Validation of Nonlinear PCA

Matthias Scholz

Abstract   Linear principal component analysis (PCA) can be extended to a nonlinear PCA by using artificial neural networks. But the benefit of curved components requires a careful control of the model complexity. Moreover, standard techniques for model selection, including cross-validation and more generally the use of an independent test set, fail when applied to nonlinear PCA because of its inherent unsupervised characteristics. This paper presents a new approach for validating the complexity of nonlinear PCA models by using the error in missing data estimation as a criterion for model selection. It is motivated by the idea that only the model of optimal complexity is able to predict missing values with the highest accuracy. While standard test set validation usually favours over-fitted nonlinear PCA models, the proposed model validation approach correctly selects the optimal model complexity.

Keywords   Nonlinear PCA · Neural networks · Model selection · Validation

1 Introduction

Nonlinear principal component analysis[1–3] is a nonlinear generalization of standard principal component analysis (PCA). While PCA is restricted to linear components, nonlinear PCA generalizes the principal components from straight lines to curves and hence describes the inherent structure of the data by curved subspaces. Detecting and describing nonlinear structures is especially important for analysing time series. Nonlinear PCA is therefore frequently used to investigate the dynamics of different natural processes[4–6]. But validating the model complexity of nonlinear PCA is a difficult task[7]. Over-fitting can be caused by the often limited number of available samples; moreover, in nonlinear PCA over-fitting can also occur by the intrinsic geometry of the data, as shown in Fig. 5, which cannot be solved by increasing the number of samples. A good control of the complexity of the nonlinear

M. Scholz (✉)
Edmund Mach Foundation, Research and Innovation Center, Via Edmund Mach 1,
38010 San Michele all'Adige, TN, Italy
e-mail: matthias.scholz@fmach.it

PCA model is required. We have to find the optimal flexibility of the curved components. A component with too little flexibility, an almost linear component, cannot follow the complex curved trajectory of real data. By contrast, a too flexible component fits non-relevant noise of the data (over-fitting) and hence gives a poor approximation of the original process, as illustrated in Fig. 1a. The objective is to find a model whose complexity is neither too small nor too large.

Even though the term nonlinear PCA (NLPCA) is often referred to the auto-associative neural network approach, there are many other methods which visualise data and extract components in a nonlinear manner [8]. *Locally linear embedding* (LLE) [9,10] and *Isomap* [11] visualise high dimensional data by projecting (embedding) them into a two or three-dimensional space. *Principal curves* [12] and *self organising maps* (SOM) [13] describe data by nonlinear curves and nonlinear planes up to two dimensions. *Kernel PCA* [14] as a kernel approach can be used to visualise data and for noise reduction [15]. In [16] linear subspaces of PCA are replaced by manifolds and in [17] a neural network approach is used for nonlinear mapping. This work is focused on the auto-associative neural network approach to nonlinear PCA and its model validation problem.

For supervised methods, a standard validation technique is cross-validation. But even though the neural network architecture used is supervised, the nonlinear PCA itself is an unsupervised method that requires validating techniques different from those used for supervised methods. A common approach for validating unsupervised methods is to validate the robustness of the components under moderate modifications of the original data set, e.g., by using resampling *bootstrap* [18] or by corrupting the data with a small amount of Gaussian noise [19]. In both techniques, the motivation is that reliable components should be robust and stable against small random modification of the data. In principle, these techniques could be adapted to nonlinear methods. But there would be the difficulty of measuring the robustness of nonlinear components. Robustness of linear components is measured by comparing their directions under slightly different conditions (resampled data sets or different noise-injections). But since comparing the curvature of nonlinear components is no trivial task, nonlinear methods require other techniques for model validation.

