Nonlinear principal component analysis of noisy data

William W. Hsieh, *Member, INNS*

***Abstract*— With very noisy data, overfitting is a serious problem in pattern recognition. For nonlinear regression, having plentiful data eliminates overfitting, but for nonlinear principal component analysis (NLPCA), overfitting persists even with plentiful data. Thus simply minimizing mean square error is not a sufficient criterion for NLPCA to find good solutions in noisy data.**

A new index is proposed which measures the disparity between the nonlinear principal components u and $\tilde{u}$ for a data point $\mathbf{x}$ and its nearest neighbour $\tilde{\mathbf{x}}$. This index, $1 - C_S$ (the Spearman rank correlation between u and $\tilde{u}$), tends to increase with overfitted solutions, thereby providing a diagnostic tool to determine how much regularization (i.e. weight penalty) should be used in the objective function of the NLPCA to prevent overfitting. Tests are performed using autoassociative neural networks for NLPCA on synthetic and real climate data.

I. 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 [6, 5], and the kernel PCA model [11].

When using nonlinear machine learning methods, the presence of noise in the data can lead to overfitting (i.e. fitting to the noise). When plentiful data are available (i.e. far more samples than model parameters), overfitting is not a problem when performing nonlinear regression on noisy data. Unfortunately, even with plentiful data, overfitting is a problem when applying NLPCA to noisy data [4, 2]. As illustrated in Figure 1, overfitting in NLPCA can arise from the geometry of the problem, rather than from the scarsity of data. Here for a Gaussian-distributed data cloud, a nonlinear model with enough flexibility will find the zigzag solution of Figure 1b as having a smaller MSE than the linear solution in Figure 1a. Since the distance between the point A and a, its projection on the NLPCA curve, is smaller in Figure 1b than the corresponding distance in Figure 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 [8], are projected far apart on the NLPCA curve in Figure 1b. Thus simply searching for the solution which gives the smallest MSE is not a sufficient criterion for NLPCA to find a satisfactory 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 the effective number of model parameters via the size of the weight penalty parameter(s) [1]. Typically, to find the appropriate weight penalty parameter P, a number of model runs are made with different P values. The models are tested for their MSE on independent data not used in the model building, and the best model is chosen. Alternatively, Bayesian methods have been developed to automatically estimate the size of the weight penalty parameter in nonlinear regression and classification problems [7, 3].

With NLPCA, if the overfitting arise from the data geometry (as in Figure 1b) and not from data scarsity, using independent data to validate the MSE from the various models is not a viable method for choosing the appropriate P. Instead, we propose a new index for detecting the projection of neighbouring points to distant parts of the NLPCA curve, and use the index to choose the appropriate P.

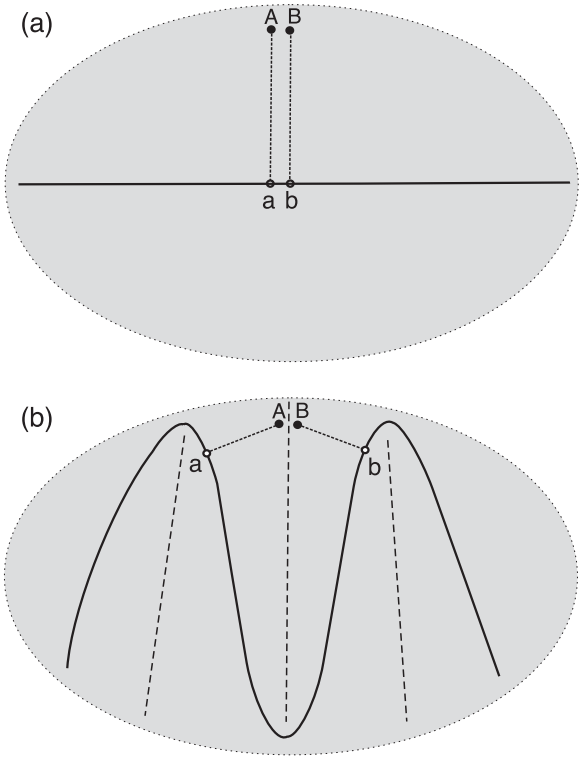

Fig. 1. Schematic diagram illustrating overfitting on noisy data. (a) PCA solution for a Gaussian data cloud, with two neighbouring points A and B shown projecting to the points a and b on the PCA straight line solution. (b) A zigzag NLPCA solution found by a flexible enough nonlinear model. Dashed lines 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.

