Helsinki University of Technology
Dissertations in Computer and Information Science
Espoo 2006 Report D16

ADVANCED SOURCE SEPARATION METHODS WITH APPLICATIONS TO SPATIO-TEMPORAL DATASETS

Alexander Ilin

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Computer Science and Engineering for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 3rd of November, 2006, at 12 o'clock noon.

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science

Distribution:
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400
FI-02015 TKK
FINLAND
Tel. +358-9-451 3272
Fax +358-9-451 3277
http://www.cis.hut.fi

Available in PDF format at http://lib.tkk.fi/Diss/2006/isbn9512284251/

Printed version:
ISBN-13 978-951-22-8424-5
ISBN-10 951-22-8424-3

Electronic version:
ISBN-13 978-951-22-8425-2
ISBN-10 951-22-8425-1

ISSN 1459-7020

Otamedia Oy
Espoo 2006

Ilin, A. (2006): **Advanced source separation methods with applications to spatio-temporal datasets**. Doctoral thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D16, Espoo, Finland.

Keywords: Bayesian learning, blind source separation, global climate, denoising source separation, frequency-based separation, independent component analysis, independent subspace analysis, latent variable models, nonstationarity of variance, post-nonlinear mixing, unsupervised learning, variational methods.

Abstract

Latent variable models are useful tools for statistical data analysis in many applications. Examples of popular models include factor analysis, state-space models and independent component analysis. These types of models can be used for solving the source separation problem in which the latent variables should have a meaningful interpretation and represent the actual sources generating data. Source separation methods is the main focus of this work.

Bayesian statistical theory provides a principled way to learn latent variable models and therefore to solve the source separation problem. The first part of this work studies variational Bayesian methods and their application to different latent variable models. The properties of variational Bayesian methods are investigated both theoretically and experimentally using linear source separation models. A new nonlinear factor analysis model which restricts the generative mapping to the practically important case of post-nonlinear mixtures is presented. The variational Bayesian approach to learning nonlinear state-space models is studied as well. This method is applied to the practical problem of detecting changes in the dynamics of complex nonlinear processes.

The main drawback of Bayesian methods is their high computational burden. This complicates their use for exploratory data analysis in which observed data regularities often suggest what kind of models could be tried. Therefore, the second part of this work proposes several faster source separation algorithms implemented in a common algorithmic framework. The proposed approaches separate the sources by analyzing their spectral contents, decoupling their dynamic models or by optimizing their prominent variance structures. These algorithms are applied to spatio-temporal datasets containing global climate measurements from a long period of time.

Preface

This work has been carried out in the Adaptive Informatics Research Centre (former Neural Networks Research Centre) hosted by the Laboratory of Computer and Information Science (CIS) at Helsinki University of Technology (HUT). Part of the work was done during my short visit to the Institut National Polytechnique de Grenoble (INPG).

I have been working under the supervision of Prof. Erkki Oja, the head of the CIS laboratory. I would like to thank him for the opportunity to do my doctoral studies in such a strong research group, for his guidance, support and encouragement. I also want to acknowledge gratefully the outstanding research facilities provided in the laboratory.

I wish to express my gratitude to Dr. Harri Valpola, who has been the instructor of this thesis, for encouraging me in this work and sharing his own practical experience. He has motivated me to conduct interesting research and his ideas have strongly influenced all the work done in this project.

I would like to thank Prof. Juha Karhunen and members of the Bayes research group, Dr. Antti Honkela, Tapani Raiko and Markus Harva, for joint work and interesting discussions. I am very grateful to Prof. Christian Jutten for hosting me at INPG and I would like to thank him and Dr. Sophie Achard for the fruitful discussions and friendly atmosphere there. I am grateful to Prof. Padhraic Smyth for pointing out climate datasets, which led to a very exciting application for the developed methods. I also wish to thank the pre-examiners of the thesis, Dr. Mark Girolami and Dr. Aki Vehtari, for their valuable feedback.

The main source of funding for the work has been HUT, the additional funding has been from the Centre for International Mobility (CIMO) and the Helsinki Graduate School in Computer Science and Engineering (HeCSE). The visit to INPG was funded by the IST Programme of the European Community, under the project BLISS, IST-1999-14190. I am also grateful to the personal grant received from the Jenny and Antti Wihuri foundation.

I would like to thank my friends in the laboratory, Karthikesh, Jan-Hendrik, Ramūnas, for sharing thoughts and social life. Many thanks to our secretary Leila Koivisto for her help in many important practical issues.

Finally, I warmly thank Tatiana and my family for their support.

Espoo, November 2006

Alexander Ilin

Contents

Publications of the thesis

1 A. Ilin and H. Valpola. On the effect of the form of the posterior approximation in variational learning of ICA models. *Neural Processing Letters*, Vol. 22, No. 2, pages 183–204, October 2005.