In a similar neural network based nonlinear PCA model, termed nonlinear factor analysis (NFA) [20], a Bayesian framework is used in which the weights and inputs are described by posterior probability distributions which leads to a good regularisation. While in such Bayesian learning the inputs (components) are explicitly modelled by Gaussian distributions, the maximum likelihood approach in this work attempts to find a single
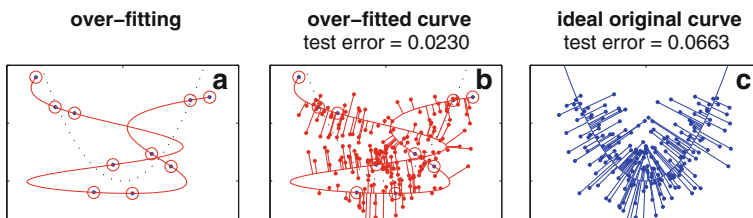

Fig. 1 The problem of mis-validation of nonlinear PCA by using a test data set. Ten training samples were generated from a quadratic function (*dotted line*) plus noise. **a** A nonlinear PCA model of too high complexity leads to an overfitted component (*solid curve*). But validating this over-fitted model with an independent test data set (**b**) gives a better (smaller) test error than using the original model from which the data were generated (**c**)

set of values for the network weights and inputs. A weight-decay regulariser is used to control the model complexity. There are several attempts to the model selection in the auto-associative nonlinear PCA. Some are based on a criterion of how good the local neighbour relation is preserved by the nonlinear PCA transformation [21]. In [22], a nearest neighbour inconsistency term that penalises complex models is added to the error function, but standard test set validation is used for model pre-selection. In [23] an alternative network architecture is proposed to solve the problems of over-fitting and non-uniqueness of nonlinear PCA solutions. Here we consider a natural approach that validates the model by its own ability to estimate missing data. Such missing data validation is used, e.g., for validating linear PCA models [24], and for comparing probabilistic nonlinear PCA models based on Gaussian processes [25]. Here, the missing data validation approach is adapted to validate the auto-associative neural network based nonlinear PCA.

2 The Test Set Validation Problem

To validate supervised methods, the standard approach is to use an independent test set for controlling the complexity of the model. This can be done either by using a new data set, or when the number of samples is limited, by performing cross-validation by repeatedly splitting the original data into a training and test set. The idea is that only the model, which best represents the underlying process, can provide optimal results on new, for the model previously unknown, data. But test set validation only works well when there exist a clear target value (e.g., class labels) as in supervised methods, it fails on unsupervised methods. In the same way that a test data set cannot be used to validate the optimal number of components in standard linear PCA, test data also cannot be used to validate the curvature of components in nonlinear PCA [7]. Even though nonlinear PCA can be performed by using a supervised neural network architecture, it is still an unsupervised method and hence should not be validated by using cross-validation. With increasing complexity, nonlinear PCA is able to provide a curved component with better data space coverage. Thus, also test data can be projected onto the (over-fitted) curve by a decreased distance and hence give an incorrect small error. This effect is illustrated in Fig. 1 using 10 training and 200 test samples generated from a quadratic function plus Gaussian noise of standard deviation $\sigma = 0.4$. The mean square error (MSE) is given by the mean of the squared distances $E = \|\hat{\mathbf{x}} - \mathbf{x}\|^2$ between the data points $\mathbf{x}$ and their projections $\hat{\mathbf{x}}$ onto the curve. The over-fitted and the well-fitted or ideal model are compared by using the same test data set. It turns out that the test error of the true original model (Fig. 1c) is almost three times larger than the test error of the overly complex model (Fig. 1b), which over-fits the data. Test set validation clearly favours the over-fitted model over the correct model, and hence fails to validate nonlinear PCA.

