BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 20 2005, pages 3887–3895
doi:10.1093/bioinformatics/bti634

Gene expression

Non-linear PCA: a missing data approach

Matthias Scholz[1,*], Fatma Kaplan[2], Charles L. Guy[2], Joachim Kopka[1] and Joachim Selbig[1,3]

[1]Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany, [2]University of Florida, Plant Molecular and Cellular Biology Program, Department of Environmental Horticulture, Gainesville, Florida 32611, USA and [3]University of Potsdam, Bioinformatics, Germany

Received on January 5, 2005; revised on August 2, 2005; accepted on August 15, 2005
Advance Access publication August 18, 2005

ABSTRACT

Motivation: Visualizing and analysing the potential non-linear structure of a dataset is becoming an important task in molecular biology. This is even more challenging when the data have missing values.

Results: Here, we propose an inverse model that performs non-linear principal component analysis (NLPCA) from incomplete datasets. Missing values are ignored while optimizing the model, but can be estimated afterwards. Results are shown for both artificial and experimental datasets. In contrast to linear methods, non-linear methods were able to give better missing value estimations for non-linear structured data.

Application: We applied this technique to a time course of metabolite data from a cold stress experiment on the model plant *Arabidopsis thaliana*, and could approximate the mapping function from any time point to the metabolite responses. Thus, the inverse NLPCA provides greatly improved information for better understanding the complex response to cold stress.

Contact: scholz@mpimp-golm.mpg.de

1 INTRODUCTION

Non-linear principal component analysis (NLPCA) is generally seen as a non-linear generalization of standard linear principal component analysis (PCA) (Jolliffe, 1986; Diamantaras and Kung, 1996). The principal components are generalized from straight lines to curves. Here, we focus on a neural network based NLPCA, the auto-associative neural network (Kramer, 1991; DeMers and Cottrell, 1993; Hecht-Nielsen, 1995; Kirby and Miranda, 1996; Malthouse, 1998). It is successfully applied in the fields of atmospheric and oceanic sciences (Hsieh, 2004; Monahan *et al*., 2003), in astronomy and even in biomedical research. In Scholz and Vigário (2002) a hierarchically extended version of NLPCA was applied to spectral data from stars and to electromyographic (EMG) recordings for different muscle activities.

There is a wide variety of methods for visualizing data and extracting meaningful components (also termed features, factors or sources) in a non-linear way. Locally linear embedding (LLE) (Roweis and Saul, 2000; Saul and Roweis, 2004) and Isomap (Tenenbaum *et al*., 2000) were developed to visualize high dimensional data by projecting (embedding) them into a two- or low-dimensional space. A mapping function as a non-linear model is not explicitly given. Principal curves (Hastie and Stuetzle, 1989) and self-organizing maps (SOM) (Kohonen, 2001) are useful for detecting non-linear curves and two-dimensional non-linear planes. Both methods are limited to extraction of two components at most, due to high computational costs. Kernel PCA (Schölkopf *et al*., 1998), when used as pre-processing, can improve classification results.

Here, we consider the neural network approach. It provides a non-linear model of the mapping function and we will show that it can be applied to incomplete datasets by modelling only the second part of the auto-associative network, the reconstruction or generation part. The difficulty is to estimate both the model weights and the inputs which are now the required components.

For this approach Hassoun and Sudjianto (1997) optimized the weights and the inputs in two alternate steps by minimization of an error function which is equivalent to maximum likelihood. A similar approach was also used by Oh and Seung (1998). As the inputs can be represented by weights, we propose to optimize the inputs and weights simultaneously.

The same network architecture is also used by Valpola for a non-linear factor analysis (NFA) and a non-linear independent factor analysis (NIFA) (Lappalainen and Honkela, 2000; Honkela and Valpola, 2005), also applicable to incomplete datasets (Raiko and Valpola, 2001). The weights and inputs are optimized by Bayesian learning. The inputs (components) are explicitly modelled by a plain Gaussian distribution in NFA and a mixture of Gaussian distribution in NIFA. Although Bayesian inference in NFA and maximum likelihood in NLPCA often lead to similar results, their conceptual basis is rather different. Maximum likelihood attempts to find a single set of values for the network weights and inputs. In contrast, in the Bayesian approach the weights and inputs are described by posterior probability distributions which lead to a good regularisation. There are some relations: the Gaussian prior distribution for the weights corresponds to the use of a weight-decay regularizer in the maximum likelihood approach. Minimization of a mean square error function is equivalent to one maximum a posteriori (MAP) with additive Gaussian observation noise. In the proposed inverse

*To whom correspondence should be addressed.

NLPCA model a single error function is minimized. The model weights and inputs (components) are optimized simultaneously and the model is extended to be applicable to incomplete datasets. There are many methods for estimating missing values (Little and Rubin, 2002). Here, we focus on detecting non-linear components from incomplete datasets, so our approach involves ignoring missing values not a priori estimating them. However, once the non-linear mapping is effectively modelled, the missing values can then be estimated as well. This is shown for an artificial dataset and for experimental data. Estimation results were compared with results of state-of-the-art estimation techniques. There are two PCA based linear techniques: the recently published Bayesian missing value estimation method for gene expressions (Oba *et al*., 2003) which is based on Bayesian principal component analysis (BPCA) (Bishop, 1999) and probabilistic PCA (PPCA) (Verbeek *et al*., 2002) based on Roweis, (1997). Furthermore, there are the *k*-nearest neighbour based approach, KNNimpute (Troyanskaya *et al*., 2001), and a non-linear estimation by SOM.

There are many other approaches which are not further considered; for example, there are methods based on non-linear regression among variables (Zhou *et al*., 2003) or on modelling a dynamical system (Simeka and Kimmel, 2003). The latter takes the time information into account. It belongs, therefore, to supervised methods where it is much more difficult to avoid over-fitting than in the previously mentioned unsupervised methods.

Cold stress to the cell can cause rapid changes in metabolite levels. Here, we have analysed the temporal metabolite response to cold stress in the model plant *Arabidopsis thaliana*. The proposed inverse NLPCA model was applied to these, partly incomplete, metabolite data (Kaplan *et al*., 2004). Thus, we model the cold stress adaptation by a mapping function from a given time point to the metabolite responses. For each time point we are able to give the metabolites in the order of importance, i.e. the metabolites are ranked by the relative change in their concentration level. This procedure is analogous to ranking in PCA by the eigenvector values (also termed loadings or weights).

