Nonlinear PCA by Neural Network Models

Yoshio Takane
McGill University

1 Introduction

Feed-forward neural network (NN) models and statistical models have much in common (e.g., Cheng & Titterington, 1994; Ripley, 1993). The former can be viewed as approximating nonlinear functions that connect inputs to outputs. Many statistical techniques can also be viewed as approximating functions (often linear) connecting predictor variables to criterion variables. It is quite natural then that nonlinear extensions of linear statistical techniques exploit various developments in NN models. In this paper we discuss one particular technique, nonlinear principal component analysis (PCA) by NN models, and examine its properties focussing on its ability to recover underlying structures. Previous work by the present author focussed on the mechanism of nonlinear function approximations by the Cascade Correlation Learning Network (Fahlman & Lebiere, 1990) for two specific tasks, the continuous xor problem (Takane, et al., 1994) and the learning of first and second pronouns (Oshima-Takane, et al., 1995).

2 Nonlinear PCA by the 5-Layer Neural Network Model

It is well known (e.g., Baldi & Hornik, 1989; Diamantras & Kung, 1994) that a 3-layer (including the input and the output layers) neural network model with linear transfer functions at the hidden layer has the rank reducing capability, where the specific rank is effected by the number of units at the hidden layer. This is a network version of reduced-rank (RR) regression (Anderson, 1951), which is also known as PCA of instrumental variables (Rao, 1964) and redundancy analysis (van den Wollenberg, 1977). The usual PCA follows when inputs and outputs coincide in RR regression analysis (e.g., ten Berge, 1993). The network version of PCA is not interesting in itself, because there are other more effcient and precise algorithms available. It becomes interesting when the model is extended to nonlinear PCA by including two additional hidden layers with nonlinear transfer functions, one between the input layer and the middle layer and the other between the middle layer and the output layer. Fig. 1 shows the basic design of this extended 5-layer network model. The 5-layer network model was proposed (apparently independently) by several authors at about the same time (Irie & Kawato, 1989; Katayama & Ohyama, 1989; Kramer, 1991; Morishima, et al., 1990; Oja, 1991; Usui, et al., 1991), and has been applied to PCA of faces (DeMers & Cottrell, 1993) and facial expressions of emotions (Ueki, et al., 1994).

The model is interesting because it allows joint multivariate nonlinear transformations of input variables, thereby capturing interaction effects among them. Previous methods (e.g., Kruskal & Shepard, 1974; Young et al., 1978; Gifi, 1990) allow only variablewise monotonic transformations.

3 Example 1

In the first example, we constructed a two-cycle helix by first generating a vector of x ranging from 0 to 4π in steps of $\pi/10$, then defining a matrix of $\mathbf{X} = [sin(\mathbf{x})cos(\mathbf{x})\mathbf{x}]$. The helix is depicted in Fig. 2. Matrix $\mathbf{X}$ was fed into the network algorithm with 6 units each in the second and the fourth layers and 1 unit in the middle layer. The network recovered the original helix (i.e., $\mathbf{X}$) almost perfectly at the output layer. Fig. 3 displays the plot of component scores recovered at the middle layer against the original $\mathbf{x}$ used to generate the helix data. The recovered $\mathbf{x}$ is strictly monotonic with the original $\mathbf{x}$. That is, $\mathbf{x}$ is recoverable only up to a monotonic transformation. This makes sense, because only a one-to-one function preserves nonlinear information, and the only continuous one-to-one function is strictly monotonic.

4 Example 2

When a set of binary items form a perfect unidimensional scale in Guttman's (1941) sense, a group of subjects responding to the items show such a characteristic pattern as shown in Table 1 (Iwatsubo, 1987). For such data the third kind of quantification method (Q3) yields the constant first eigenvector (order 0), the second eigenvector whose

[1] *In the Proceeding of the 63rd Annual Meeting of the Japan Statistical Society (pp. 258-260).*

elements are linear (order 1) with the item order, and the remaining eigenvectors which are successive polynomial functions of the second eigenvector. More than one eigenvalue (excluding the unit eigenvalue associated with the constant eigenvector) are nonzero despite the fact that the underlying mechanism that generates the data set is known to be unidimensional. (This phenomenon is known as the Guttman effect.)