To understand this contradiction, we have to distinguish between an error in supervised learning and the fulfilment of specific criteria in unsupervised learning. Test set validation works well for supervised methods because we measure the error as the difference from a known target (e.g., class labels). Since in unsupervised methods the target (e.g., the correct component) is unknown, we optimize a specific criterion. In nonlinear PCA the criterion is to project the data by the shortest distance onto a curve. But a more complex over-fitted curve covers more data space and hence can also achieve a smaller error on test data than the true original curve.
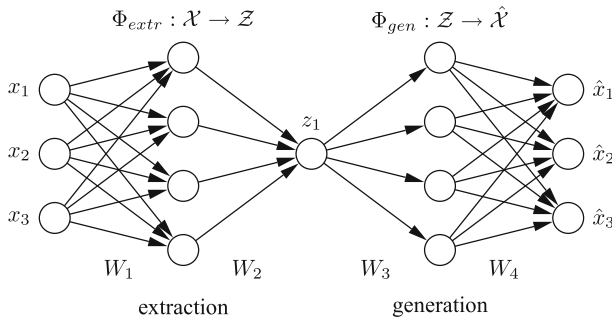
Fig. 2 The standard auto-associative neural network for nonlinear PCA. The network output $\hat{\mathbf{x}}$ is required to approximate the input $\mathbf{x}$. Illustrated is a 3-4-1-4-3 network architecture. Three-dimensional samples $\mathbf{x}$ are compressed to one component value $\mathbf{z}$ in the middle by the extraction part. The inverse generation part reconstructs $\hat{\mathbf{x}}$ from $\mathbf{z}$. The sample $\hat{\mathbf{x}}$ is a noise-reduced representation of $\mathbf{x}$, located on the component curve

3 The Nonlinear PCA Model

Nonlinear PCA (NLPCA) can be performed by using a multi-layer perceptron (MLP) of an auto-associative topology, also known as auto-encoder, replicator network, bottleneck, or sand-glass type network, see Fig. 2.

The auto-associative network performs an identity mapping. The output $\hat{\mathbf{x}}$ is forced to approximate the input $\mathbf{x}$ by minimising the squared reconstruction error $E = \parallel \hat{\mathbf{x}} - \mathbf{x} \parallel^2$. The network can be considered as consisting of two parts: the first part represents the extraction function $\Phi_{extr} : \mathcal{X} \rightarrow \mathcal{Z}$, whereas the second part represents the inverse function, the generation or reconstruction function $\Phi_{gen} : \mathcal{Z} \rightarrow \hat{\mathcal{X}}$. A hidden layer in each part enables the network to perform nonlinear mapping functions. By using additional units in the component layer in the middle, the network can be extended to extract more than one component. Ordered components can be achieved by using a hierarchical nonlinear PCA[26].

For the proposed validation approach, we have to adapt nonlinear PCA to be able to estimate missing data. This can be done by using an inverse nonlinear PCA model[27] which optimises the generation function by using only the second part of the auto-associative neural network. Since the extraction mapping $\mathcal{X} \rightarrow \mathcal{Z}$ is lost, we have to estimate both the weights $\mathbf{w}$ and also the inputs $\mathbf{z}$ which represent the values of the nonlinear component. Both $\mathbf{w}$ and $\mathbf{z}$ can optimised simultaneously to minimise the reconstruction error, as shown in[27].

The complexity of a model can be controlled by a weight-decay penalty term[28] added to the error function $E_{total} = E + \nu \left(\sum_i w_i^2 \right)$, w are the network weights. By varying the coefficient ν, the impact of the weight-decay term can be changed and hence we modify the complexity of the model which defines the flexibility of the component curves in nonlinear PCA.

4 The Missing Data Validation Approach

Since classical test set validation fails to select the optimal nonlinear PCA model, as illustrated in Fig. 1, I propose to evaluate the complexity of a model by using the error in missing data estimation as the criterion for model selection. This requires to adapt nonlinear PCA for missing data as done in the inverse nonlinear PCA model[27]. The following model

selection procedure can be used to find the optimal weight-decay complexity parameter ν of the nonlinear PCA model:

(1)   Choose a specific complexity parameter ν.
(2)   Apply inverse nonlinear PCA to a training data set.
(3)   Validate the nonlinear PCA model by its performance on missing data estimation of an independent test set in which one or more elements x_i^n of a sample $\mathbf{x}^n$ are randomly rejected. The mean of the squared errors $e_i^n = \| \hat{x}_i^n - x_i^n \|^2$ between the randomly removed values x_i^n and their estimations $\hat{x}_i^n$ by the nonlinear PCA model is used as the validation or generalization error.