The observed experimental time information is not used in this unsupervised model. Thus, the risk of over-fitting is much lower than in a supervised regression model. Furthermore, the response time and developmental state of plant individuals in any experiment differs from the exact physical time measurement. Hence we cannot absolutely trust the physical experimental time for the description of biological experiments. An unsupervised model will be superior in accommodating the unavoidable individual variability of biological samples such as plants.

2 AUTO-ASSOCIATIVE NEURAL NETWORKS

The NLPCA, proposed by Kramer (1991), is based on a multi-layer perceptron (MLP) with an auto-associative topology, also known as an autoencoder, replicator network, bottleneck or sand glass type network. A good introduction to multi-layer perceptrons can be found in Bishop (1995), Haykin (1998).

The auto-associative network performs the identity mapping, the output $\hat{x}$ has to be equal to the input x, by minimizing the square error $\|x-\hat{x}\|^2$.

This is no trivial task, as there is a 'bottleneck' in the middle, a layer of fewer nodes than at input or output, where the data have to be projected or compressed into a lower dimensional space Z, (Fig. 1).

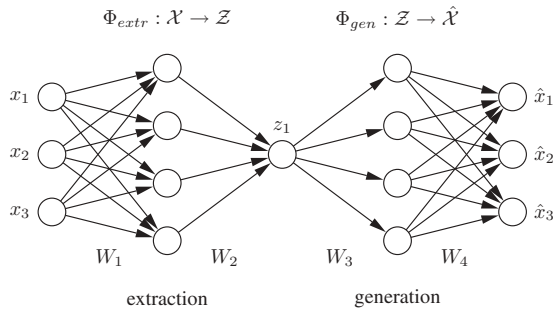

Fig. 1. The standard auto-associative neural network. The network output $\hat{x}$ is required to be equal to the input x. Illustrated is a [3-4-1-4-3] network architecture. Biases have been omitted for clarity. Three-dimensional samples x are compressed (projected) to one component z by the extraction part. The inverse generation part reconstructs $\hat{x}$ from z. The sample $\hat{x}$ is usually a noise-reduced representation of x.

The network can be divided into two parts: the first part represents the extraction function $\Phi_{\text{extr}}: \mathcal{X} \to \mathcal{Z}$, whereas the second part represents the inverse function, the generation or reconstruction function $\Phi_{\text{gen}} : \mathcal{Z} \to \hat{\mathcal{X}}$. A hidden layer in each part enables the network to perform non-linear mapping functions.

3 INVERSE NLPCA MODEL

The inverse model of NLPCA extracts the required components by only modelling the generation function $\Phi_{\text{gen}} : \mathcal{Z} \to \hat{\mathcal{X}}$of the auto-associative network. It is the inverse function to the component extraction function $\Phi_{\text{extr}}: \mathcal{X} \to \mathcal{Z}$.

The inverse model presents a set of advantages; we only have to train the second part of the auto-associative network, which is more efficient than training both parts. Also, we model the natural process, which has generated the observed samples, hence we can be sure that such a function exists, which is not necessarily the case for the extraction model. And, most importantly, the inverse NLPCA can be extended to handle incomplete datasets, as we do not need the sample data as input, the data are needed only as required output.

As the desired components are now unknown inputs, the blind inverse problem is to estimate both the inputs and the parameters of the model by only given outputs. This makes sense only with the additional constraint of a lower dimensional input.

The output $\hat{x}$ depends on the input z and the network weights $w \in W_3$, W_4, as illustrated in Figure 2,

$$\hat{x} = \Phi_{\text{gen}}(w, z) = W_4 g(W_3 z)$$

The non-linear activation function g (e.g. tanh) is applied element-wise. Biases are not explicitly considered; however, they can be included by introducing an extra unit, or input, with activation fixed at one. The mean square error depends on z and w as well:

$$E(w,z) = \frac{1}{dN} \sum_{n}^{N} \sum_{i}^{d} \left[x_i^n - \sum_{j}^{h} w_{ij} g\left(\sum_{i}^{m} w_{jk} z_k^n \right) \right]^2,$$

d is the dimensionality of the data (the number of metabolites), N is the number of samples.

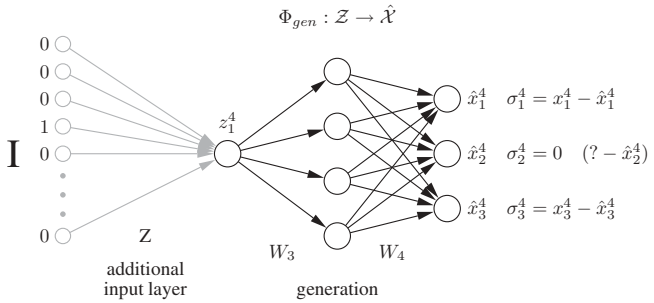

Fig. 2. The proposed inverse NLPCA model as [1-4-3] network. Only the generation part (black) of the auto-associative network (Fig. 1) is used. The inputs z can be optimized by propagating the partial errors back to the input layer z. This is equivalent to the illustrated prefixed input layer (grey), where the weights are representing the component values z. The input is now a (sample x sample) identity matrix I. For the 4th sample ($n = 4$), as illustrated, all inputs are zero except the 4th, which is one. On the right, the second element x_2^4 of the 4th sample x^4 is missing. Therefore, the partial error σ_2^4 is set to zero, identical to ignoring or non-back-propagating.