Since all subsequent eigenvectors are one-to-many nonlinear transformations of the second eigenvector (and since the constant vector is taken care of by bias parameters in network models), a single component strictly monotonic with the second eigenvector should be sufficient to acccount for all variations in the data set in Table 1. The input matrix for Q3 was submitted to the network algorithm, with 10 units in each of the second and the fourth layers and a single unit in the middle layer, which almost perfectly recovered the original input matrix. Fig. 4 shows the component obtained in the middle layer, which is strictly monotonic with the second eigenvector. One may argue that all eigenvectors are also some nonlinear transformations of all other eigenvectors (except the constant vector), so why should we always obtain a monotonic function of the second eigenvector as the component? Aren't monotonic functions of all other eigenvectors also legitimate for the component? No, because no other eigenvectors are one-to-one with the second eigenvector, so that they fail to produce the second eigenvector component needed to reproduce the input matrix.

It would be of interest to examine how the NN model responds to the kind of structures discussed by Okamoto (1994).

5 Example 3

Kruskal & Shepard (1974) used a set of 30 cylinders in their demonstration of nonmetric PCA. They generated the cylinders by systematically varying their altitude (a) and base area (b), which were in turn generated by pairs of b' and a' supplied by Coombs & Kao (1960) according to $a = exp[c(b' + .6)]$ and $b = exp[c(a' + .6)]$ where $c = 1.7325$. They then defined 12 variables which are functions of a & b, and measured the cylinders on the 12 variables. These variables are listed in Table 3. The derived data matrix was subjected to nonmetric PCA which almost perfectly recovered pairs of a' and b' used to generate the data. The recovered a' and b' in turn reproduced monotonically transformed data found by the method.

We generated our third example in a similar way to the above except that 169 pairs of a' and b' were obtained by factorial combinations of each varying from $-.6$ to $.6$ in steps of .1, and that after the cylinders were measured on the same 12 variables, they were subjected to an arbitrary linear transformation to obtain a completely different set of 12 variables. Two examples of the latter variables are shown in Fig. 5 (Small $\circ$'s indicates cylinders used in Shepard & Kruskal.), which are no longer monotonic with a' or b'. Nonmetric PCA will certainly have a great deal of difficulty for this kind of data. However, the 5-layer NN model with 18 units in each of the second and the fourth layers and two units in the middle layer could almost perfectly recover the input data. (Little changes were observed when the number of units in the second and the fourth layers was reduced to 12 each.) Fig. 6 depicts the component scores recovered at the middle layer (shown by +) plotted against the original a' and b' (shown by $\circ$). Again, the recovered components are monotonic functions of the original a' and b', (and are in fact more like a and b derived from a' and b').

6 References

Anderson, T.W. (1951). AMS; Baldi, P. & Hornik, K. (1989). *Neural Networks*; Cheng, B. & Titternington, D.M. (1994). *Stat. Sci.*; Coombs, C.H. & Kao, R.C. (1960). *Psyhometrika*; DeMers, D., & Cottrell, G. (1993). *NIPS 5*; Diamantras, K. I. & Kung, S-Y., (1994). *IEEE Trans. on NN*; Fahlman, S.E. & Lebiere, C. (1990). *NIPS 2*; Gifi, A. (1990). Wiley; Guttman, L. (1941). In Horst, P. Soc. Sci. Res. Council; Irie, B. & Kawato, M. (1989). *Shingakugiho*; Iwatsubo, S. (1987). Asakura; Katayama, Y. & Ohyama, K. (1989). *Shingakushunkizendai*; Kramer, M. (1991). *AIChE Journal*; Kruskal, J.B. & Shepard, R.N. (1974). *Psychometrika*; Morishima, S., Nakayama, H., Shimizu, S., Katayama, Y. & Harashima, H. (1990). *Shingakugiho*; Oja, E. (1991). In T. Kohonen, et al. (Eds.) *ANN*; Okamoto, M. (1994). *Ottemondaigaku Kiyo*; Oshima-Takane, Y., Takane, T. & Shultz, T.R. (1995). Submitted to *NIPS*. Rao, C.R. (1964). *Sankhya A*; Ripley, B. (1993). In O.E. Barndorff-Nielsen, et al. (Eds.) Chapman & Hall; Takane, Y., Oshima-Takane, Y. & Shultz, T.R. (1994). *The Proceedings of Japan Classification Soc. Meeting*; ten Berge, J.M.S. (1993). DSWO Press; Ueki, N., Morishima, S., Yamada, H. & Harashima, H. *Shingakuronbun*; Usui, S., Nakauchi, S. & Nakano, M. (1991). In T. Kohonen, et al. (Eds.), *ANN*; van den Wollenberg, A.L. (1977). *Psychometrika*; Young, F.W., Takane, Y. & de Leeuw (1978). *Psychometrika.*