Applied to a range of different weight-decay complexity parameters ν, the optimal model complexity ν is given by the lowest missing value estimation error. To get a more robust result, for each complexity setting, nonlinear PCA can be repeatedly applied by using different weight-initializations of the neural network. The median can then be used for validation as shown in the following examples.

5 Validation Examples

The first example of a nonlinear data set shows that model validation based on missing data estimation performance provides a clear optimum of the complexity parameter. The second example demonstrates that the proposed validation ensures that nonlinear PCA does not describe data in a nonlinear way when the inherent data structure is, in fact, linear.

5.1 Helix Data

The nonlinear data set consist of data $\mathbf{x} = (x_1, x_2, x_3)^T$ that lie on a one-dimensional manifold, a helical loop, embedded in three dimensions, plus Gaussian noise η of standard deviation $\sigma = 0.4$, as illustrated in Fig. 3. The samples $\mathbf{x}$ were generated from a uniformly distributed factor t over the range $[-0.8, 0.8]$, t represents the angle:

$$x_1 = \sin(\pi t) + \eta$$
$$x_2 = \cos(\pi t) + \eta$$
$$x_3 = t + \eta$$

Nonlinear PCA is applied by using a 1-10-3 network architecture optimized in 5,000 iterations by using the *conjugate gradient descent* algorithm [29].

To evaluate different weight-decay complexity parameters ν, nonlinear PCA is applied to 20 complete samples generated from the helical loop function and validated by using a missing data set of 1,000 incomplete samples in which randomly one value of the three dimensions is rejected per sample and can be easily estimated from the other two dimensions when the nonlinear component has the correct helical curve. For comparison with standard test set validation, the same 1,000 (complete) samples are used. This is repeatedly done 100 times for each model complexity with newly generated data each time. The median of missing data estimation over all 100 runs is finally taken to validate a specific model complexity.

Figure 4 shows the results of comparing the proposed model selection approach with standard test set validation. It turns out that only the missing data approach is able to show a clear minimum in the performance curve. Test set validation, by contrast, shows a small error even for very complex (over-fitted) models. This is contrary to our experience with supervised
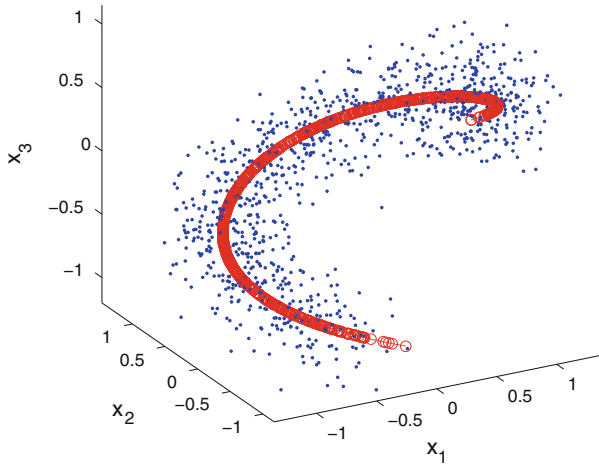
Fig. 3 Helical data set. Data are generated from a one dimensional helical loop embedded in three dimensions and additive Gaussian noise. The nonlinear component is plotted as a *red line*. *Red circles* show the projections of the data (*blue dots*) onto the curve. (Color figure online)
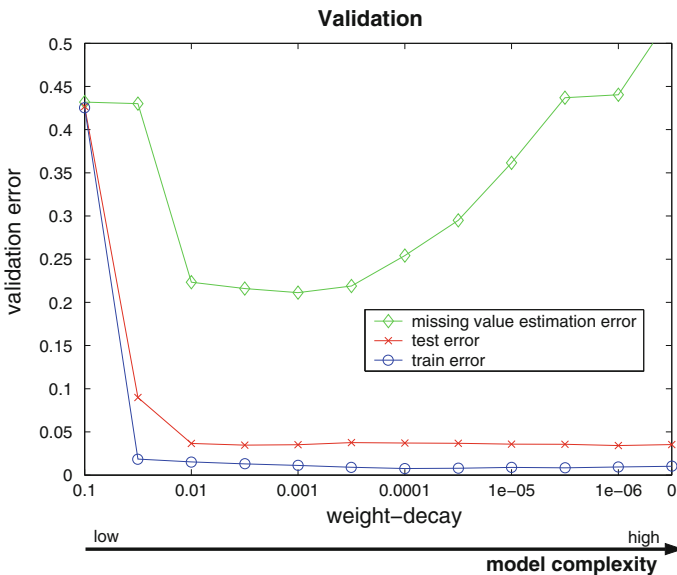
Fig. 4 Model selection. Missing data estimation is compared with standard test set validation by using the helical data set (Fig. 3). A nonlinear PCA network model of low complexity which is almost linear (*left*) results in a high error as expected for both the training and the test data. Only the missing data approach shows the expected increase in validation error for over-fitted models (*right*)