The error can be minimized by a gradient optimization algorithm. e.g. conjugate gradient descent (Hestenes and Stiefel, 1952; Press *et al*., 1992). The gradients are obtained by propagating the partial errors σ_i^n back to the input layer. For the input gradients it is simply one step further than usual. The gradients of the weights $w_{ij} \in W_4$, $w_{jk} \in W_3$ and inputs z_k^n are the partial derivatives:

$$\frac{\partial E}{\partial w_{ij}} = \sum_n \sigma_i^n g(a_j^n); \quad \sigma_i^n = \hat{x}_i n - x_i^n$$
$$\frac{\partial E}{\partial w_{jk}} = \sum_n \sigma_j^n z_k^n; \quad \sigma_j^n = g'(a_j^n) \sum_i w_{ij} \sigma_i^n$$
$$\frac{\partial E}{\partial z_k^n} = \sigma_k^n; \quad \sigma_k^n = \sum_j w_{kj} \sigma_j^n$$

For the bias, additional weights w_{i0} and w_{j0} can be used, with associated constants $z_0 = 1$ and $g(a_0) = 1$. The weights w and the inputs z can be optimized simultaneously, by considering (w, z) as one vector to optimize with given gradients. This would be equivalent to an approach where an additional input layer is representing the components z as weights, and new inputs are given by a (sample x sample) identity matrix, as illustrated in Figure 2. However, this layer is not needed for implementation. The purpose of the additional input layer is only to explain that the inverse NLPCA model can be converted to a conventionally trained multi-layer perceptron, with known inputs and simultaneously optimized weights, including the weights z, representing the desired components. Hence, an alternating approach as done by Hassoun and Sudjianto (1997) is unnecessary. Beside a more efficient optimization, it also avoids the risk of oscillating while training in an alternating approach.

A disadvantage of such an inverse approach is that there is no mapping function $\mathcal{X} \to \mathcal{Z}$, required for new data x. However, we can approximate the mapping by searching for an optimal input z to a given new sample x. For that, the network weights w have to be fixed and the input z has to be optimized to minimize the square error $\|x - \hat{x}(z)\|^2$. This is only a line search (in case of one component) or low dimensional optimization with given gradients, efficiently done by a gradient optimization algorithm.

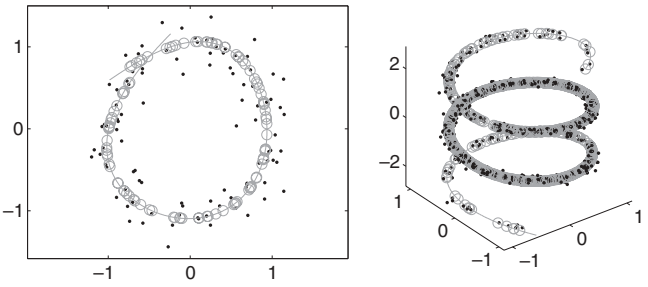

Fig. 3. Approximation of a circular (left) and a helical (right) structure by the proposed inverse NLPCA model. The noisy data x (dots) are projected onto a one-dimensional non-linear component (line). The projection or de-noised reconstruction $\hat{x}$ is marked by a circle. Note that an inverse model is able to extract self-intersecting components (left).

The inverse NLPCA is able to extract components of higher non-linear complexity than the standard NLPCA, even self-intersecting components can be modelled. This is shown in Figure 3 for a circular structure in two dimensions, generated from a uniformly distributed factor t (the angle) and a helical structure embedded in three dimensions, generated from a Gaussian distributed factor t. For the uniformly distributed 100 circular data points (plus noise), a [1-3-2] network is trained in 3000 iterations. The noisy helical structure of 1000 Gaussian distributed data points, is modelled with a [1-8-3] network in 10 000 iterations.

The inverse NLPCA is not restricted to one component. It can be extended to m components by increasing the number of units in the input layer, the component layer z, to m. With an additional hierarchical error function (Scholz and Vigário, 2002), the non-linear components $1, \ldots, m$ can be extracted in a hierarchical order, which is a natural non-linear extension to the hierarchical ordered components of the standard linear PCA.

3.1 Regularization

As we usually have a large number of dimensions (metabolites) and a relatively small number of samples, a regularization of the non-linear model is very important.

Standard methods for regularization in neural networks reduce the number of hidden units or add a weight decay term to the error function. Furthermore, auto-associative neural networks have a kind of self-regularization, caused by the fact that for each mapping function the inverse function has to be estimated as well. A complex function has usually a much more complex inverse function or the inverse function does not even exist. Therefore, the auto-associative neural network is constrained to keep the functions as simple as possible. A similar effect is observed when extracting non-linear components in a hierarchical order, where subsequent components are extracted in respect to the previous components. A complex first component would strongly increase the complexity of the second or later components. Thus, the network is constrained to generate very smooth first components.

4 MISSING VALUE ESTIMATION

The inverse NLPCA model can be easily extended to be applicable to incomplete datasets. If the ith element x_i^n of the nth sample vector x^n is missing, the partial error σ_i^n is set to zero before

