Nonlinear Dimensionality Reduction by Multi Layer Perceptron Using Superposed Energy

Takashi TAKAHASHI† and Ryuji TOKUNAGA‡

†*JSPS Research Fellow, University of Tsukuba, Ibaraki 305-8573 Japan*
Takashi@Takahashi.com
‡*Institute of Information Sciences and Electronics, University of Tsukuba, Ibaraki 305-8573 Japan*

Abstract— We investigate an energy function for MLP called *superposed energy*. Applying to autoassociative learning of a sandglass-type MLP, it can adaptively adjust the effective number of the bottleneck-layer units to the intrinsic dimensionality of nonlinear data, and the optimal dimensionality reduced representation can be extracted after learning.

I. Introduction

Dimensionality reduction is an important technique to extract a small number of significant features from large-scale data scattered around high dimensional space. In a variety of methods, the linear method called Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) has been successfully applied to wide areas. Nonlinear time series analysis is one of such examples[1]. For nonlinear data, however, the application of such linear methods is problematic in some cases. We consider a nonlinear dimensionality reduction using autoassociative learning of MLP (Multi Layer Perceptron).

On autoassociative learning of MLP, it is shown that a three-layer linear perceptron realizes dimensionality reduction which is equivalent to PCA[2], and it is also known that a sandglass-type (typically, five-layer) MLP consisting of nonlinear units can perform nonlinear dimensionality reduction[3, 4]. In order to obtain good performance in such autoassociative learning, it is crucial to choose the appropriate number of units for the bottleneck-layer adjusting to the intrinsic dimensionality of data. In general, however, it is impossible since the intrinsic dimensionality is not known a priori. Therefore it is necessary to estimate the optimal number of the bottleneck-layer units.

In this paper, we investigate an energy function for sandglass-type MLP called *superposed energy*. It is shown that the energy can adjust the effective number of bottleneck-layer units adaptively, and that the optimal dimensionality reduced representation can be extracted after learning.

II. Nonlinear dimensionality reduction by MLP

Dimensionality reduction can be expressed by the following estimation problem.

A set of N dimensional values

$$\left\{ \boldsymbol{z}_t \in R^N \right\}_{t=1}^{T} \tag{1}$$

is on an unknown L dimensional surface $S^* \subset R^N$, and a set of noise-added data

$$\{\boldsymbol{x}_t = \boldsymbol{z}_t + \boldsymbol{n}_t\}_{t=1}^{T} \tag{2}$$

is observed, where $\boldsymbol{n}_t$ is the noise which follows the spherical Gaussian distribution $N(\mathbf{o}, \sigma^2 I)$. From the observed values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, estimate the surface S^* and the true values $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$.

If the true surface S^* is considered to be a hyperplane, conventional linear methods such as PCA (Principal Component Analysis) are applicable to this problem. However, such linear methods fail to estimate the true surface if S^* has some nonlinear properties. For instance, if S^* is a parabola in R^2, PCA cannot extract any significant one-dimensional components. Hence, in order to deal with nonlinear data, it is necessary to develop nonlinear dimensionality reduction methods which are able to extract and represent nonlinear relations behind data.

Autoassociative learning of sandglass-type MLP is one of the methods to realize nonlinear dimensionality reduction[3, 4]. Let us consider a five-layer sandglass-type MLP which consists of N input units, N output units, and H middle(bottleneck)-layer units. These three layers consists of linear units, while the second and the fourth layers consist of nonlinear units. The number of the middle layer units, H, is set to be less than N, and this layer works as a bottleneck representing input data in lower dimensionality. The MLP is trained to approximate the identity mapping of data by minimizing the error

$$E = \frac{1}{T} \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{y}_t||^2 \tag{3}$$

where $\boldsymbol{y}_t$ denotes the output of the MLP corresponding to the input data $\boldsymbol{x}_t$. This MLP then fits each data on a H dimensional curved surface S, and outputs an N dimensional value approximating the data. Thus dimensionality reduced data representation is organized as the outputs of the bottleneck-layer.

