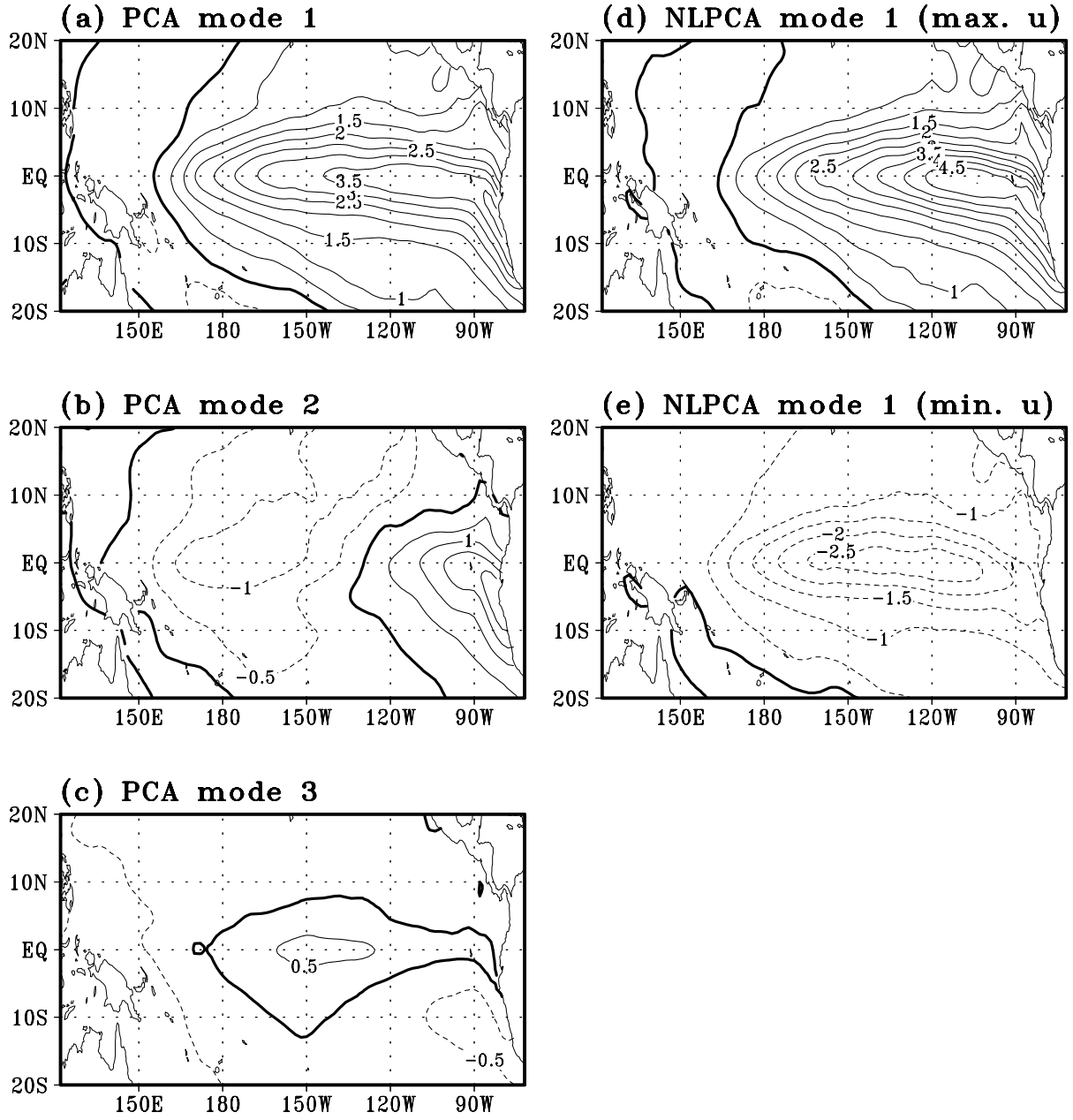

Figure 5: The tropical Pacific SST anomaly field associated with (a) PCA mode 1, (b) PCA mode 2, (c) PCA mode 3, and NLPCA mode 1 when the NLPC u is (d) maximum and (e) minimum. In (a), (b) and (c), the EOF was multiplied by the maximum value of the corresponding PC. Negative contours are dashed, and zero contours thickened.