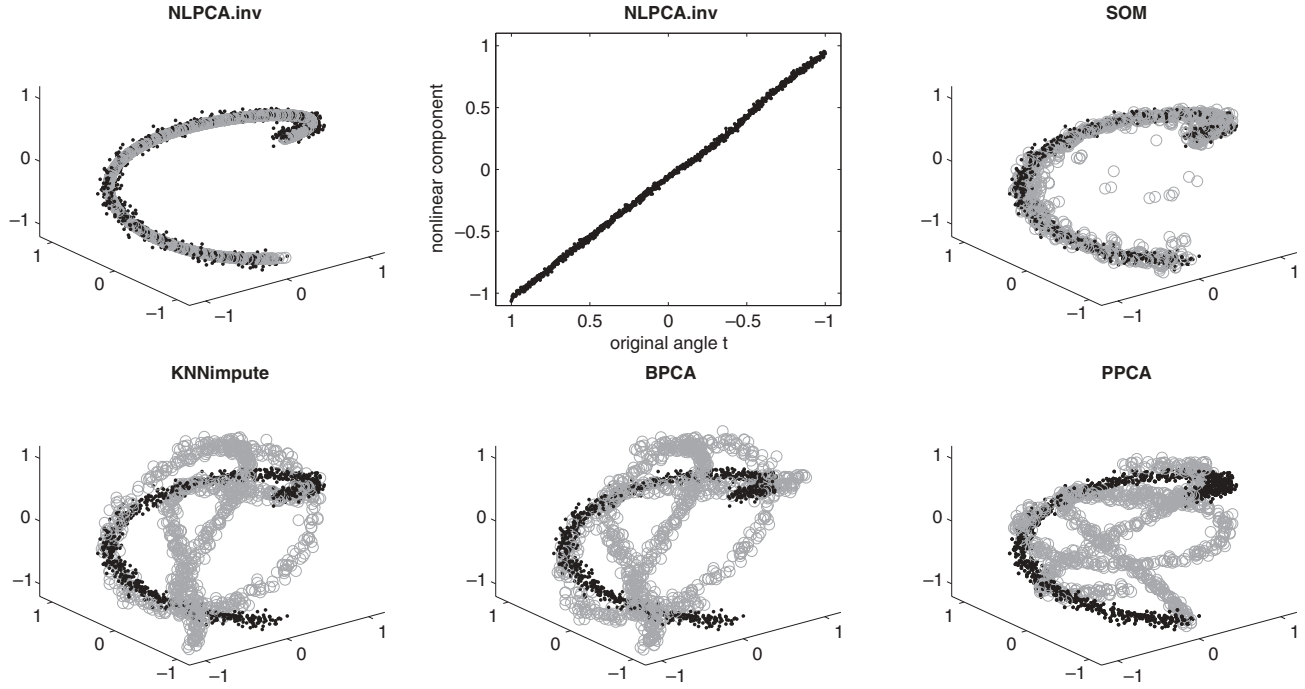

Fig. 4. Artificial data were generated to test different missing value algorithms. The samples form a helical loop. From each of the three-dimensional samples, one value is removed and then estimated by each missing value algorithm. The known complete samples are plotted as dots and the estimated values as circle. Above: the inverse NLPCA is able to extract the non-linear component from this highly incomplete dataset, and hence it can give a very good estimation of the missing values. SOM also gives a reasonably good estimation, but the linear approaches BPCA and PPCA, as well as the *k*-nearest neighbour based approach KNNimpute, fail with this non-linear dataset, see also Table 1.

back-propagating; hence this error is ignored, and it has no contribution to the gradients. Thus, the non-linear components are extracted from all the available observations. With these components the original data can be reconstructed, including the missing values. The network output x_i^n gives the estimation of the missing element $x_i{}^n$.

4.1 Missing data: artificial data

The inverse NLPCA approach was first applied to an artificial dataset and the results were compared with other missing value estimation techniques, the linear techniques BPCA[1] and PPCA[2], the *k*-nearest neighbour based approach KNNimpute[3], and the non-linear SOM[4]. The data x lie on a one-dimensional manifold (a helical loop) embedded in three dimensions, plus Gaussian noise with standard deviation 0.05, see Figure 4. 1000 samples x were generated from an uniformly distributed factor t over the range $[-1, 1]$, t represents the angle:

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

From each three-dimensional sample, one value is randomly removed and is regarded as missing. This gives a high missing value rate of 33.3 percent. However, if the non-linear component (the helix) is known, the estimation of a missing value is given exactly by the two other coordinates, except at the first and last positions of the helix loop, where in the case of missing vertical coordinate x_3, the sample can be assigned either to the first or to the last position. There are two possible optimal solutions; consequently, missing value estimation is not always unique in the non-linear case.

In Figure 4 and Table 1 it is shown that even if the datasets are incomplete for all samples, the inverse NLPCA model is able to detect the non-linear component and gives a very good missing

Table 1. MSE of missing value estimation

	Noise	Noise-free
NLPCA.inv	0.0021	0.0013
SOM	0.0405	0.0384
KNNimpute	0.4435	0.4429
BPCA	0.4191	0.4186
PPCA ($k = 3$)	0.4354	0.4347
Mean	0.4429	0.4422

Mean square error (MSE) of different missing value estimation techniques, applied to the helical data (Fig. 4). The inverse NLPCA model gives a very good estimation of the missing values. Although the model was trained with noisy data, the noise-free data were better represented than the noisy data, confirming the de-noising ability of the model.

Also SOM gives a good estimation, but the linear techniques BPCA and PPCA, as well as KNNimpute are not able to give good estimations, the results are similar to the results of naive substitution by the mean over the residuals of one variable.

[1]http://hawaii.aist-nara.ac.jp/~shige-o/tools/
[2]http://carol.science.uva.nl/~jverbeek/software/
[3]http://smi-web.stanford.edu/projects/helix/pubs/impute/
[4]http://www.cis.hut.fi/projects/somtoolbox/

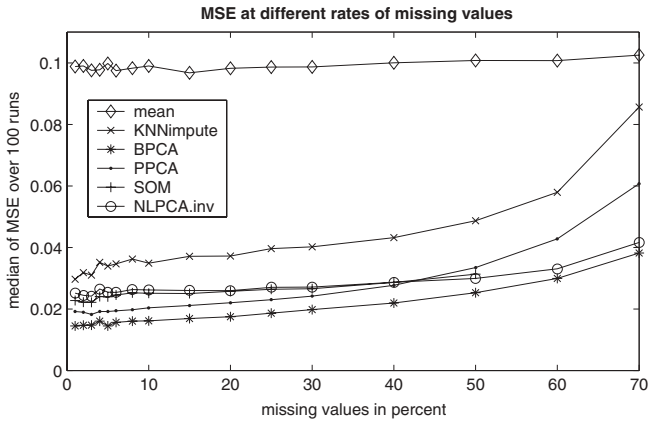

Fig. 5. From an experimental dataset of completely available 140 metabolites, different percentages of values were removed randomly and estimated by different missing value algorithms. This is done 100 times with differently removed values. The MSE over all the runs is plotted. An estimation by mean over the residual values gives the worst result. It is used as a base line. BPCA gives the best result. However, this is the case only when all 140 metabolites are considered including the large number of non-relevant metabolites with small relative variances.

