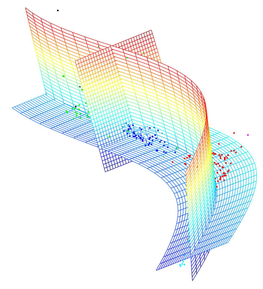
A missing data approach to validate nonlinear PCA

Matthias Scholz

Edmund Mach Foundation - Research and Innovation Center
Via Edmund Mach 1, 38010 San Michele all'Adige (TN), Italy



Many biological processes behave in a nonlinear way. Observations over time usually show a curved trajectory in the data space. To understand the dynamics of biological processes we have to identify and analyze the time trajectory. This can be done by using a nonlinear extension of principal component analysis (PCA) which provides a noise-reduced description of the curved data structure. To avoid over-fitting a careful control of the model complexity is required for which we need a strategy to validate unsupervised nonlinear methods [1].

Time trajectory

Nonlinear PCA [2] is a frequently applied technique to analyse the nonlinear data structure of experimental time courses. This includes studies of metabolite stress response [3] and gene expression analysis of reproductive cycles [4]. Nonlinear PCA recovers the trajectory from noisy data and provides a model of the time course for investigating the underlying biological process.

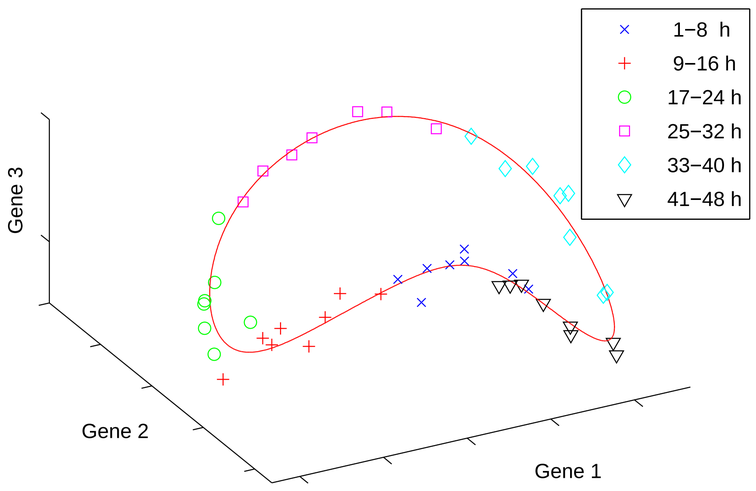


Figure 1: Nonlinear PCA is used to identify the time trajectory (red line). The nonlinear component approximates the trajectory of the data and hence gives a noise-reduced and continuous model of the biological process.

Linear PCA

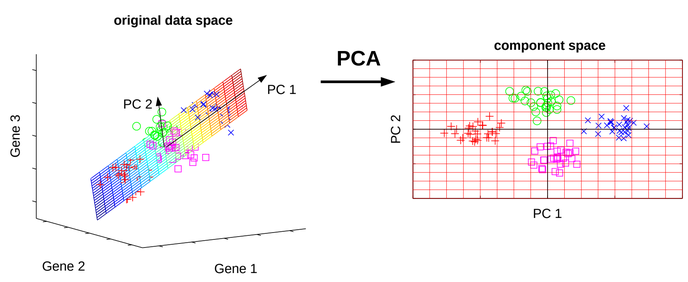


Figure 2: Standard linear PCA is restricted to describe a data structure by linear components (straight lines).

Nonlinear PCA

Nonlinear PCA generalizes the principal components of linear PCA from straight lines to curves [2].