William W. Hsieh is with the Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (phone: 604-822-2821; fax: 604-822-6088; email: whsieh@eos.ubc.ca).

II. Autoassociative NN model for NLPCA

To perform NLPCA, the NN model (Figure 2) 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 Appendix for details).

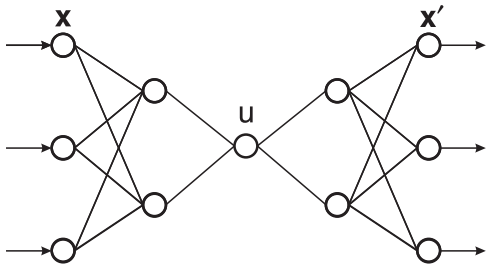

Fig. 2. Schematic diagram of the autoassociativie NN model for performing NLPCA.

Using the Bayesian NN code trainbr.m [3] 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.

III. Results using Synthetic Data

To introduce an 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 u is standardized (i.e. mean removed and divided by the standard deviation). Two indices based on the squared distance between u and $\tilde{u}$ (the standardized NLPC of $\mathbf{x}$ and $\tilde{\mathbf{x}}$, respectively) are introduced:

$$I_1 = \left(\sum |u - \tilde{u}| \right)^2 , \tag{1}$$

$$I_2 = \sum (u - \tilde{u})^2 , \tag{2}$$

where the sum is over all samples, and the distance between u and $\tilde{u}$ is measured by the L_1 norm in (1) and the L_2 norm in (2). When some neighbouring points are projected to distant parts of the NLPCA curve, the difference between u and $\tilde{u}$ becomes large for such pairs, leading to an increase in I_1 and I_2. With C denoting the (Pearson) correlation coefficient and C_S, the Spearman rank correlation coefficient [10], two more indices are introduced:

$$I_P = 1 - C(u, \tilde{u}) , \tag{3}$$

$$I_S = 1 - C_S(u, \tilde{u}) . \tag{4}$$

When u and $\tilde{u}$ are very different for some pairs, both C and C_S would drop, leading to a rise in the indices I_P and I_S.

The proposed strategy is to make a series of model runs with the weight penalty parameter ranging from large to small, then choose the model based on the minimum of one of the I indices, i.e. the model which did the least amount of divergent projection of nearest neighbours is selected as the model with the appropriate P.

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

$$x_1^{(s)} = t, \qquad x_2^{(s)} = \frac{1}{2} t^2. \tag{5}$$

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. Twelve noisy data sets (containing the same signal) were generated. NLPCA was performed on the data with weight penalty parameter P at various values $(10, 1, 0.1, \ldots, 10^{-5}, 0)$. The four indices I_1, I_2, I_P and I_S were computed. Over the twelve noisy data sets, the computed indices showed considerable fluctuation, with I_S showing the least amount. That I_S turned out to be the most reliable is probably because the Spearman rank correlation is a robust statistic.

Figure 3 shows the behaviour of I_S from NLPCA (normalized by I_S computed from the linear PCA method) as P decreased. For large P, the NLPCA solution is relatively close to the linear PCA solution, so the normalized I_S is close to 1. As P decreased, the NLPCA finds a nonlinear solution with lower I_S than the linear method; however, going to even smaller P eventually leads to overfitting, resulting in zigzig solutions and relatively large I_S. Also note the increased scatter among the 12 runs as P becomes small. Thus a viable strategy for choosing the most appropriate P is to proceed from large to small P, locate the first minimum of I_S, and choose the P at this minimum, or more generally choose the smallest P with I_S within $1+\epsilon$ times this minimum I_S value. A heuristic choice of ϵ in our tests is 0.02.

Figure 4 shows (for one of the 12 datasets) the NLPCA solution chosen by the I_S criterion, indicating a successful retrieval of the underlying quadratic signal. Figure 5 displays a zigzag solution for the same dataset as a smaller P is used. This confirms that the I_S criterion is effective in preventing overfitting. Additional tests replacing $x_2^{(s)}$ in (5) by $x_2^{(s)} = t^2$ (a stronger quadratic signal) and by $x_2^{(s)} = 0$ (a linear signal) also yielded satisfactory results from using the I_S criterion for model selection.