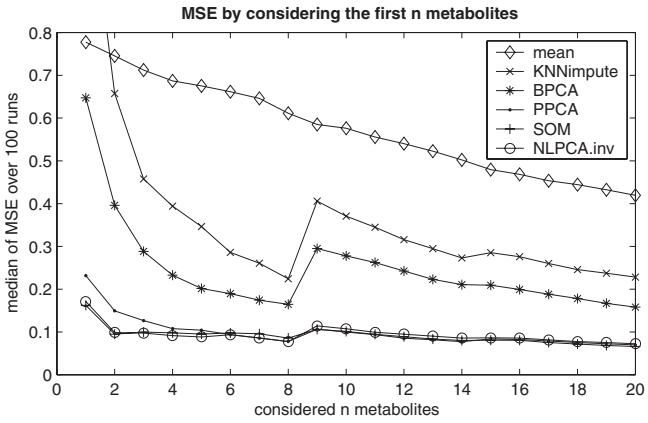

Fig. 6. In contrast to Figure 5 we have considered only the top n metabolites of highest variance, $n = 1, \ldots, 20$, at a fixed missing value rate of 10%. As the dataset contains ratios, a metabolite with a high variance is assumed to be important. The results differ from those in Figure 5. Here, BPCA gives no very good result, but still better than KNNimpute ($k = 10$ neighbours). The best result of PPCA was given with $k = 5$ components. However, at the first five metabolites, this result could still be outperformed by the non-linear techniques, the inverse NLPCA and SOM, which perform almost equally well. All techniques show an abrupt rise at the 9th metabolite (citramalic acid), caused by badly distributed data.

value estimation. The SOM also gives a reasonably good estimation, but the linear approaches BPCA and PPCA, as well as the k-nearest neighbour based approach KNNimpute, fail with this non-linear dataset.

4.2 Missing data: metabolite data

The performance of the missing value estimation techniques was also assessed using a real experimental dataset. For that we used a completely available set of 140 metabolites from our cold stress experiment, see section 5 for more details. Different percentages of values were randomly removed and regarded as missing for the estimation techniques. A good overall missing value estimation is obtained for up to 50 percent missing values. This unexpectedly high rate might be caused by the high redundancy in the data, possibly due to high connectivity or dependency among the metabolites. By comparing the different techniques, we first found that BPCA gives the best average over all 140 metabolites, (Fig. 5). But instead of a good average we are interested in a good estimation of the most important metabolites. As our data values are ratios, see section 5.1, a high variance indicates an important metabolite. Therefore, we compared the performance on the first n metabolites of highest variance which mostly also show a strong non-linear behaviour. Now the results are different, (Fig. 6). The inverse NLPCA and SOM, which perform almost equally well, give the best result at the first five most important metabolites, and perform almost as equally well as the result of PPCA with the remaining metabolites.

4.3 Missing data: gene expression data

To obtain a fair and comprehensive comparison, we also tested the performance of the missing data estimation using a larger set of gene expression data obtained from the same cold stress experiment. The data were again transformed to $\log_2$ ratios, relative to the median of control samples at time zero. In total, 16 996 genes were reduced to 1000 of highest log ratio variance. These genes are expected to be most important as they show the largest relative expression change. Twenty-one samples were measured at seven different time points.

Again, instead of a good averaged missing value estimation over all genes, we are interested in a good estimation of the most important genes, those of highest relative variance. Therefore, the cumulative mean square error (MSE) for the first 30 genes of highest ratio variance is shown (Fig. 7). The results differ from those on the metabolite dataset in Figure 6. All methods give quite similar, but significantly better, results than naive substitution by the mean of the residual values of each gene. However, BPCA which was developed for this kind of high-dimensional datasets, gave the best result for both the averaged estimation (not shown) and the estimation for the first n genes as shown in Figure 7. BPCA is successful because it uses principal components in the lower dimensional data space given by the small number of samples and not by the genes. Similar results can therefore also be obtained by the similar technique of PPCA when applied to the transposed dataset. However, the advantage of BPCA is that no parameter k, the number of used components, has to be chosen as is necessary with PPCA. The results of NLPCA were also improved when applied to the transposed matrix, and with the use of more than one non-linear component ($k = 4$). However, there might be no advantage of a non-linear technique applied to the transposed dataset as a non-linear data structure in gene data space does not necessarily lead to a non-linear structure in sample space (where genes are data points).

Consequently, for estimating missing values in large gene expression datasets BPCA is a good choice. In datasets with a smaller number of variables, as is typical for metabolite or protein datasets, other methods are more suitable. These include non-linear techniques, such as NLPCA or SOM, when the data are non-linearly distributed. Both the gene expression and metabolite datasets, are

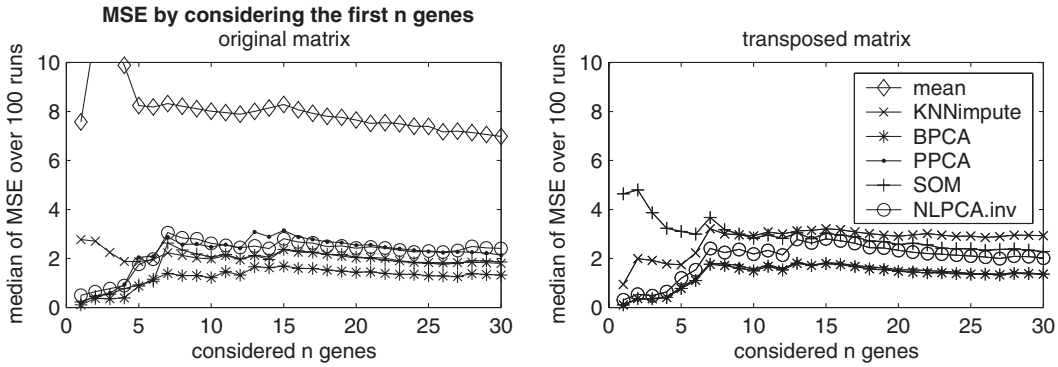

Fig. 7. Missing data algorithms applied to gene expression data of 1000 genes with 10% randomly removed values. The results differ from those on metabolite data in Figure 6. Again, we consider the most important genes of highest ratio variance. The cumulative MSE is given for the first 30 genes of highest ratio variance. All algorithms give significantly better results than the naive substitution by mean. The best result, though, is given by BPCA. Right: the results of most methods can be improved when applied to the transposed matrix. PPCA with $k = 5$ components is then almost as good as BPCA, which was applied alone without transposition because it has already an internal transposition.