learning, where the test error becomes large again when the model over-fits. Thus, test set validation cannot be used to determine the optimal model complexity of unsupervised methods. In contrast, the missing value validation approach shows that the optimal complexity setting of the weight-decay coefficient is in the range $0.01 < \nu < 0.0001$.

5.2 Linear Data

Nonlinear PCA can also be used to answer the question of whether high-dimensional observations are driven by an unimodal or a multimodal process, e.g., in atmospheric science for analysing the El Niño-Southern Oscillation[30]. But applying nonlinear PCA can be misleading if the model complexity is insufficiently controlled: multimodality can be incorrectly detected in data that are inherently unimodal, as pointed out by Christiansen[7]. Figure 5c, d illustrates that if the model complexity is too high, even linear data is described by nonlinear components. Therefore, to obtain the right description of the data, controlling the model complexity is very important. Figure 5 shows the validation error curves of the standard test set and the proposed missing data validation for different model complexities. The median of 500 differently initialized 1-4-2 networks is plotted. Again, it is shown that standard test set validation fails in validating nonlinear PCA. With increasing model complexity, classical test set validation shows an decreasing error, and hence favours over-fitted models. By contrast, the missing value estimation error shows correctly that the optimum would be a strong penalty which gives a linear or even a point solution, thereby confirming the absence of nonlinearity in the data. This is correct because the data consists, in principle, of Gaussian noise centred at the point (0,0).

While test set validation favours over-fitted models which produce components that incorrectly show multimodal distributions, missing data validation confirms the unimodal characteristics of the data. Nonlinear PCA in combination with missing data validation can therefore be used to find out whether a high-dimensional data set is generated by a unimodal or a multimodal process.