2 A. Ilin, H. Valpola, and E. Oja. Nonlinear dynamical factor analysis for state change detection. *IEEE Transactions on Neural Networks*, Vol. 15, No. 3, pages 559–575, May 2004.

3 A. Ilin, S. Achard, and C. Jutten. Bayesian versus constrained structure approaches for source separation in post-nonlinear mixtures. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, pages 2181–2186, Budapest, Hungary, July 2004.

4 A. Ilin and A. Honkela. Post-nonlinear independent component analysis by variational Bayesian learning. In *Proceedings of the Fifth International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, pages 766–773, Granada, Spain, September 2004.

5 A. Ilin, H. Valpola, and E. Oja. Semiblind source separation of climate data detects El Niño as the component with the highest interannual variability. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2005)*, pages 1722–1727, Montréal, Québec, Canada, August 2005.

6 A. Ilin and H. Valpola. Frequency-based separation of climate signals. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, pages 519–526, Porto, Portugal, October 2005.

7 A. Ilin, H. Valpola, and E. Oja. Exploratory analysis of climate data using source separation methods. *Neural Networks*, Vol. 19, No. 2, pages 155–167, March 2006.

8 A. Ilin. Independent dynamic subspace analysis. In *Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN 2006)*, pages 345–350, Bruges, Belgium, April 2006.

9 A. Ilin, H. Valpola, and E. Oja. Extraction of climate components with structured variance. In *Proceedings of the IEEE World Congress on Computational Intelligence, (WCCI 2006)*, pages 10528–10535, Vancouver, BC, Canada, July 2006.

List of abbreviations

ADF	Assumed-density filtering
BSS	Blind source separation
DCT	Discrete cosine transform
DFA	Dynamic factor analysis
DSS	Denoising source separation
EEG	Electroencephalogram
EM	Expectation maximization
ENSO	El Niño–Southern Oscillation
EOF	Empirical orthogonal functions
EP	Expectation propagation
FA	Factor analysis
ICA	Independent component analysis
IDSA	Independent dynamics subspace analysis
IFA	Independent factor analysis
ISA	Independent subspace analysis
i.i.d.	Independently and identically distributed
KL	Kullback–Leibler (divergence)
LVM	Latent variable model
MAP	Maximum a posteriori
MEG	Magnetoencephalogram
MLP	Multilayer perceptron (network)
MoG	Mixture of Gaussians
NCEP	National Centers for Environmental Prediction
NCAR	National Center for Atmospheric Research
NDFA	Nonlinear dynamic factor analysis
NFA	Nonlinear factor analysis
NIFA	Nonlinear independent factor analysis
NH	Northern Hemisphere
NSSM	Nonlinear state-space model
PCA	Principal component analysis
pdf	Probability density function
PNL	Post-nonlinear
PNFA	Post-nonlinear factor analysis
RBF	Radial basis function (network)
SSM	State-space model
TJ	Taleb and Jutten's (algorithm)
VB	Variational Bayesian

Mathematical notation

lower- or upper-case letter	scalar, constant or scalar function
bold-face lower-case letter	column vector, vector-valued function
bold-face upper-case letter	matrix

$\langle \cdot \rangle$	Expectation over the aproximating distribution q
$\overline{\theta}$	Mean parameter of the approximating posterior distribution $q(\theta)$
$\widetilde{\theta}$	Variance parameter of the approximating posterior distribution $q(\theta)$
$\mathbf{A}$	Mixing matrix in linear mixtures, $N \times M$
$\mathbf{a}$, $\mathbf{a}_j$	Mixing vector, $N \times 1$
a_{ij}	Mixing coefficient of the j-th source in the i-th observation
$\mathbf{B}$	Matrix of autoregressive dynamics, $M \times M$
$\mathbf{C}$	Sample covariance matrix, $N \times N$
$\mathbf{C}_f$	Sample covariance of filtered data, $N \times N$
$\mathcal{C}$	Cost function
$\mathcal{C}(t)$	The value of the variational Bayesian cost function calculated after obtaining data $\mathbf{x}(1), \ldots, \mathbf{x}(t)$
$D(q\|\|p)$	The Kullback–Leibler divergence between the two distributions q and p
$\mathbf{f}$	Nonlinear generative mapping
f_i	Post-nonlinear distortions in post-nonlinear mixing model
$\mathcal{F}$	Maximized objective function
$\mathbf{F}$, $\mathbf{F}_j$	Filtering matrix
$\mathbf{g}$, $\mathbf{g}_j$	Nonlinear mapping of autoregressive dynamics
$H(x)$	Differential entropy of a continuous random variable x
$h(s)$	The differential entropy rate of a stochastic process $\{s_t\}$
$\widehat{h}_L(t)$	The estimate of the differential entropy rate using a process realization at time instants $t - L + 1, \ldots, t$
$\mathbf{I}$	Identity matrix
$\mathcal{L}$	Logarithm of likelihood or logarithm of posterior
$\mathcal{M}, \mathcal{M}_i$	The model
M	Number of sources (dimensionality of $\mathbf{s}$)
$\mathbf{m}(t), \mathbf{m}_k(t), m_j(t)$	Noise terms in the autoregressive model (innovation process)
N	Number of observations (dimensionality of $\mathbf{x}$)
$\mathbf{n}, \mathbf{n}(t), n_i, n_i(t)$	Observation noise terms