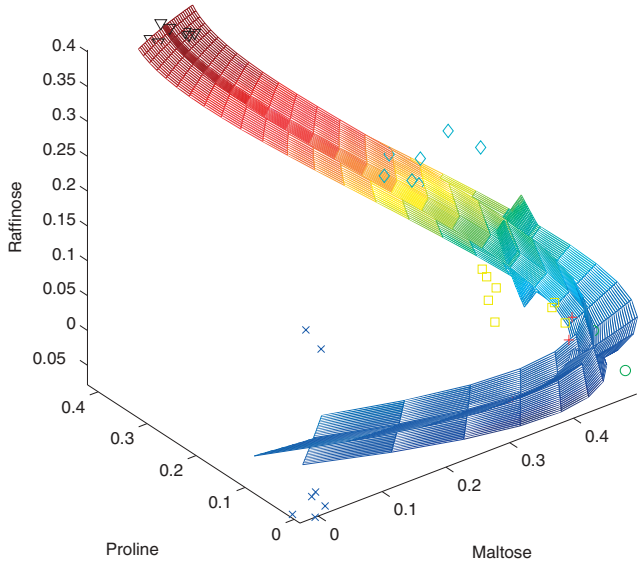

Fig. 8. The first three extracted non-linear components are plotted into the space, given by the top three metabolites of highest variance. The grid represents the new coordinate system after the non-linear transformation. The main curvature, the first non-linear component, shows the trajectory over time in the cold stress experiment. The additional second and third components represent only the noise in the data, but they are useful for regulating the first component.

available at http://nlpca.mpimp-golm.mpg.de. However, our major objective is to detect non-linear components in incomplete datasets. As these components should explain the experimental factors in the data space given by genes (where samples are data points) a transposed matrix is of no use.

5 APPLICATION

The proposed inverse NLPCA model was used to analyse the metabolite response of *A.thaliana* to cold stress at 4°C. This gives us an approximation of the mapping function from a given time point t_i to the metabolite responses x, and hence we obtain a 'noise-free' model of the biological cold stress response.

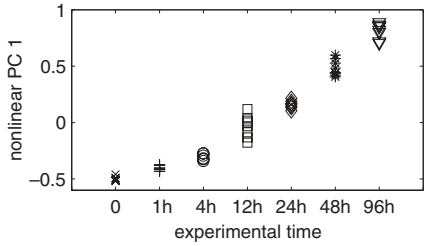

Fig. 9. The extracted first non-linear component represents the time factor. This relation is shown by plotting the first component against the observed experimental time.

5.1 Data acquisition

We have used gas chromatography/mass spectrometry (GC/MS) to measure 497 metabolites at seven different time points, at 0,1, 4,12,24,48 and 96 h, time point zero represents the control samples; Only 140 metabolites had available measurements for all samples, these metabolites were used in the previous section 4.2 to test the different methods for missing value estimation. In this experimental section the inverse NLPCA is applied to all metabolites which have <1/3 missing values. After removing 109 metabolites, the final dataset contains 388 metabolites (140 complete, 248 incomplete) and 52 samples at seven different time points (7–8 samples per time point).

The data are transformed to log fold changes (log ratios). All measurements of each metabolite $x_i = \left(x_i^1, \ldots, x_i^{52}\right)^T$ are divided by the median of the control samples at time point zero. Consequently, we are analysing ratios of metabolite concentrations with respect to a control time point. The logarithm $\log_2$ is used to get symmetric changes: $x_{\text{normed}} = \log_2\left(\frac{x}{\text{median}(x_{\text{control}})}\right)$.

5.2 Model parameters

As inverse NLPCA model, we have used a network with a [3-20-388] architecture. This means we have extracted three non-linear components; 20 non-linear hidden units were used to perform the

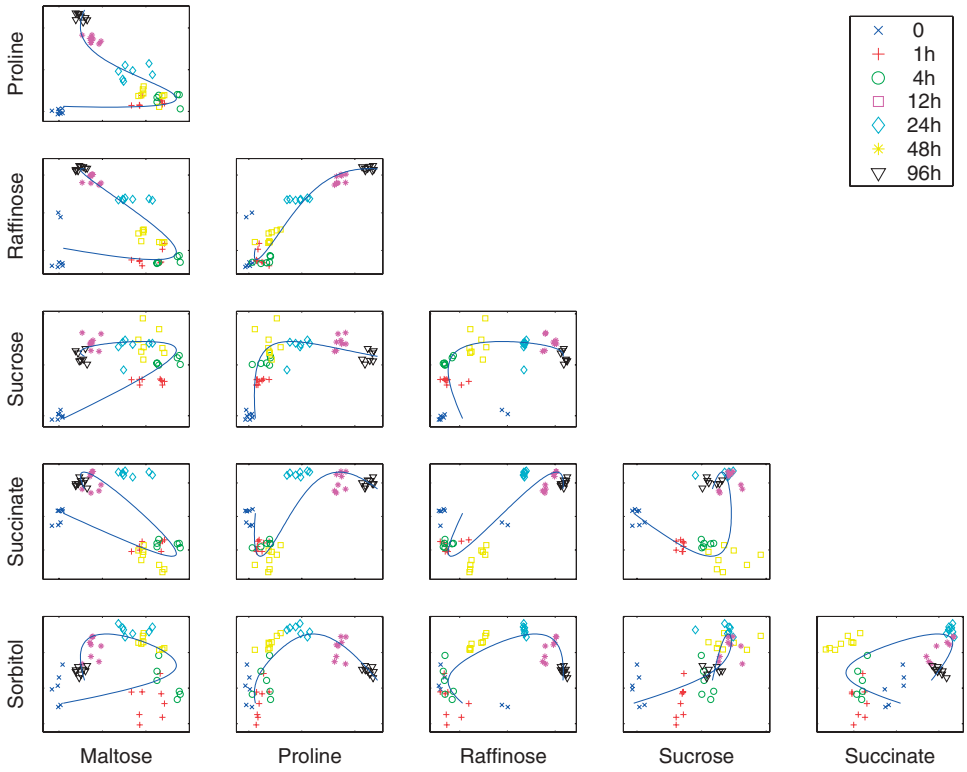

Fig. 10. Scatter plot of six selected metabolites of highest relative variance. The extracted time component (non-linear PC 1) is marked by a curve, which shows a strong non-linear behaviour.

non-linear transformation, and 388 metabolites were approximated. The training was done in 300 iterations. To limit the complexity of the model we also added a weight decay term to the error function $E_{\text{total}} = E + \nu(\sum_i w_i^2 + \sum_j z_j^2)$ with $\nu = 0.001$ and we have extracted the second and third component in a hierarchical order (Scholz and Vigário, 2002), which stabilizes the first component.