6 Test Set Versus Missing Data Approach

In standard test set validation, the nonlinear PCA model is trained using a training set $\mathcal{X}$. An independent test set $\mathcal{Y}$ is then used to compute a validation error as $E = \parallel \hat{\mathbf{y}} - \mathbf{y} \parallel^2$, where $\hat{\mathbf{y}}$ is the output of the nonlinear PCA given the test data $\mathbf{y}$ as the input. The test set validation reconstructs the test data from the test data itself. The problem with this approach is that increasingly complex functions can give approximately $\hat{\mathbf{y}} = \mathbf{y}$, thus favouring complex models. While test set validation is a standard approach in supervised applications, in unsupervised techniques it suffers from the lack of a known target (e.g., a class label). Highly complex nonlinear PCA models, which over-fit the original training data, are in principle also able to fit test data better than would be possible by the true original model. With higher complexity, a model is able to describe a more complicated structure in the data space. Even for new test samples, it is more likely to find a short projecting distance (error) onto a curve which covers the data space almost complete than by a curve of moderate complexity (Fig. 1). The problem is that we can project the data onto *any* position on the curve. There is no further restriction in pure test set validation. In missing data estimation, by contrast, the required position on the curve is *fixed*, given by the remaining available values of the same sample. The artificially removed missing value of a test sample gives an exact target which have to be predicted from the available values of the same test sample. While test set validation predicts the test data from the test data itself, the missing data validation predicts removed values from the remaining values of the same sample. Thus, we transform the unsupervised validation problem into a kind of supervised validation problem.
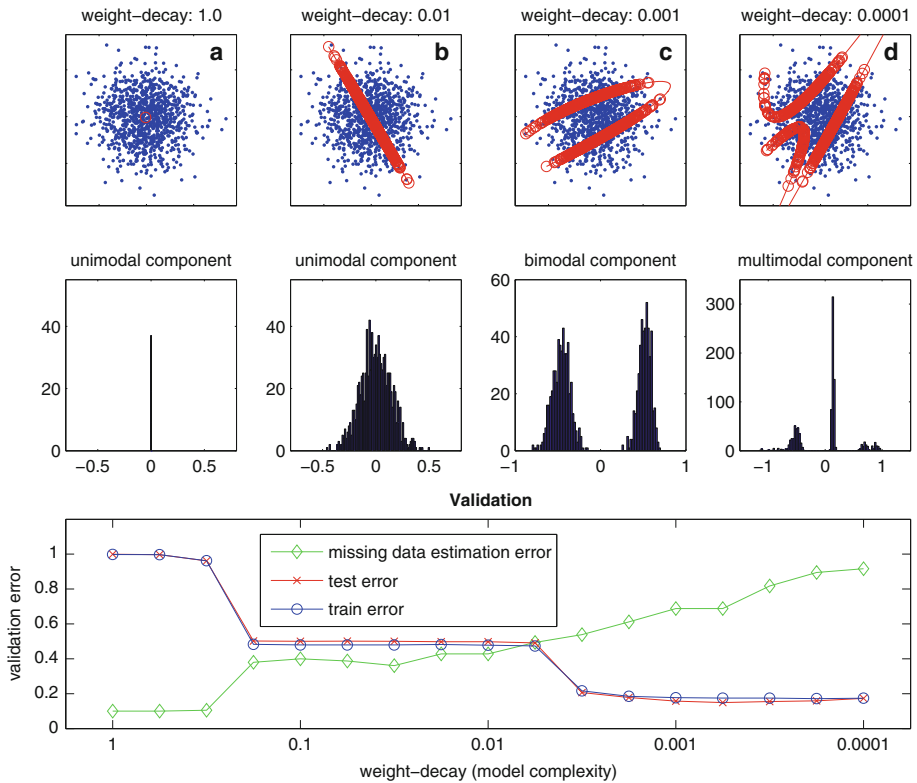
Fig. 5 Nonlinear PCA is applied to data of a two-dimensional Gaussian distribution. **c**, **d** Over-fitted models. **b** A weight-decay of 0.01 forces nonlinear PCA to describe the data by a linear component. **a** An even stronger penalty of 1.0 forces a single point solution. *Below* Only the missing data approach shows that it would be best to impose a very strong penalty that forces the network into a linear solution

7 Conclusion

In this paper, the missing data validation approach to model selection is proposed to be applied to the auto-associative neural network based nonlinear PCA. The idea behind this approach is that the true generalization error in unsupervised methods is given by a missing value estimation error and not by the classical test set error. The proposed missing value validation approach can therefore be seen as an adaptation of the standard test set validation so as to be applicable to unsupervised methods. The absence of a target value in unsupervised methods is replaced by using artificially removed missing values as expected target values that have to be predicted from the remaining values of the same sample. It can be shown that standard test set validation clearly fails to validate nonlinear PCA. In contrast, the proposed missing data validation approach was able to validate correctly the model complexity.

Availability of Software

A MATLAB® implementation of nonlinear PCA including the inverse model for estimating missing data is available at: http://www.NLPCA.org/matlab.html.

An example of how to apply the proposed validation approach can be found at: http://www.NLPCA.org/validation.html.