In the above estimation problem, the true dimensionality L is generally unknown, so that it needs to be estimated for performing efficient dimensionality reduction. Therefore the number of the bottleneck-layer units of sandglass-type MLP should be optimized comparing all possible values of H according to some criteria. As such a criterion, however, the learning error E is inappropriate, since it does not reflect the true performance of dimensionality reduction. The true error which we must optimize to find L is the distance between the true value $\boldsymbol{z}$ and its estimate $\boldsymbol{y}$:

$$\mathcal{E}^* = \mathrm{E}\left[||\boldsymbol{y} - \boldsymbol{z}||^2\right] \quad . \tag{4}$$

This equation includes the non-observable value $\boldsymbol{z}$, and we must estimate the value. Hiraoka and Yoshizawa derived the estimates of $\mathcal{E}^*$ under the condition that both S and S^* are well approximated to planar surfaces[5]. When the variance of the noise, σ^2, is known, the estimate of $\mathcal{E}^*$ is given by

$$\mathcal{E}_K = E + (2H - N)\sigma^2 \quad . \tag{5}$$

On the other hand, in the case of unknown σ^2, it is estimated as

$$\tilde{\sigma}^2 = \frac{E}{N - H} \quad , \tag{6}$$

and then the estimate of $\mathcal{E}^*$ is obtained by substituting $\tilde{\sigma}^2$ into Eq. (5) :

$$\mathcal{E}_U = \frac{HE}{N - H} \quad . \tag{7}$$

It is shown that both of $\mathcal{E}_K$ and $\mathcal{E}_U$ are minimum when the dimensionality of MLP, H, is equal to L. Therefore, according to this criteria, it is able to choose the MLP which has the optimal dimensionality among the MLPs which consist of different numbers of bottleneck-layer units. However, this can become intractable if H is large.

III. Superposed energy

In order to enable sandglass-type MLP to adaptively adjust the effective dimensionality (the number of the bottleneck-layer units to be used), we have investigated an energy function called *superposed energy*[6, 7]. When the energy is used for autoassociative learning of three-layer linear perceptron, it is shown that each bottleneck-layer unit gives the output corresponding to the principal component of input data, and that these outputs are arranged in the order of their magnitude of contribution[6]. In this section, a brief description of superposed energy is given, and its application to nonlinear dimensionality reduction is discussed including the use of Hiraoka–Yoshizawa's criteria.

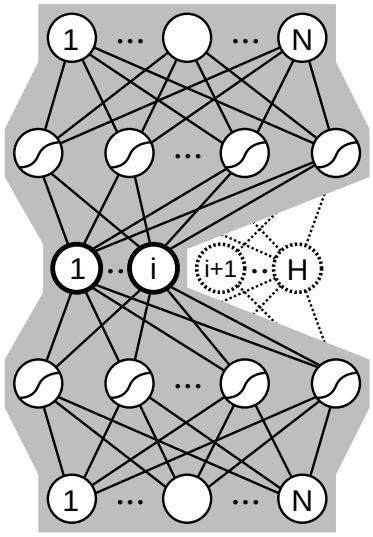

Figure 1: A sandglass-type MLP and its sub-perceptron. The shading region represents the i-th sub-perceptron.

Superposed energy is defined as follows. Let f_i $(i = 1, 2, \ldots, H)$ denote a function corresponding to the MLP which approximates the identity map only with the first i units among all the bottleneck-layer units. The function f_i is referred as the i-th *sub-perceptron* (Fig. 1). In the same way as the error (3) is defined, the error of the sub-perceptron f_i can also be defined[1] by

$$E_i = \frac{1}{T}\sum_{t=1}^{T} ||x_t - f_i(x_t)||^2 \quad . \tag{8}$$

Then *superposed energy* is defined by a weighted sum of all E_is :

$$F = \sum_{i=1}^{H} \beta_i E_i \tag{9}$$

where β_i is a positive constant. By using superposed energy, the bottleneck-layer units with smaller indices tend to extract more important feature components, and it is able to adjust the number of bottleneck-layer units to be used for generating the outputs. Therefore one can select the dimensionality of the fitting surface S after learning.