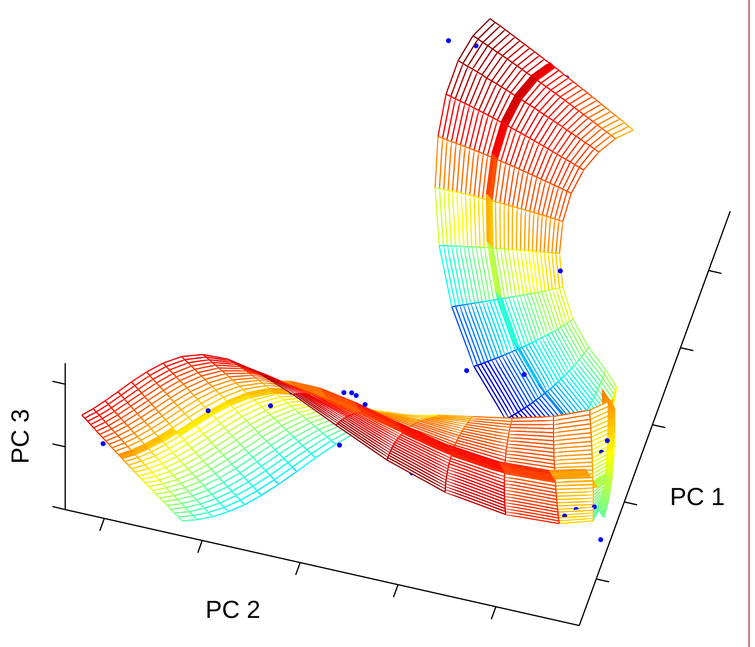


Figure 3: Nonlinear PCA can describe the inherent structure of the data by a curved subspace.

Validation problem

Test set validation cannot be used to determine the optimal model complexity of unsupervised methods including nonlinear PCA. 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).

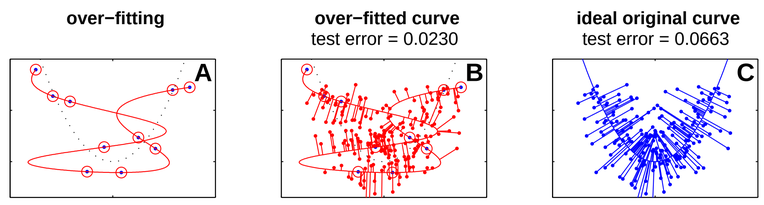


Figure 4: Standard test set validation clearly fails to validate nonlinear PCA. Validating a over-fitted model with an independent test data set (B) gives even a better (smaller) test error than using the original model from which the data were generated (C).

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. 4). 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.

Neural network for nonlinear PCA

The network output $\hat{\vec{x}}$ is required to approximate the input $\vec{x}$. Illustrated is a 3-4-1-4-3 network architecture. Three-dimensional samples $\vec{x}$ are compressed to one component value $\vec{z}$ in the middle by the extraction part. The inverse generation part reconstructs $\hat{\vec{x}}$ from $\vec{z}$. The sample $\hat{\vec{x}}$ is a noise-reduced representation of $\vec{x}$, located on the component curve.

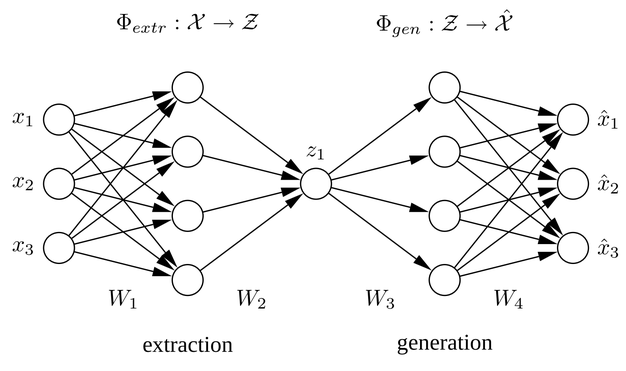


Figure 5: The standard auto-associative neural network architecture for nonlinear PCA.

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 [3] 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 $\vec{w}$ and also the inputs $\vec{z}$ which represent the values of the nonlinear component. Both $\vec{w}$ and $\vec{z}$ can optimised simultaneously to minimise the reconstruction error, as shown in [3].

A missing data approach for model validation

Motivated by the idea that only the model of optimal complexity is able to predict missing values with highest accuracy, it is used here as a natural approach for model selection. 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.