Fig. 1

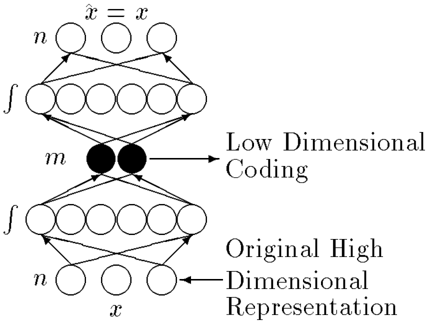


Fig. 2

The helix (two cycles)

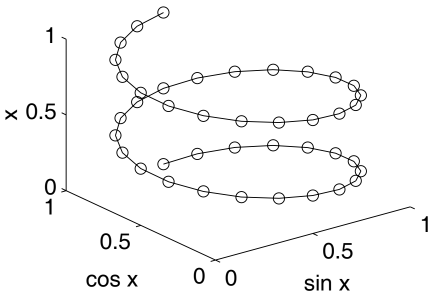


Fig. 3

The helix problem

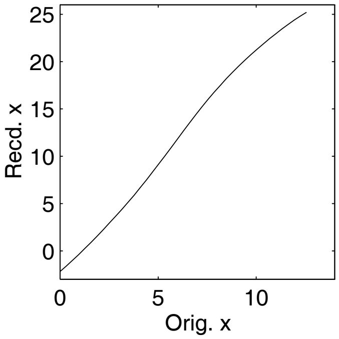


Table 1

i \ j	1	2	3	4	5	6	7	8	9	10
1	1	1								
2		1	1							
3			1	1						
4				1	1					
5					1	1				
6						1	1			
7							1	1		
8								1	1	
9									1	1
10										1

Fig. 4

Guttman scale

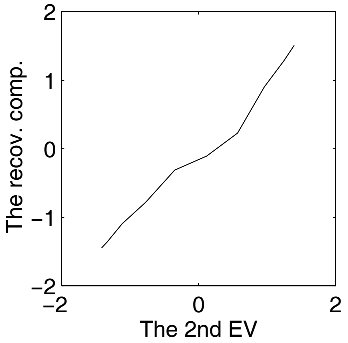


Table 2

	2 0.9729	3 0.8946	4 0.7735	5 0.6227	6 0.4587	7 0.2992	8 0.1614	9 0.0603	10 0.0068
1	-1.4142	1.4142	1.4142	-1.4142	1.4142	-1.4142	-1.4142	-1.4142	-1.4142
2	-1.3376	1.1160	0.7735	-0.3472	-1.1168	0.5681	0.9578	1.2438	1.3949
3	-1.1160	0.3472	-0.5681	1.2438	-1.3949	0.9578	0.1168	-0.7735	-1.3376
4	-0.7735	-0.5681	-1.3949	0.9578	0.3472	-1.3376	-1.1160	0.1168	1.2438
5	-0.3472	-1.2438	-0.9578	-0.7735	1.3376	0.1168	1.3949	0.5681	-1.1160
6	0.1168	-1.3949	0.3472	-1.3376	-0.5681	1.2438	-0.7735	-1.1160	0.9578
7	0.5681	-0.9578	1.3376	0.1168	-1.2438	-1.1160	-0.3472	1.3949	-0.7735
8	0.9578	-0.1168	1.1160	1.3949	0.7735	-0.3472	1.2438	-1.3376	0.5681
9	1.2438	0.7735	-0.1168	0.5681	1.1160	1.3949	-1.3376	0.9578	-0.3472
10	1.3949	1.3376	-1.2438	-1.1160	-0.9578	-0.7735	0.5681	-0.3472	0.1168