Hiraoka–Yoshizawa's criteria are also applicable to a MLP trained by superposed energy. The most immediate way is to calculate Eqs. (5) and (7) for each

[1] Note that E_H is Eq. (3) itself, which is the error function used for conventional autoassociative learning.

sub-perceptron. The estimates of $\mathcal{E}^*$ of the i-th sub-perceptron, $\mathcal{E}_K(i)$ and $\mathcal{E}_U(i)$, are then given by

$$\mathcal{E}_K(i) = E_i + (2i - N)\sigma^2 \quad (10)$$

$$\mathcal{E}_U(i) = \frac{HE_i}{N - i} \quad , \quad (11)$$

respectively. However, Eq. (11), as well as Eq. (7), has the problem of underestimating $\mathcal{E}^*$ in the case of $i < L$, since Eq. (6) overestimates σ^2 in that condition[5]. To reduce the estimation error, we consider the following equation as an alternative of Eq. (11).

$$\widehat{\mathcal{E}}_U(i) = E_i + (2i - N)\tilde{\sigma}^2_{min} \quad (12)$$

where

$$\tilde{\sigma}^2_{min} = \min_i \frac{E_i}{N - i} \quad . \quad (13)$$

IV. Experiment

This section shows the results of an application of the proposed method to estimate the intrinsic dimensionality of a nonlinear time series data and to remove its noise. Data sets were generated from the Rössler system :

$$\begin{cases} \dot{x}_1 = -y - z \\ \dot{y} = x + ay \\ \dot{z} = bx - cz + xz \end{cases} \quad (14)$$

where the parameters are set at the values $a = 0.36$, $b = 0.4$, $c = 4.5$. To obtain signals with noise, the x coordinate of the trajectory[2] was sampled and added Gaussian noise whose standard deviation is 10% of that of the signal. From this one-dimensional time series, an 8-dimensional input data for a MLP was constructed by the method of delays with an interval of four sample times. The MLP has five layers composed of 8–32–7–32–8 units, respectively, therefore, $N = 8$ and $H = 7$. It is trained using superposed energy, provided that the value of each β_i is modified adaptively in learning (see appendix).

After learning converged, the learning error E_i, the true error $\mathcal{E}^*$, and the estimate values $\mathcal{E}_U$ and $\widehat{\mathcal{E}}_U$ were calculated with each sub-perceptron. Tab. 1 shows the results. While E_i decreases monotonically, $\mathcal{E}^*$, $\mathcal{E}_U$ and $\widehat{\mathcal{E}}_U$ take minimum values at $i = 4$. It is seen that the diferences between $\mathcal{E}^*$ and $\widehat{\mathcal{E}}_U$ are small enough showing the significant improvement in comparison with $\mathcal{E}_U$.

Next, signal-to-noise ratio (SNR) is calculated for the time series which is reconstructed from the output of each sub-perceptron. In order to make a comparison, the conventional dimensionality reduction method using PCA is also applied to the same data. The results are given in Fig. 2. The dimensionality giving the best SNR coincides with that giving the minimum of $\widehat{\mathcal{E}}_U$. It is shown that noise reduction can be done by selecting the appropriate dimensionality. That is also illustrated in Fig. 3. On the other hand, little enhancement is seen in the case of PCA.

[2]The trajectory of the Rössler attractor was generated using the fourth order Runge–Kutta method with step-size 0.01, and every 20-th point was used.

Table 1: Results of autoassociative learning applied to nonlinear data

i	E_i	$\mathcal{E}_U$	$\widehat{\mathcal{E}}_U$	$\mathcal{E}^*$
1	0.6735	0.0952	0.6638	0.7343
2	0.0896	0.0272	0.0831	0.0716
3	0.0302	0.0155	0.0270	0.0245
4	0.0066	0.0067	0.0066	0.0070
5	0.0049	0.0083	0.0081	0.0081
6	0.0034	0.0105	0.0098	0.0093
7	0.0019	0.0166	0.0117	0.0106

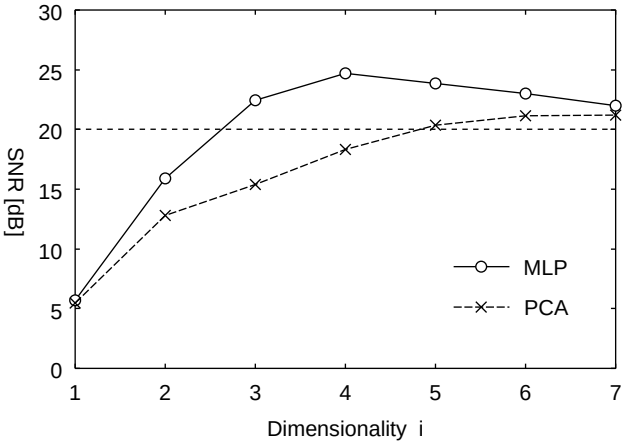

Figure 2: SNR of the reconstructed time series. The horizontal dashed line shows the value of the observed time series.