IV. Results using Climate Data

The method is also tested on two real climate datasets, the tropical Pacific sea surface temperature (SST) and the North American surface air temperature (SAT). The monthly SST data on a $2^\circ \times 2^\circ$ grid for the period 1948-2005 came from the Extended Reconstructed Sea Surface Temperatures (ERSST version 2) dataset [13] (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

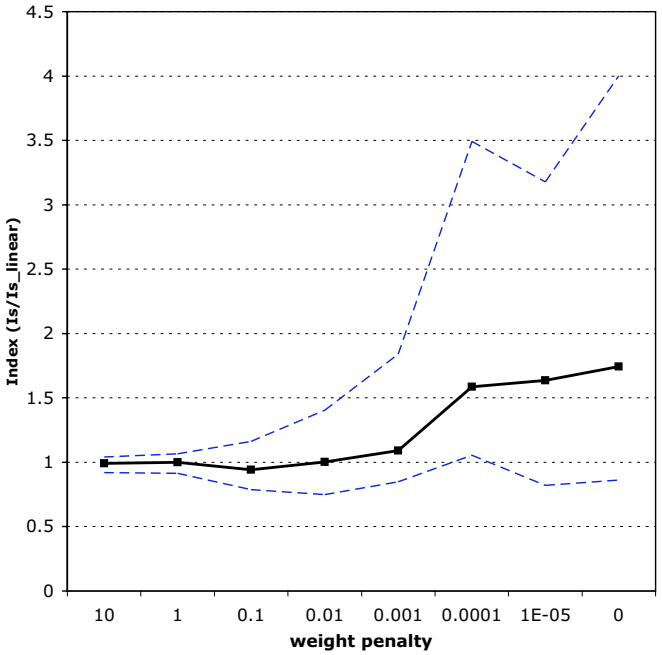

Fig. 3. The index I_S from NLPCA (normalized by I_S computed from the linear PCA method) for various values of the weight penalty parameter P. The solid curve is the average over 12 noisy datasets, with the dashed curves indicating the maxium and minimum values over the set of 12.

tropical Pacific domain of 124°E-70°E, 20°S-20°N. The 7 leading PCs containing 86.5% of the variance were retained.

The monthly land SAT data, from the Climate Research Unit (CRU) at the University of East Anglia, UK [9] (http://www.cru.uea.ac.uk/cru/data/hrg.htm), on a $0.5° \times 0.5°$ grid were chosen for N. America (north of 20°N) over the period 1950-2004. The climatological seasonal cycle was removed to yield the SAT anomalies, with only the 5 winter months (November–March) used. PCA was used to compress the data, with 7 leading PCs (containing 87.7 % of the variance) retained. Because of mid-latitude weather systems, the winter SAT dataset is much noisier than the tropical Pacific SST, which is dominated by the El Niño-Southern Oscillation phenomenon.

The leading PCs were input into the NLPCA model. For SST, the minimum in I_S occurred at $P = 10^{-4}$, while for the noisier SAT, the minimum occurred at $P = 0.1$ (Figure 6). The NLPCA solutions for SST and SAT, displayed in Figures 7 and 8, respectively, showed that a more curved solution was justified by the I_S criterion for SST, but because of the larger P selected for SAT, only a less curved solution can be justified for the noisier SAT data. Note that fluctuations in I_S may produce a second (and possibly lower) minimum at smaller P, as is found for SAT in Figure 6. In general, these addition minima yield zigzag solutions and should be ignored.

V. Conclusions

For NLPCA, the overfitting problem with noisy data is much more serious than for nonlinear regression, since for

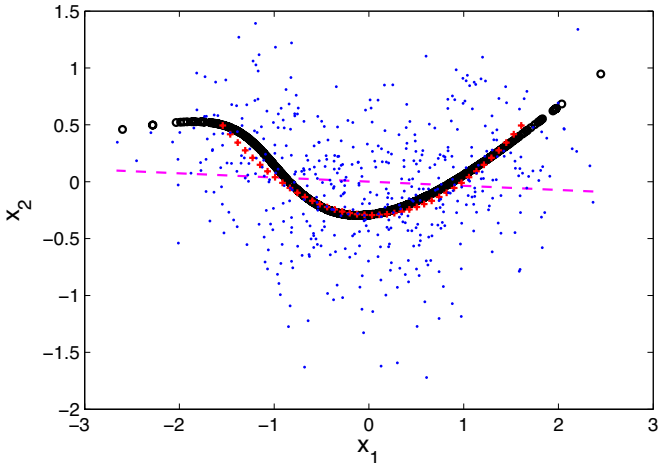

Fig. 4. The NLPCA solution (shown as densely overlapping circles) for the synthetic dataset (dots), with $P = 0.1$. The quadratic signal curve is indicated by "+" and the linear PCA solution by the dashed line.