Table 3

VARIABLE		FORMULA
1	ALTITUDE	a
2	BASE AREA	b
3	CIRCUMFERENCE	$(2\sqrt{\pi})\, b^{1/2}$
4	SIDE AREA	$(2\sqrt{\pi})\, ab^{1/2}$
5	VOLUME	ab
6	MOMENT OF INERTIA	$(1/2\pi)\, ab^2$
7	SLENDERNESS RATIO	$(1/\sqrt{2\pi})\, ab^{-1/2}$
8	DIAGONAL-BASE ANGLE	$TAN^{-1}[(\sqrt{\pi}/2)\, ab^{-1/2}]$
9	DIAGONAL-SIDE ANGLE	$COT^{-1}[(\sqrt{\pi}/2)\, ab^{-1/2}]$
10	ELECTRICAL RESISTANCE	ab^{-1}
11	CONDUCTANCE	$a^{-1}b$
12	TORSIONAL DEFORMABILITY	$(2\pi)\, ab^{-2}$

Fig. 5a

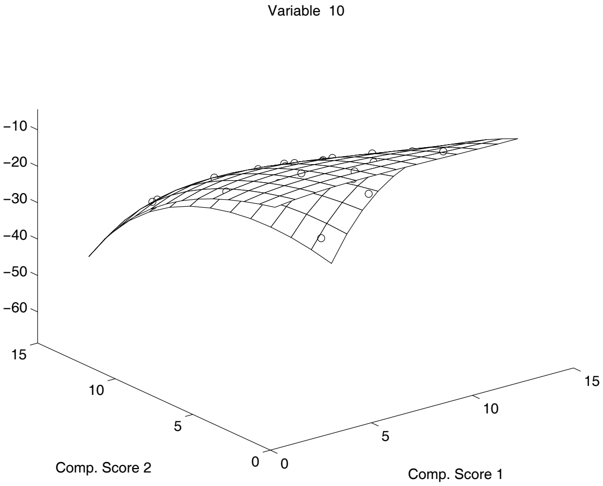


Fig. 5b

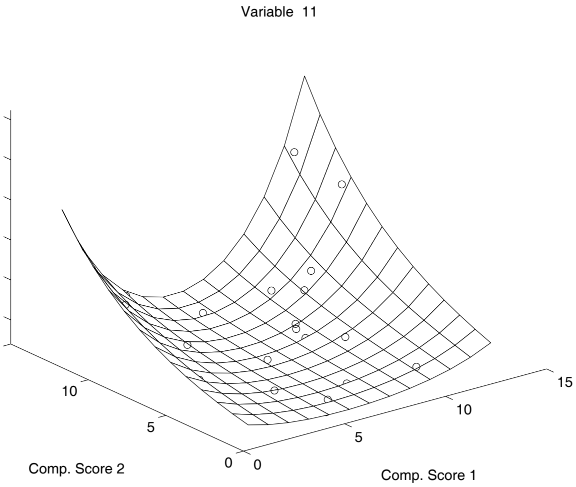
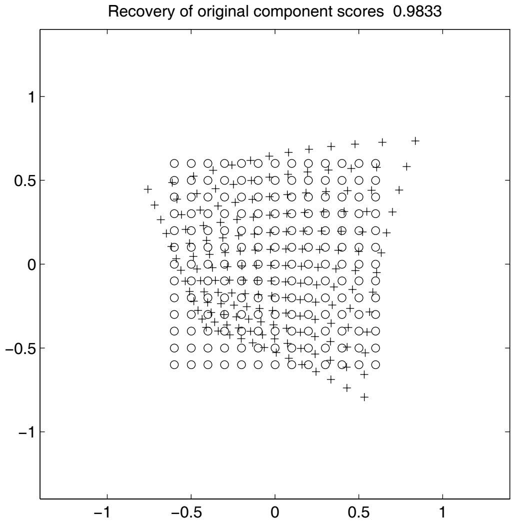


Fig. 6


Recovery of original component scores 0.9833