Missing Values in Hierarchical Nonlinear Factor Analysis

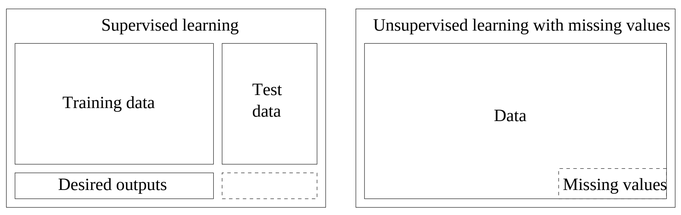
Tapani Raiko, Harri Valpola, Tomas Östman and Juha Karhunen
Helsinki University of Technology, Neural Networks Research Centre

ABSTRACT

The properties of hierarchical nonlinear factor analysis (HNFA) recently introduced by Valpola and others [3] are studied by reconstructing values. The variational Bayesian learning algorithm for HNFA has linear computational complexity and is able to infer the structure of the model in addition to estimating the parameters. To compare HNFA with other methods, we continued the experiments with speech spectrograms in [1] comparing nonlinear factor analysis (NFA) with linear factor analysis (FA) and with the self-organising map. Experiments suggest that HNFA lies between FA and NFA in handling nonlinear problems. Furthermore, HNFA gives better reconstructions than FA and it is more reliable than NFA.

Introduction

- Assume data $\mathbf{X}$ to be a set of real valued vectors
- Missing values are components of the data that are not observed
  - Example: supervised learning can be seen as reconstructing missing values



- A generative model for the data can handle and reconstruct missing values easily
  - A typical model: Factors $\mathbf{s}(t)$ have generated the observations $\mathbf{x}(t)$ through a (possibly) nonlinear mapping $\mathbf{f}(\cdot)$ and noise $\mathbf{n}(t)$:

$$\mathbf{x}(t) = \mathbf{f}[\mathbf{s}(t)] + \mathbf{n}(t) \quad (1)$$

- We wish to demonstrate the properties of hierarchical nonlinear factor analysis (HNFA)
  - High-dimensional problems need to be studied indirectly
  - Synthetic missing value patterns allow controlled comparison
  - Real-world data makes the experiments realistic

Variational Bayesian learning for nonlinear models

- Approximate the true posterior density $p(\boldsymbol{\theta} \mid \mathbf{X})$ of the unknown variables $\boldsymbol{\theta}$ by $q(\boldsymbol{\theta})$
- The unknown variables $\boldsymbol{\theta}$ include the factors $s(t)$, the parameters determining the mapping $\mathbf{f}(\cdot)$ and other parameters as the amount of noise
- The functional form of $q(\boldsymbol{\theta})$ is restricted
  - In our case $q(\boldsymbol{\theta})$ is a diagonal Gaussian distribution
- The misfit between $p(\boldsymbol{\theta} \mid \mathbf{X})$ and $q(\boldsymbol{\theta})$ is measured by a Kullback-Leibler divergence based cost function

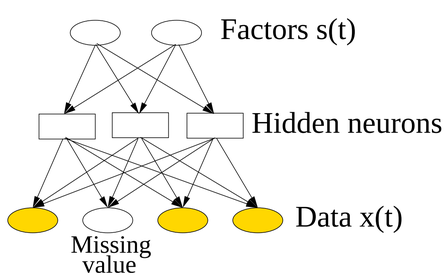
$$\mathcal{C} = D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathbf{X})) - \log p(\mathbf{X}) = \left\langle \log \frac{q(\boldsymbol{\theta})}{p(\mathbf{X}, \boldsymbol{\theta})} \right\rangle \quad (2)$$

- The cost function relates to the model evidence $p(\mathbf{X} \mid \text{model})$
  - Allows comparison of model structures
- Missing values are handled as part of $\boldsymbol{\theta}$ instead of $\mathbf{X}$
  - The posterior approximation $q(\boldsymbol{\theta})$ can be used as a reconstruction for the missing values
  - No substantial increase in computational complexity

Nonlinear factor analysis (NFA)