The inverse NLPCA model gives us a non-linear transformation from three estimated non-linear components to a 388 dimensional metabolite dataset. This is shown in Figure 8 for the top three metabolites of highest variance.

5.3 Results

The extracted first non-linear component is directly related to the experimental time factor, see Figure 9. This means that the global or main information, represented by variance, is the metabolite change over time. This time trajectory clearly has a non-linear behaviour, see Figure 10. The time component gives a strong curve in the original metabolite data space. It can be seen as a noise-reduced representation of the cold stress response. The inverse model gives us a mapping function $\mathcal{R}^1 \rightarrow \mathcal{R}^{388}$ from a time point t to the response x of all considered 388 metabolites $x=(x_1, \ldots, x_{388})^T$. Thus, we can analyse the approximated response curves for each metabolite, shown in Figure 11. The cold stress is reflected in almost all metabolites; however, the response behaviour is quite different. Some metabolites have a very early positive or negative response, e.g. maltose and raffinose, whereas other metabolites show only a moderate increase.

In classical PCA we can select the metabolites that are most important to a specific component by a rank order of the absolute values from the corresponding eigenvector, also termed loadings or weights. As the components are curves in non-linear PCA, no global ranking is possible. The rank order is different for different positions on the curved component, hence different at different time points in our case. However, we can give a rank order for each individual time point by computing the gradient $q_i = \frac{dx_i}{dt}$ on the non-linear time curve at this time point. The rank order of the top 20 metabolites is shown in Table 2 for an early time point t_1 and a late time point t_2. The influence values $\hat{q}_i$ are the l_2-normalized gradients q_i, $\sum_i (\hat{q}_i)^2 = 1$. The gradient curves over time are shown in Figure 11. We found that even at the last time point of the experiment, 96 hours, there are still some metabolites with significant changes in their concentrations.

6 CONCLUSIONS

NLPCA was achieved by an inverse neural network model that was applicable to incomplete datasets. With this inverse NLPCA we were able to extract non-linear (curved) components from datasets with a large number of missing values. These extracted components can be used, together with the model, to reconstruct the original data, including the missing values. We have shown that in the case of non-linearly structured datasets, both non-linear techniques, the inverse NLPCA and SOM, can improve the missing value estimation performance on the most important metabolites. We have shown that in the case of non-linearly structured datasets, both non-linear techniques, the inverse NLPCA and SOM, can improve the missing value estimation performance for the most important metabolites of the lower dimensional metabolite dataset. In the

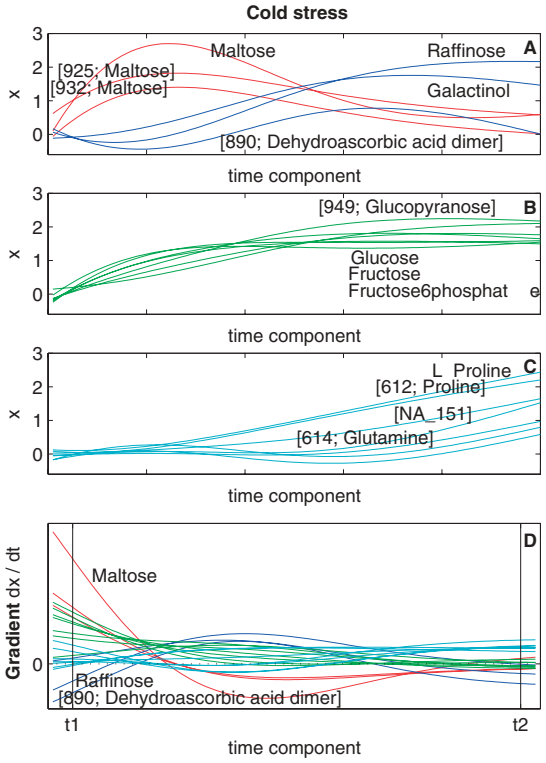

Fig. 11. The top three figures show the different shapes of the approximated metabolite response curves over time. (**A**) Early positive or negative transients, (**B**) increasing metabolite concentrations up to a saturation level, or (**C**) a delayed increase, and still increasing at the last time point. (**D**) The gradients give us the influence of all metabolites at any time point, analogous to loading factors in PCA. A high positive or high negative gradient would be interesting. There is a strong early dynamic, which is quickly moderated, except for some metabolites that are still not stable at the end. Plotted are the top 20 metabolites with the highest absolute gradients. The rank order for the marked early time t_1 and late time t_2 is given in Table 2.

larger gene expression dataset the best missing data estimations were obtained by BPCA and PPCA.

Applied to our cold stress experiment, the first non-linear component was directly related to the experimental time factor. Thus, the inverse NLPCA model gives us the continuous metabolite response over the time frame of the experiment. This trajectory over time is helpful to get a better understanding of the cold stress response. For each time point, including interpolated time points, we are able to give a ranked list of the most important metabolites, analogous to a global ranking in PCA.

The cold stress response clearly showed a non-linear behaviour over time, at the metabolite level (Kaplan *et al*., 2004). A similar non-linear behaviour was also found in gene expression data from the same cold stress experiment (data not shown). This non-linear analysis can therefore be done in the same way for such data.

Non-linearities are not restricted to temporal experiments, they can also be caused by other continuously changing factors, e.g. different temperatures at a fixed time point. Even natural phenotypes often take the form of a continuous range (Fridman *et al*., 2004), where non-linearities could exist.