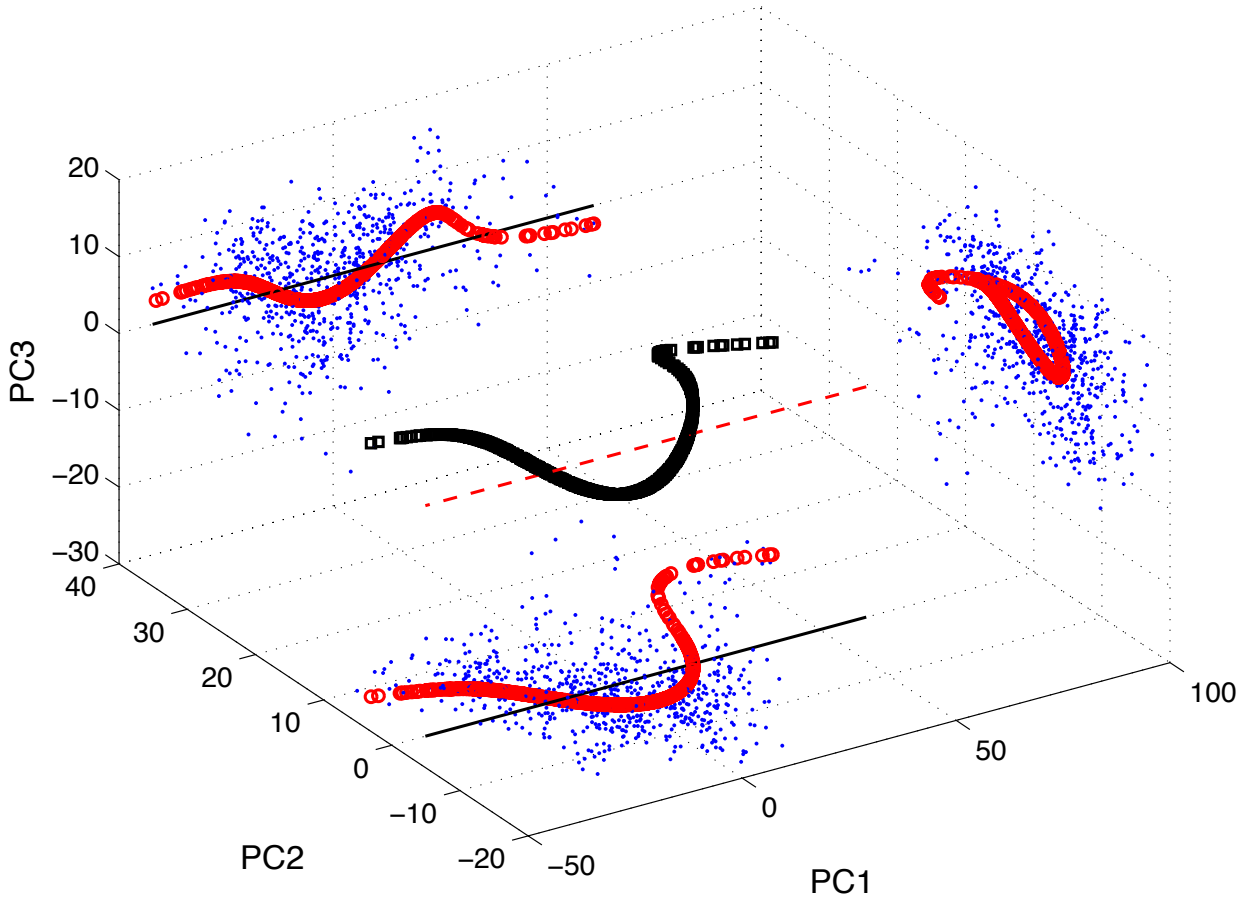

Figure 6: The NLPCA mode 1 of the SST anomaly data plotted as (overlapping) squares in the PC_1-PC_2-PC_3 3-D space. The linear (PCA) mode 1 is also shown as a dashed line. The NLPCA mode and the PCA mode are also projected onto the PC_1-PC_2 plane, the PC_1-PC_3 plane, and the PC_2-PC_3 plane, where the projected NLPCA solution is indicated by (overlapping) circles, and the PCA by thin solid lines, and the projected data points (during 1948-2005) by the scattered dots.