References

1. Kramer MA (1991) Nonlinear principal component analysis using auto-associative neural networks. AIChE J 37(2):233–243
2. DeMers D, Cottrell GW (1993) Nonlinear dimensionality reduction. In: Hanson D, Cowan J, Giles L (eds) Advances in neural information processing systems 5. Morgan Kaufmann, San Mateo, pp 580–587
3. Hecht-Nielsen R (1995) Replicator neural networks for universal optimal source coding. Science 269:1860–1863
4. Hsieh WW, Wu A, Shabbar A (2006) Nonlinear atmospheric teleconnections. Geophys Res Lett 33(7):L07714
5. Herman A (2007) Nonlinear principal component analysis of the tidal dynamics in a shallow sea. Geophys Res Lett 34:L02608
6. Scholz M, Fraunholz MJ (2008) A computational model of gene expression reveals early transcriptional events at the subtelomeric regions of the malaria parasite, *Plasmodium falciparum*. Genome Biol 9:R88. doi:10.1186/gb-2008-9-5-r88
7. Christiansen B (2005) The shortcomings of nonlinear principal component analysis in identifying circulation regimes. J Clim 18(22):4814–4823
8. Gorban AN, Kégl B, Wunsch DC, Zinovyev A (2007) Principal manifolds for data visualization and dimension reduction. LNCSE, vol 58. Springer, Berlin
9. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
10. Saul LK, Roweis ST (2004) Think globally, fit locally: unsupervised learning of low dimensional manifolds. J Mach Learn Res 4(2):119–155
11. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
12. Hastie T, Stuetzle W (1989) Principal curves. J Am Stat Assoc 84:502–516
13. Kohonen T (2001) Self-organizing maps, 3rd edn. Springer, New York
14. Schölkopf B, Smola A, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10:1299–1319
15. Mika S, Schölkopf B, Smola A, Müller KR, Scholz M, Rätsch G (1999) Kernel PCA and de-noising in feature spaces. In: Kearns M, Solla S, Cohn D (eds) Advances in neural information processing systems 11.. MIT Press, Cambridge, pp 536–542
16. Girard S, Iovleff S (2005) Auto-associative models and generalized principal component analysis. J Multivar Anal 93(1):21–39. doi:10.1016/j.jmva.2004.01.006
17. Demartines P, Herault J (1997) Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. IEEE Transac Neural Netw 8(1):148–154
18. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman and Hall, New York
19. Harmeling S, Meinecke F, Müller KR (2004) Injecting noise for analysing the stability of ICA components. Signal Process 84:255–266
20. Honkela A, Valpola H (2005) Unsupervised variational bayesian learning of nonlinear models. In: Saul YWL, Bottous L (eds) Advances in neural information processing systems 17 (NIPS'04), pp 593–600
21. Chalmond B, Girard SC (1999) Nonlinear modeling of scattered multivariate data and its application to shape change. IEEE Trans Pattern Anal Mach Intell 21(5):422–432
22. Hsieh WW (2007) Nonlinear principal component analysis of noisy data. Neural Netw 20(4): 434–443. doi:10.1016/j.neunet.2007.04.018
23. Lu BW, Pandolfo L (2011) Quasi-objective nonlinear principal component analysis. Neural Netw 24(2): 159–170. doi:10.1016/j.neunet.2010.10.001
24. Ilin A, Raiko T (2010) Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res 11:1957–2000
25. Lawrence ND (2005) Probabilistic non-linear principal component analysis with Gaussian process latent variable models. J Mach Learn Res 6:1783–1816
26. Scholz M, Vigário R (2002) Nonlinear PCA: a new hierarchical approach. In: Verleysen M (ed) In: Proceedings of ESANN, pp 439–444
27. Scholz M, Kaplan F, Guy CL, Kopka J, Selbig J (2005) Non-linear PCA: a missing data approach. Bioinformatics 21(20):3887–3895

28. Hinton GE (1987) Learning translation invariant recognition in massively parallel networks. In: Proceedings of the Conference on Parallel Architectures and Languages Europe (PARLE), pp 1–13
29. Hestenes MR, Stiefel E (1952) Methods of conjugate gradients for solving linear systems. J Res Nat Bureau Stand 49(6):409–436
30. Hsieh WW (2001) Nonlinear principal component analysis by neural networks. Tellus A 53(5):599–615