V. Conclusion

We have investigated a nonlinear dimensionality reduction method which is able to adjust the effective dimensionality of sandglass-type MLP to the intrinsic dimensionality of data. In comparison with conventional autoassociative learning, this method has the advantage of requiring only one MLP to be trained instead of all possible MLPs with different numbers of bottleneck-layer units. It produces a reduction of computational costs for learning. However, the costs of training a sandglass-type MLP is still expensive. Future work will include the investigation of the method which require less computational costs.

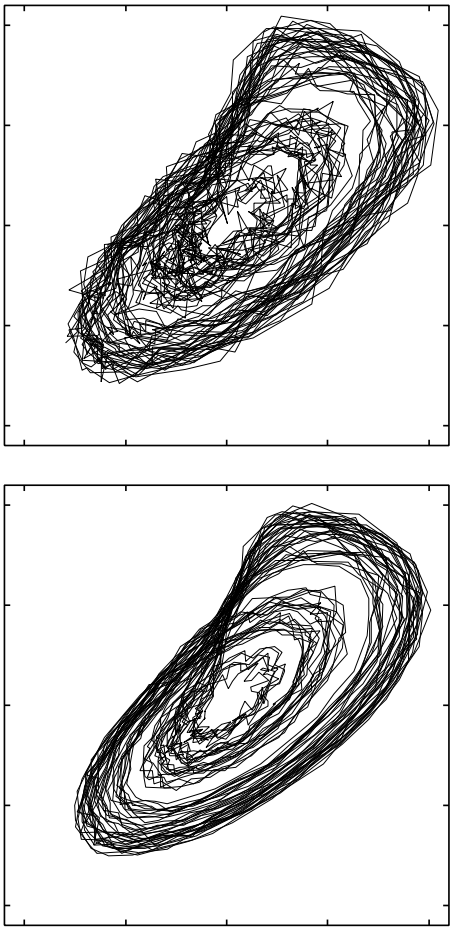

Figure 3: The trajectories of the observed data(upper) and its reconstruction using the fourth subperceptron(lower).

Appendix

In the case of nonlinear MLP, learning speed depends on the choice of the values of β_i $(i = 1, 2, \ldots, H)$. It was empirically found that better results are obtained by adapting these values according to the estimated true errors which are calculated while learning proceeds, for instance,

$$\beta_i \propto \frac{i}{\widehat{\mathcal{E}}_U} \quad . \tag{15}$$

Thus we adopted the following scheme: modify the values of β_i according to Eq. (15) and

$$\sum_{i=1}^{H} \beta_i = 1 \tag{16}$$

at an appropriate learning interval.

Acknowledgments

This research was supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

References

[1] D. S. Broomhead, and G. P. King, "Extracting qualitative dynamics from experimental data," *Physica* 20D, pp.217–236, 1986.

[2] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Networks*, vol.2, pp.53–58, 1989.

[3] S. Usui, S. Nakauchi, and M. Nakano, "Internal color representation acquired by a five-layer neural network," *Proc. Int. Conf. Artificial Neural Networks*, vol.1, pp.867–872, 1991.

[4] D. DeMers and G. Cottrell, "Non-linear dimensionality reduction," *Advances in Neural Information Processing Systems 5*, pp.580–587, 1993.

[5] K. Hiraoka and S. Yoshizawa, "On selection of the bottleneck size in autoassociative learning," *Technical report of IEICE*, NC97–139, 1998 (in Japanese).

[6] T. Takahashi and R. Tokunaga, "Removing the Redundancy of Perceptrons in Terms of a Simple Energy Function," *Proc. Int. Conf. Neural Information Processing*, vol.1, pp.271–274, 1997.

[7] T. Takahashi and R. Tokunaga, "Energy Functions for Efficient Nonlinear Dimensionality Reduction by Multi Layer Perceptrons," *Proc. Int. Conf. Neural Information Processing*, vol.1, pp.494–497, 1998.