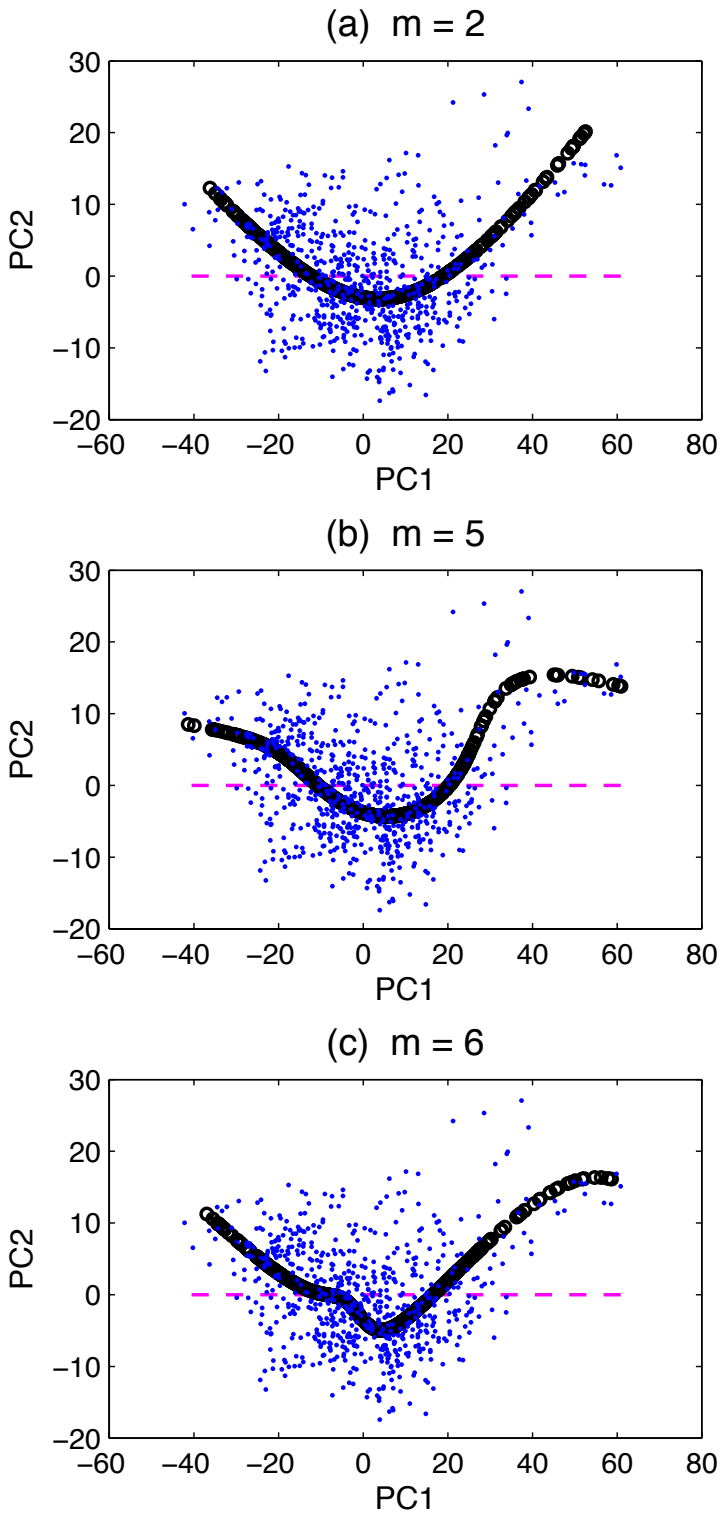

Figure 7: The best NLPCA mode 1 solution (for the SST anomaly data) selected for (a) $m = 2$ (b) $m = 5$ and (c) $m = 6$. The solution is shown only in the PC1-PC2 plane, with the linear PCA mode 1 solution indicated by the dashed line.