- Nonlinear factor analysis (NFA) has a multi-layer perceptron (MLP) network that connects the factors to the data

$$\mathbf{f}(\mathbf{s}(t); \mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}) = \mathbf{A} \tanh[\mathbf{B}\mathbf{s}(t) + \mathbf{b}] + \mathbf{a} \quad (3)$$



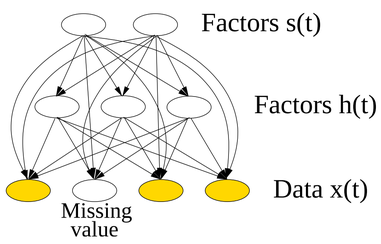
- Learning is unsupervised and thus differs in many ways from standard backpropagation
- The posterior mean and variance of $\mathbf{f}(\cdot)$ over $q(\boldsymbol{\theta})$ need to be approximated
  - This causes unreliability

Hierarchical Nonlinear Factor Analysis (HNFA)

- The key idea in HNFA is to introduce latent variables $\mathbf{h}(t)$ before the nonlinearities and thus split the mapping (3) into two parts:

$$\mathbf{h}(t) = \mathbf{B}\mathbf{s}(t) + \mathbf{b} + \mathbf{n}_h(t) \quad (4)$$
$$\mathbf{x}(t) = \mathbf{A}\phi[\mathbf{h}(t)] + \mathbf{C}\mathbf{s}(t) + \mathbf{a} + \mathbf{n}_x(t) \quad (5)$$

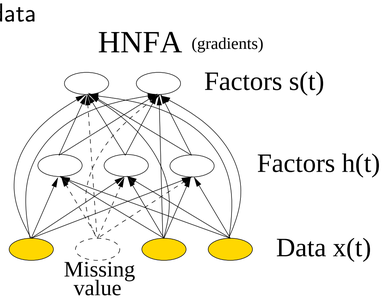
  - $\mathbf{n}_h(t)$ and $\mathbf{n}_x(t)$ are Gaussian noise terms and
  - The nonlinearity $\phi(\xi) = \exp(-\xi^2)$ operates on each element



- The nodes in the middle layer are assumed independent in $q(\boldsymbol{\theta})$
  - The posterior means and variances have analytic expressions
  - The solution is pushed into a direction with fewer simultaneously active middle layer nodes
    - Leads to conservative estimates of the nonlinearity of the model
- Each step in learning tries to minimise the cost function (2)
- Computational complexity is linear to the size of the model in HNFA and quadratic in NFA

Handling Missing Values

- During learning, the factors are updated based on gradients from below
- Missing values do not contribute to the gradients
- The factors (and other parameters) are thus estimated based only on observed data



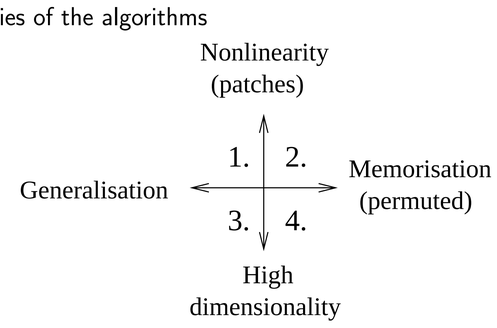
- In the feedforward phase, the missing values behave like any other ones (see Figure in the previous box)

Other Comparison Methods

- Factor analysis (FA)
  - FA is similar to principal component analysis (PCA) but it has an explicit noise model
  - The mapping $\mathbf{f}(\cdot)$ is linear
  - Large dimensionality is not a problem
  - Equivalent to HNFA without hidden nodes $\mathbf{h}(t)$
- The self-organising map (SOM)
  - SOM can be presented in terms of (1): The factor vector $\mathbf{s}(t)$ contains discrete map coordinates which select the active map unit
  - SOM captures nonlinearities and clusters, but has difficulties with data of high intrinsic dimensionality and with generalisation
  - Reconstructions are done here by associating a Gaussian kernel to each map unit