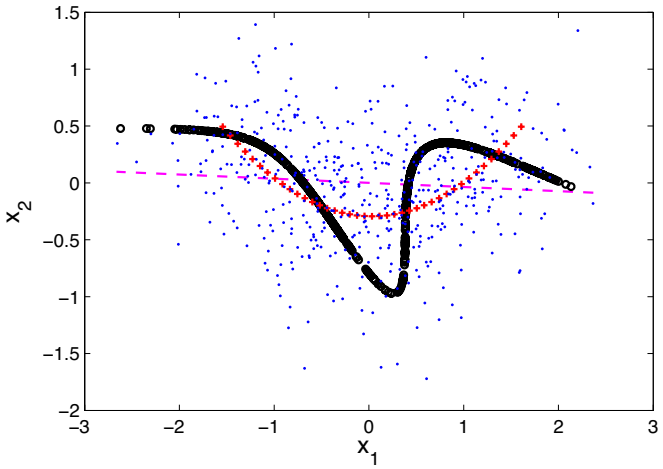

Fig. 5. NLPCA of the same dataset as in Figure 4, but with $P = 10^{-5}$.

NLPCA, overfitting can arise not only from data scarsity, but also from the geometry of the data. Using independent data to validate the MSE from the various models is not a viable method for choosing the appropriate weight penalty parameter P to control overfitting. Instead, we propose a new index I_S for detecting the projection of neighbouring points to distant parts of the NLPCA curve, and use the index to choose the appropriate P. The strategy for choosing the most appropriate P is to run the model repeatedly from large to small P, locate the first minimum of I_S, and choose the P at this minimum, or more generally choose the smallest P with I_S within $1 + \epsilon$ times this minimum I_S value. Tests with synthetic data and with climate data indicated that this criterion is effective in model selection, discarding unjustifiable zigzag solutions. Although the tests were performed with an NN model, this criterion for NLPCA model selection should work well with other implementations of nonlinear PCA (e.g. using kernel methods [11]). Finally, for future work, NLPCA may be made more resistant to

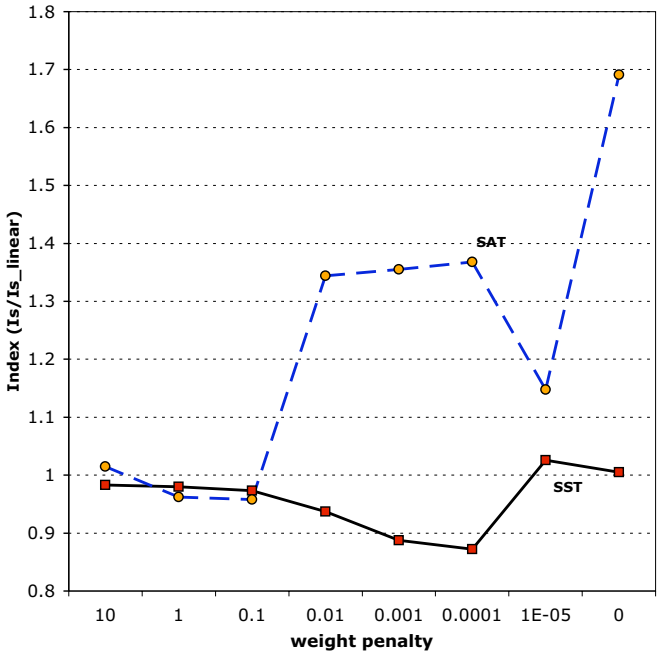

Fig. 6. The index I_S from NLPCA (normalized by I_S from the linear PCA) for various values of the weight penalty parameter P, computed for SST (solid) and SAT (dashed).

outliers by replacing the L_2 norm by an L_1 in the objective function, as is done in nonlinear regression by support vector machines [12].

APPENDIX

With the input variables forming the 0th layer of the network in Figure 2, 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.