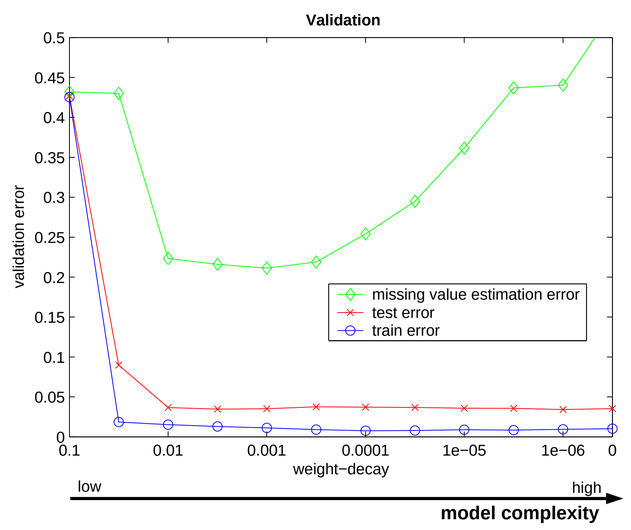


Figure 6: While standard test set validation usually favours over-fitted nonlinear PCA models, model validation based on the correctness of missing data estimation provides a clear optimum.

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).

Incorrectly identified nonlinearities

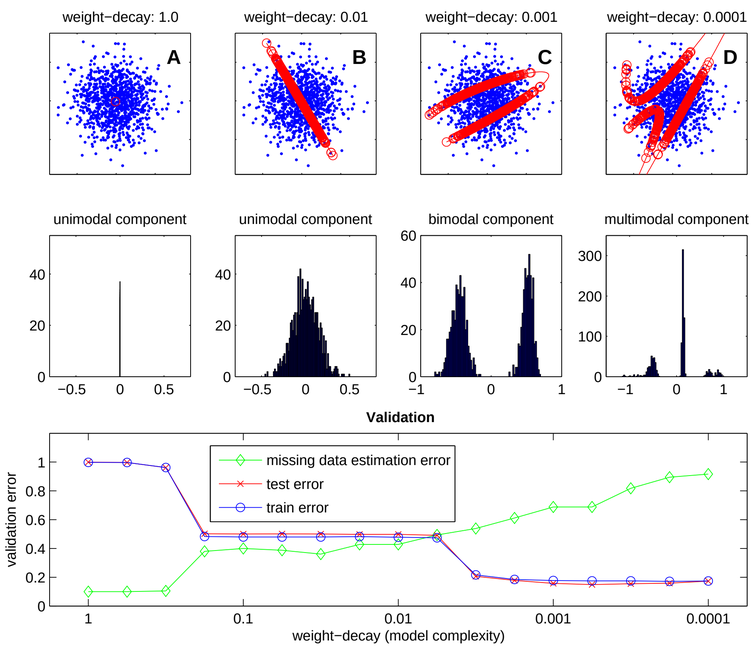


Figure 7: Model validation based on missing data estimation ensures that nonlinear PCA does not describe data in a nonlinear way when the inherent data structure is, in fact, linear. By contrast, standard test set validation favours over-fitted nonlinear PCA models.

Linear data can easily described incorrectly by nonlinear components when the model complexity is too high [5]. While classical test set validation shows an decreasing error for over-fitted models, 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.

References

[1] M. Scholz. Validation of nonlinear PCA. *Neural Processing Letters*, 2012
[2] M. A. Kramer. Nonlinear principal component analysis using auto-associative neural networks. *AIChE Journal*, 37(2):233–243, 1991
[3] M. Scholz, F. Kaplan, C.L. Guy, J. Kopka, and J. Selbig. Non-linear PCA: a missing data approach. *Bioinformatics*, 21(20):3887–3895, 2005
[4] M. Scholz and M.J. Fraunholz. A computational model of gene expression reveals early transcriptional events at the subtelomeric regions of the malaria parasite, *Plasmodium falciparum*. *Genome Biology* 9:R88, 2008
[5] B. Christiansen. The shortcomings of nonlinear principal component analysis in identifying circulation regimes. *J Clim* 18(22):48144823, 2005

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`

Contact: matthias.scholz@fmach.it