Table 2. Top 20 metabolites at time points t_1 and t_2

t_1, approx. 0.5 h		t_2, approx. 96 h	
$\hat{q}$	metabolite	$\hat{q}$	metabolite
0.43	Maltose methoxyamine	0.24	[614; Glutamine]
0.23	[932; Maltose]	−0.20	[890; Dehydroascorbic acid dimer]
0.21	Fructose methoxyamine	0.18	[NA_293]
0.19	[925; Maltose]	0.18	[NA_201]
0.19	Fructose-6-phosphate	0.17	[NA_351]
0.17	Glucose methoxyamine	0.16	[NA_151]
0.17	Glucose-6-phosphate	0.16	L-Arginine
0.16	[674; Glutamine]	0.16	L-Proline
0.16	[NA_1]	−0.14	Sorbitol
0.15	[NA_154]	−0.13	4-Aminobutyric acid
0.14	[NA_341]	0.13	[612; Proline]
0.14	[NA_19]	0.12	[NA_42]
0.14	L-Arginine	−0.11	[NA_118]
0.13	Glycine	−0.11	[NA_37]
0.13	[NA_160]	−0.11	[NA_70]
0.12	[949; Glucopyranose]	0.11	[529; Indole-3-acetic acid]
0.12	[NA_84]	0.10	[NA_210]
−0.12	[890 Dehydroascorbic acid dimer]	0.10	[NA_68]
0.12	[880; Maltose methoxyamine]	−0.10	Galactinol
0.12	L-Glycerol-3-phosphate	−0.10	[NA_117]

The most important metabolites are given for an early time point t_1 of around 0.5 h (interpolated) cold stress and a very late time point t_2 of around 96 h.
The metabolites are ranked by their influences at a specific time point, given by the gradient of the non-linear time component at this time point. As expected maltose, fructose and glucose give a strong early response to cold stress; however, even after 96 hours there are still some metabolites with significant changes in their activity. Brackets '[. . .]' denote an unknown metabolite, e.g. [925; Maltose] denotes a metabolite with high mass spectral similarity to maltose.

ACKNOWLEDGEMENT

We would like to thank John Lunn for helpful comments on the manuscript. Funding to pay the Open Access publication charges for this article was provided by the Max Planck Society.

Conflict of interest: none declared.

REFERENCES

Bishop,C. (1995) Neural Networks for Pattern Recognition. Oxford University Press, .

Bishop,C. (1999) Variational principal components. *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99* pp. 509–514.

DeMers,D. and Cottrell,G. W. (1993) Nonlinear dimensionality reduction. In Hanson,D., Cowan,J. and Giles,L. (eds), *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, San Mateo, CA, pp. 580–587.

Diamantaras,K. and Kung,S. (1996) Principal Component Neural Networks. Wiley, NY.

Fridman,E. *et al*. (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science*, **305**, 1786–1789.

Hassoun,M. H. and Sudjianto,A. (1997) Compression net-free autoencoders. *Workshop on Advances in Autoencoder/Autoassociator-Based Computations at the NIPS 97 Conference*.

Hastie,T. and Stuetzle,W. (1989) Principal curves. *J. American Statistical Association*, **84**, 502–516.

Haykin,S. (1998) Neural Networks-A Comprehensive Foundation, 2nd edn. Prentice Hall.

Hecht-Nielsen,R. (1995) Replicator neural networks for universal optimal source coding. *Science*, **269**, 1860–1863.

Hestenes,M.R. and Stiefel,E. (1952) Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, **49**, 409–436.

Honkela,A. and Valpola,H. (2005) Unsupervised variational bayesian learning of nonlinear models. To appear in *Advances in Neural Information Processing Systems 17*.

Hsieh,W.W. (2004) Nonlinear multivariate and time series analysis by neural network methods. *Reviews of Geophysics*, **42**, RG1003, doi:10.1029/2002RG000112.

Jolliffe,I.T. (1986) Principal Component Analysis. Springer-Verlag, NY.

Kaplan,F. *et al*. (2004) Exploring the temperature-stress metabolome of *arabidopsis*. *Plant Physiology*, **136**, 4159–4168.

Kirby,M. J. and Miranda,R. (1996) Circular nodes in neural networks. *Neural Computation*, **8**, 390–402.

Kohonen,T. (2001) Self-Organizing Maps, 3rd edn, Springer.

Kramer,M. A. (1991) Nonlinear principal component analysis using auto-associative neural networks. *AIChE Journal*, **37**, 233–243.

Lappalainen,H. and Honkela,A. (2000) Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Girolami,M. (ed.), *Advances in Independent Component Analysis*. Springer-Verlag, pp. 93–121.

Little,R.J.A. and Rubin,D.B. (2002) Statistical Analysis with Missing Data. John Wiley & Sons, NY, second edition.

Malthouse,E. C. (1998) Limitations of nonlinear pca as performed with generic neural networks. *IEEE Transactions on Neural Networks*, **9**, 165–173.

Monahan,A.H. *et al*. (2003) The vertical structure of wintertime climate regimes of the northern hemisphere extratropical atmosphere. *J. Climate*, **16**, 2005–2021.

Oba,S. *et al*. (2003) A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Oh,J.-H. and Seung,H. (1998) Learning generative models with the up-propagation algorithm. In Jordan,M.I., Kearns,M.J. and Solla,S.A. (eds), *Advances in Neural Information Processing Systems, volume 10*. The MIT Press, pp. 605–611.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B. P. (1992) Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, second edition.

Raiko,T. and Valpola,H. (2001) Missing values in nonlinear factor analysis. In *Proc. of the 8th Int. Conf. on Neural Information Processing (ICONIP'01)*. Shanghai, pp. 822–827.

Roweis,S. (1997) Algorithms for PCA and SPCA. In *Neural Information Processing Systems 10 (NIPS'97)*. pp. 626–632.

Roweis,S.T. and Saul,L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326.

Saul,L.K. and Roweis,S.T. (2004) Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res*., **4**, 119–155.

Schölkopf,B. *et al*. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319.

Scholz,M. and Vigário,R. (2002) Nonlinear PCA: a new hierarchical approach. In Verleysen,M. (ed.), *Proceedings ESANN*. pp. 439–444.

Simeka,K. and Kimmel,M. (2003) A note on estimation of dynamics of multiple gene expression based on singular value decomposition. *Math. Biosci*., **182**, 183–199.

Tenenbaum,J.B. *et al*. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Verbeek,J.J., Vlassis,N. and Kröse,B. (2002) Procrustes analysis to coordinate mixtures of probabilistic principal component analyzers. Technical report, Computer Science Institute, University of Amsterdam, The Netherlands.

Zhou,X. *et al*. (2003) Missing-value estimation using linear and non-linear regression with bayesian gene selection. *Bioinformatics*, **19**, 2302–2307.