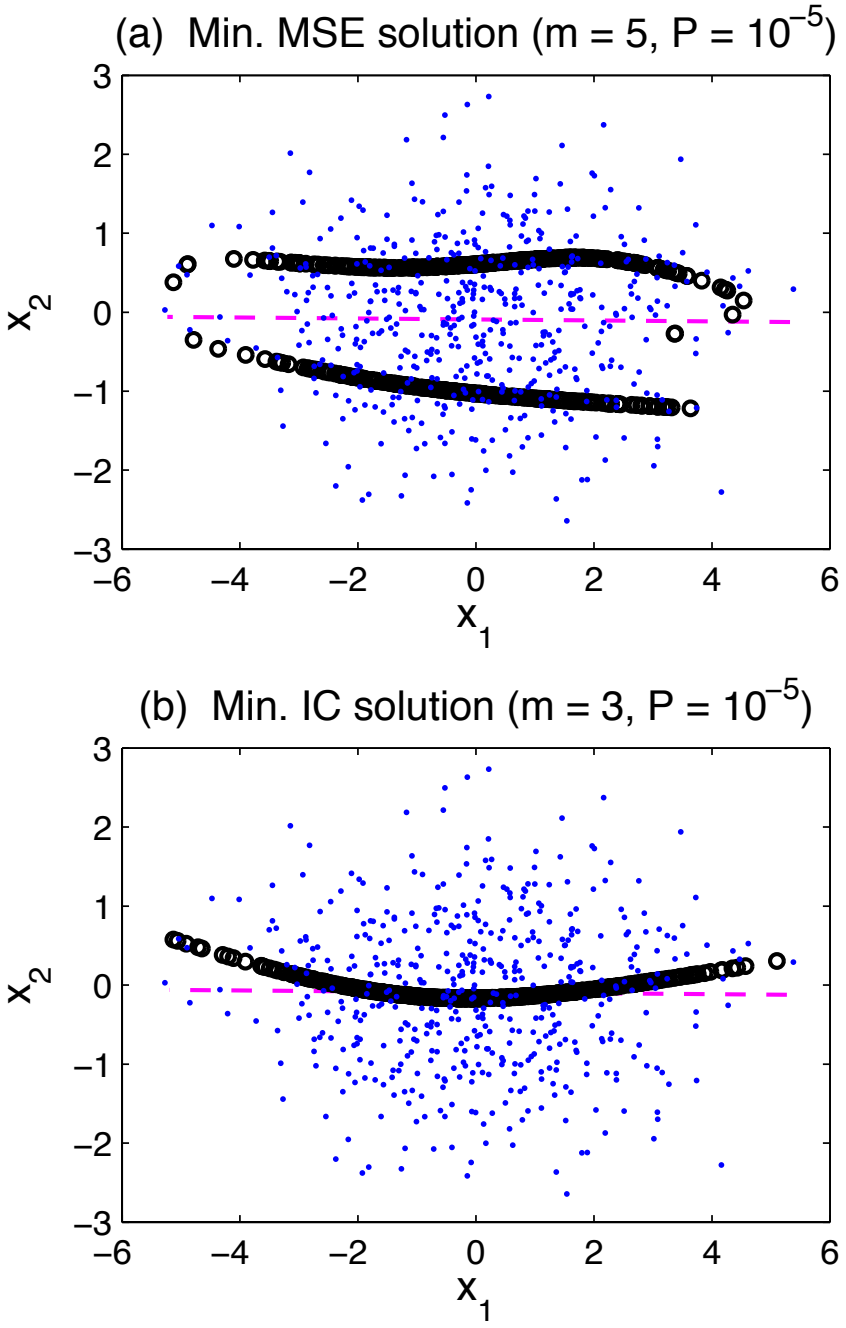

Figure 8: The NLPCA(cir) mode 1 for a Gaussian dataset, with the solution selected based on (a) minimum MSE and (b) minimum IC. The PCA mode 1 solution is shown as a dashed line.