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

where $\mathbf{w}_j^{(i)}$ is a vector of weight parameters and $b_j^{(i)}$ a bias parameter, and the 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, [5] used

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

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. [4] found that

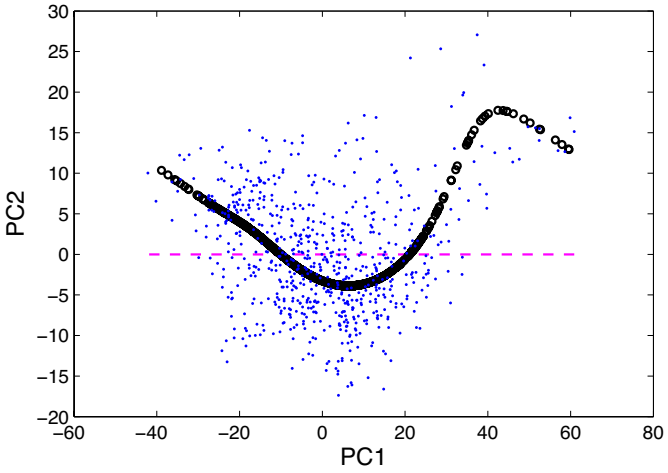

Fig. 7. The NLPCA solution (shown as densely overlapping circles) for the SST anomaly data (dots), with $P = 10^{-4}$. The solution is shown only in the PC1-PC2 plane, though it is actually a curve in the 7-dimension space spanned by the 7 leading PCs.

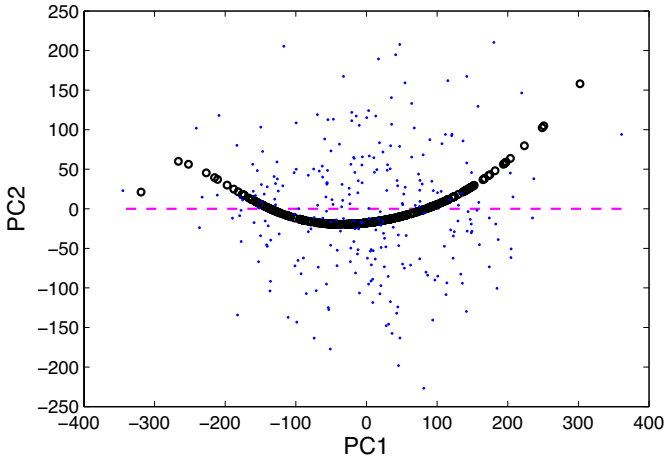

Fig. 8. The NLPCA solution for the SAT anomaly data (dots), with $P = 0.1$, shown in the PC1-PC2 plane.

penalizing just the first layer of weights is sufficient to limit the nonlinear modelling capability of the model. Through nonlinear optimization, the values of the weight and bias parameters are solved (see [5] for more details).

ACKNOWLEDGMENT

The SST and SAT data were obtained via Aiming Wu. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford: Clarendon press, 1995.

[2] B. Christiansen, "A cautionary note on the use of Nonlinear Principal Component Analysis to identify circulation regimes," *Journal of Climate*, vol. 18, pp. 4814-4823, 2005.

[3] F. D. Foresee and M. T. Hagan, "Gauss-Newton approximation to Bayesian regularization", *Proceedings of the 1997 International Joint Conference on Neural Networks*, 1997.

[4] W. W. Hsieh, "Nonlinear principal component analysis by neural networks" *Tellus*, vol. 53A, pp. 599-615, 2001.

[5] W. W. Hsieh, "Nonlinear multivariate and time series analysis by neural network methods", *Reviews of Geophysics*, vol. 42, RG1003, doi:10.1029/2002RG000112, 2004.
[6] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks", *AIChE Journal*, vol. 37, pp. 233-243, 1991.
[7] D. J. C. MacKay, " Bayesian interpolation", *Neural Computation*, vol. 4, pp. 415-447, 1992.
[8] E. C. Malthouse, "Limitations of nonlinear PCA as performed with generic neural networks", *IEEE Transactions on Neural Networks*, vol. 9, pp. 165-173, 1998.
[9] T. D. Mitchell, T. R. Carter, P. D. Jones, M. Hulme, and M. New, "A comprehensive set of climate scenarios for Europe and the globe", *Tyndall Centre Working Paper 55*, pp. 30, 2004.
[10] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes*, Cambridge: Cambridge Univ. Press, 1986.
[11] B. Schölkopf, A. Smola, A. and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation*, vol. 10, pp. 1299-1319, 1998.
[12] J. Shawe-Taylor and N. Cristianini *Kernel Methods for Pattern Analysis*, Cambridge: Cambridge Univ. Pr., 2004.
[13] T. M. Smith and R. W. Reynolds, "Improved extended reconstruction of SST (1854-1997)", *Journal of Climate*, vol. 17, pp. 2466-2477, 2004.