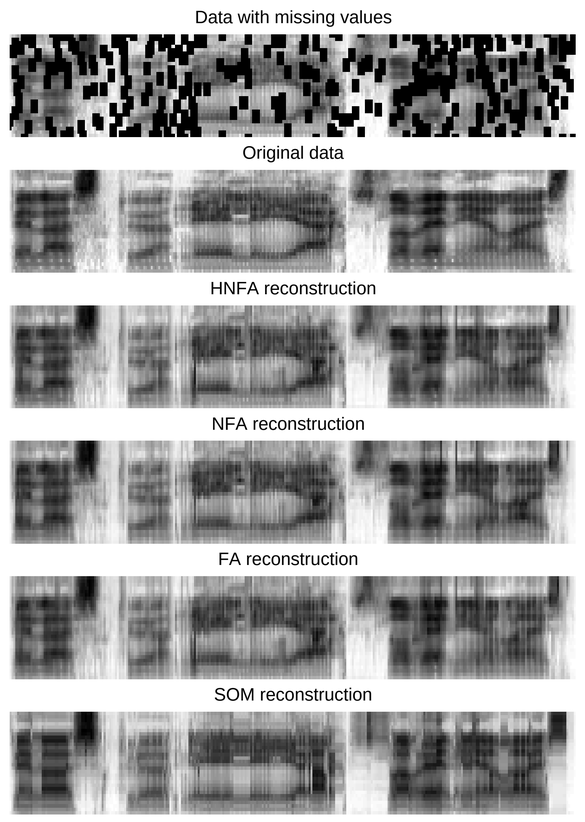
Experiments

- The goal is to study nonlinear models by measuring the quality of reconstructions of missing values
- The data set consists of 30-dimensional speech spectrograms
- Temporal information is left out to ease the comparison of the models
- Missing values are set in four different ways to measure different properties of the algorithms



Experiments, continued

- In settings 1 and 2, the values are set to miss randomly in 4×4 patches and in settings 3 and 4 independently of any neighbours
- In settings 2 and 4, the samples are randomly permuted
- Reconstructions from the setting 1:



- The mean and the std of the mean square reconstruction error are:

	FA	HNFA	NFA	SOM
1.	1.87	1.80 ± 0.03	1.74 ± 0.02	1.69 ± 0.02
2.	1.85	1.78 ± 0.03	1.71 ± 0.01	1.55 ± 0.01
3.	0.57	$0.55 \pm .005$	$0.56 \pm .002$	0.86 ± 0.01
4.	0.58	$0.55 \pm .008$	$0.58 \pm .004$	0.87 ± 0.01

- In Setting 1, the SOM with highest nonlinearity gives the best reconstructions and NFA, HNFA and finally FA follow
- In Setting 2, the permutation makes the test set contain vectors very similar to ones in the training set.
  - Generalisation becomes less important
  - SOM is able to memorise details better due to its high number of parameters
- The Settings 3 and 4 were quite similar to each other
  - The sparse missing value patterns makes the problem easier and accuracy in high dimensions more important
  - Nonlinear effects were not important since HNFA and NFA were only marginally better than FA
  - SOM was clearly poorer because it has only two intrinsic dimensions

Conclusions

- FA is better than the SOM when expressivity in high dimensions is important
- SOM is better than FA when nonlinear effects are more important
- The extensions of FA, NFA and HNFA, expectedly performed better than FA in each setting
- HNFA can model part of the nonlinearity without increasing the computational complexity dramatically
- HNFA is recommended over NFA because of its reliability
- New learning schemes may enhance NFA and HNFA

Python/C++ code

Python/C++ code for Bayes Blocks library used in the experiments is available at

`http://www.cis.hut.fi/projects/bayes/`

References

[1] Tapani Raiko and Harri Valpola. Missing values in nonlinear factor analysis. In *Proc. of the 8th Int. Conf. on Neural Information Processing (ICONIP'01)*, pages 822–827, Shanghai, 2001.

[2] Tapani Raiko, Harri Valpola, Tomas Östman, and Juha Karhunen. Missing values in hierarchical nonlinear factor analysis. In *Proc. of the joint 13th ICANN and 10th ICONIP*, Istanbul, Turkey, June 2003. To appear.

[3] Harri Valpola, Tomas Östman, and Juha Karhunen. Nonlinear independent factor analysis by hierarchical models. In *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, 2003.