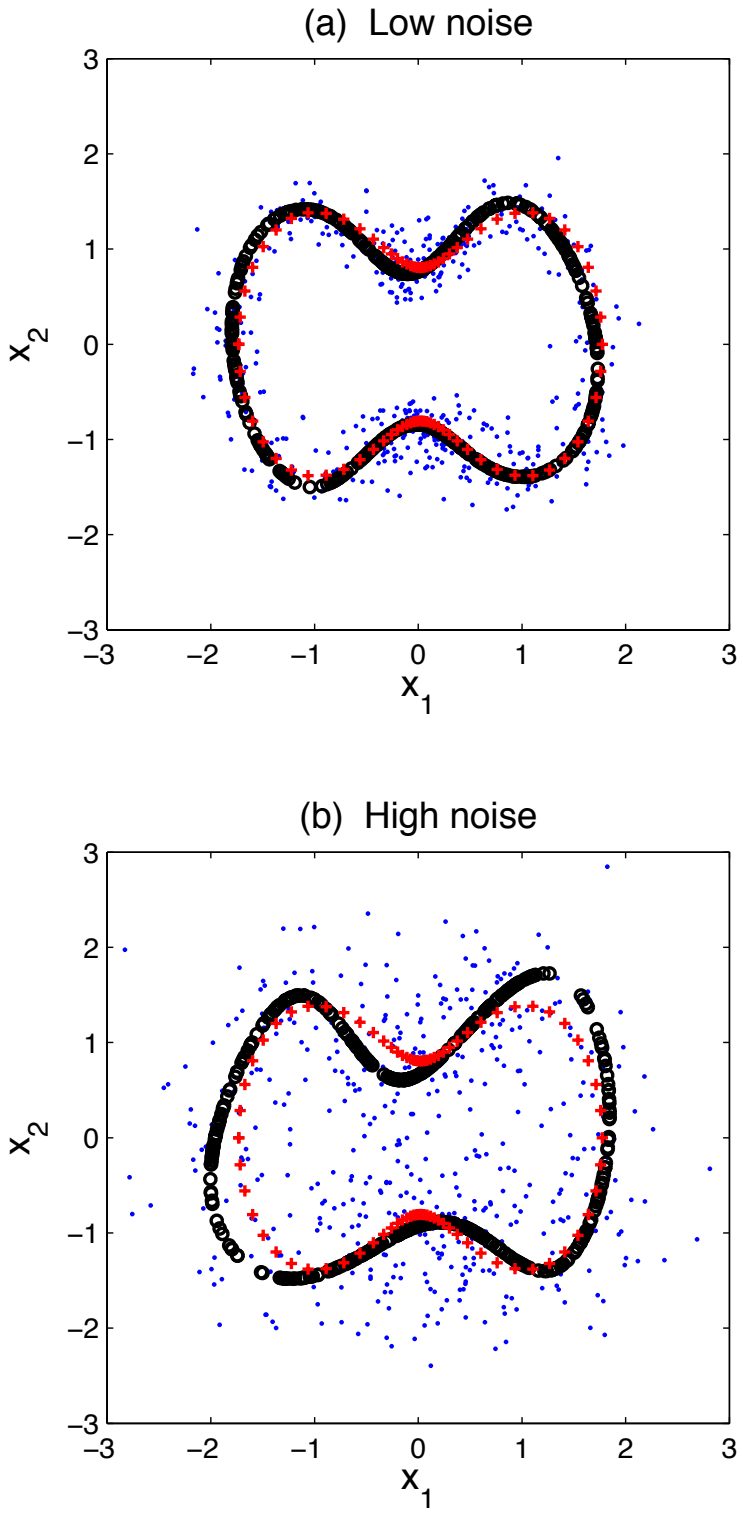

Figure 9: The NLPCA(cir) mode 1 (overlapping black circles) for the dataset containing an oval-shaped signal (indicated by "+") under (a) low noise and (b) high noise. In (a) the circles and "+" signs strongly overlapped.

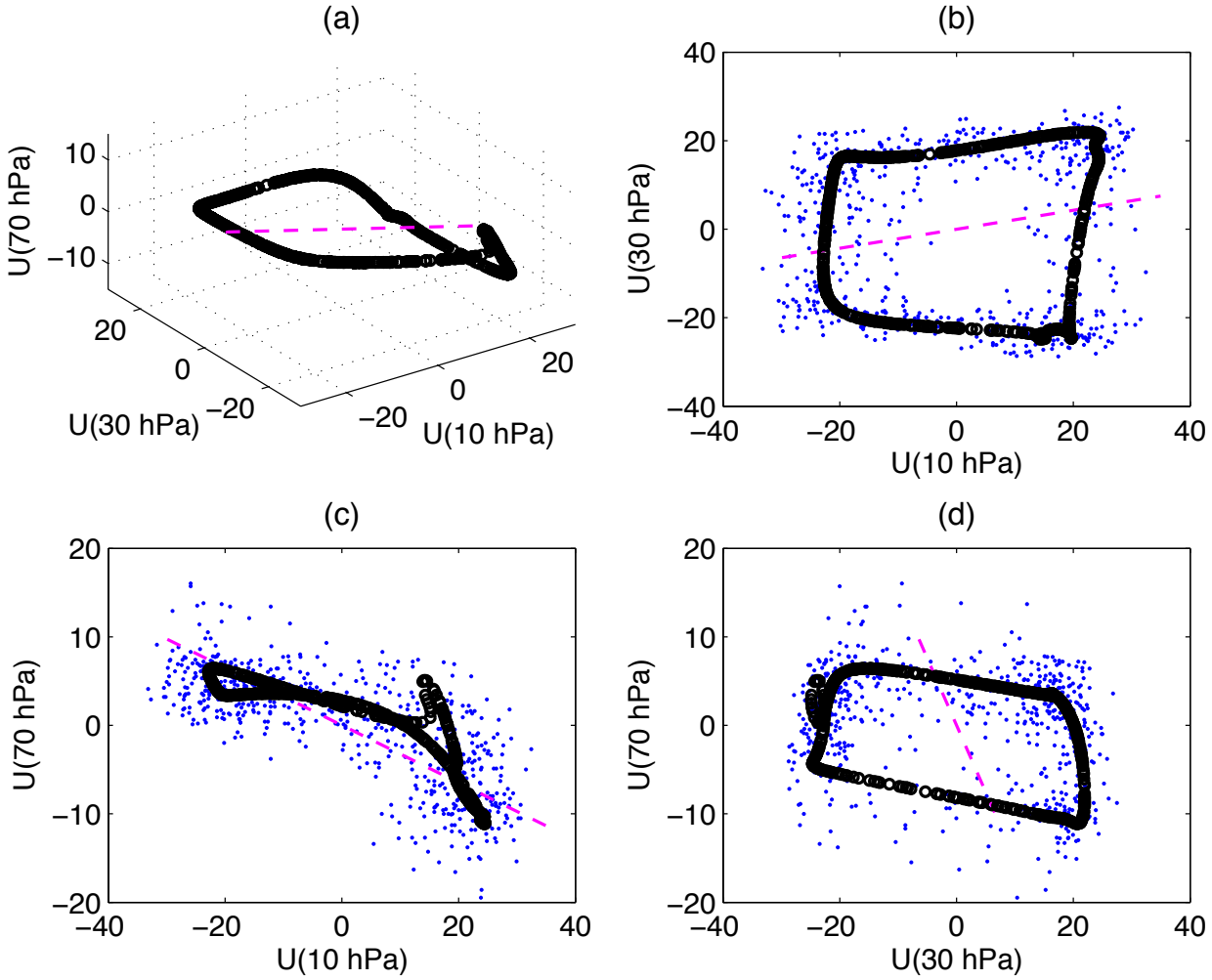

Figure 10: The NLPCA(cir) mode 1 solution for the equatorial stratospheric zonal wind anomalies. For comparison, the PCA mode 1 solution is shown by the dashed line. Only 3 out of 7 dimensions are shown, namely the zonal velocity anomaly U at the top, middle and bottom levels (10, 30 and 70 hPa). Panel (a) gives a 3-D view, while (b)-(d) give 2-D views.

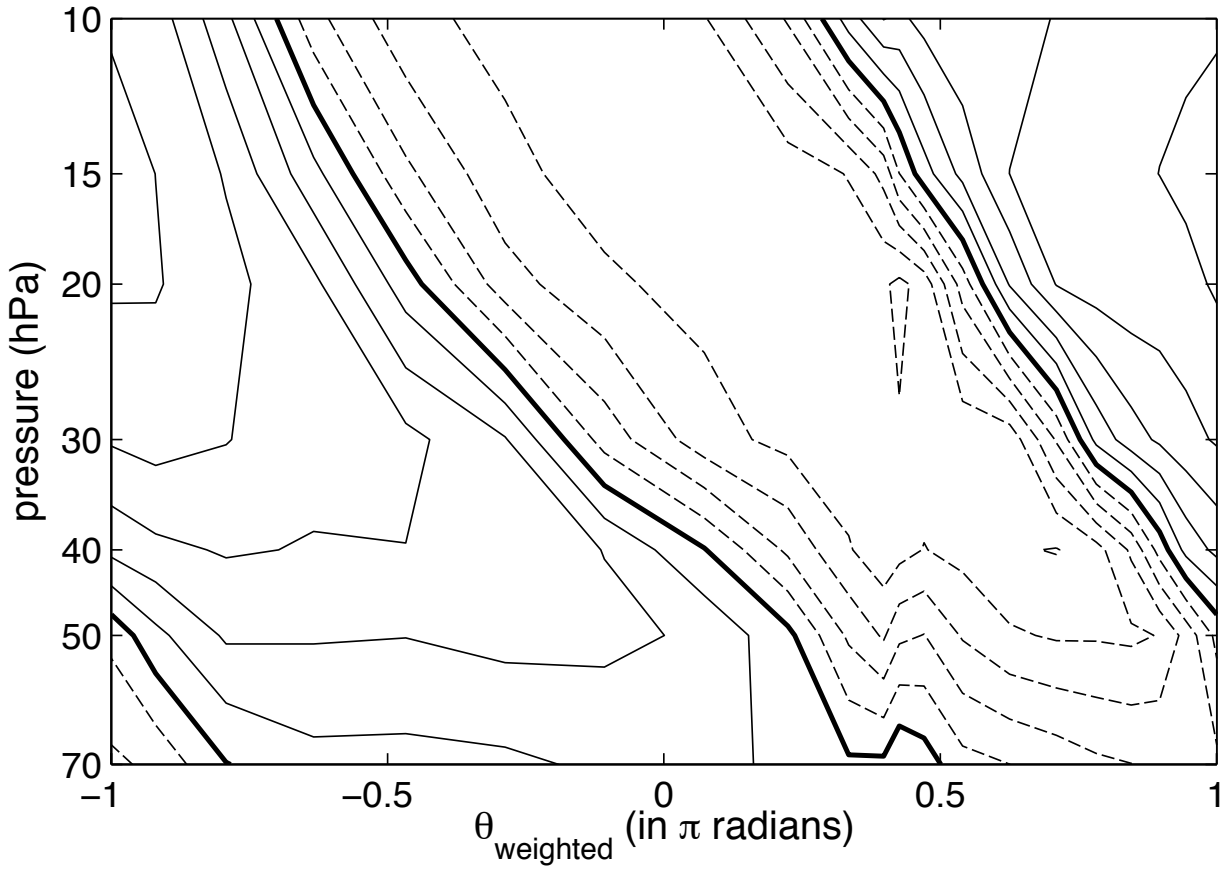

Figure 11: Contour plot of the NLPCA(cir) mode 1 zonal wind anomalies as a function of pressure and phase θ_{weighted}, where θ_{weighted} is θ weighted by the histogram distribution of θ (see Hamilton and Hsieh, 2002). Thus θ_{weighted} is more representative of actual time during a cycle than θ. Contour interval is 5 ms^{-1}, with westerly winds indicated by solid lines, easterlies by dashed lines, and zero contours by thick lines.