$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian (normal) distribution for variable $\mathbf{x}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$p(x)$	Probability density function evaluated at point x
$q(x)$	Approximating probability density function
s, s_j	Random variable representing one source
$s_j(t)$	The j-th source at (time) index t
$\mathbf{s}$	Random vector of the sources, $M \times 1$
$\mathbf{s}_k$	Random vector representing the k-th group of sources (k-th subspace)
$\mathbf{s}(t)$	Source vector corresponding to the observation vector $\mathbf{x}(t)$, $M \times 1$
$\{s_t\}$	Stochastic process (a sequence of random variables s_t) representing one source
$\mathbf{s}_{1..T}$, $\mathbf{s}_{1..T,j}$	Vector of values of one source at (time) indices $1, ..., T$ (one row of matrix $\mathbf{S}$), $T \times 1$
$\mathbf{S}$	Matrix of M sources with T samples, $M \times T$
T	Number of data samples
$v(t)$, $v_j(t)$	Variance of one source at (time) index t
$\mathbf{V}_{\mathrm{dct}}$	Orthogonal matrix of the DCT basis, $T \times T$
$\mathbf{v}_f$	One row of the matrix $\mathbf{V}_{\mathrm{dct}}$ representing the DCT component with frequency f, $T \times 1$
$\mathbf{w}$, $\mathbf{w}_j$	Demixing vector
$\mathbf{W}$	Demixing matrix
$\mathbf{x}$	Random vector of observations
$\mathbf{x}(t)$	Data vector observed at (time) index t, $N \times 1$
$x_i(t)$	The i-th observation at (time) index t
$\mathbf{x}_f(t)$	Data vector $\mathbf{x}(t)$ after temporal filtering
$\mathbf{X}$	Matrix of N observations with T samples, $N \times T$
$\mathbf{Y}$	Matrix of whitened (sphered) data
$\mathbf{0}$	Vector or matrix with all zeros
$\boldsymbol{\theta}$	Vector of model parameters
$\boldsymbol{\theta}_{\mathbf{f}}$	Parameters of the nonlinear generative mapping $\mathbf{f}$
$\boldsymbol{\Sigma}_{\mathbf{n}}$	Covariance matrix of the observation noise $\mathbf{n}$, $N \times N$
$\boldsymbol{\Sigma}_{\mathbf{m}}$	Covariance matrix of the innovation process $\mathbf{m}$, $M \times M$
$\boldsymbol{\Sigma}_{\mathbf{s}}(t)$	Covariance matrix of the Gaussian prior for $\mathbf{s}(t)$, $M \times M$
$\boldsymbol{\Sigma}_{\mathbf{s}(t),\mathrm{opt}}$	Covariance matrix of the optimal unrestricted posterior $q(\mathbf{s}(t))$, $M \times M$
$\boldsymbol{\varphi}(\cdot)$	Denoising function
$\psi(x)$	Score function evaluated at point x

Chapter 1

Introduction

1.1 Motivation and overview

Collecting data in various types of experiments is a common way of gathering information and accumulating knowledge. A great amount of data appear in all fields of human activity; examples include weather measurements, biomedical signals, economical data, and many others. Analyzing the data can help in many respects to improve the knowledge about observed natural or artificial systems.

The process of acquiring knowledge is called *learning* and this term is widely used in the data analysis literature. A classical learning problem is to estimate dependencies (mapping) between a system's inputs and outputs using some examples of the correct output responses to the given inputs provided by a teacher. Later on, the estimated mapping can be used to produce proper outputs for new input values. This concept is called *supervised learning* (Haykin, 1999; Cherkassky and Mulier, 1998) and practical examples include hand-written character recognition, speech recognition and fault diagnosis for industrial processes.

The present thesis mostly considers learning problems which fall into another domain called *unsupervised learning* (Hinton and Sejnowski, 1999; Oja, 2002). The purpose of unsupervised methods is to analyze available data in order to find some interesting phenomena, regularities or patterns that could be useful for understanding the processes reflected in the data. The knowledge obtained in an unsupervised manner can also be useful for making predictions of the future or making decisions for the purpose of controlling the processes, that is for solving supervised learning tasks.

The learning algorithms considered in this thesis are always based on a model which incorporates our prior knowledge and assumptions about the processes underlying the data. This model may sometimes be constrained by some of

the first principles (e.g., linear models in some applications are motivated by the law of superposition of electromagnetic fields) but more often it is a general mathematical model capable to capture dependencies between different variables.

The methods considered in this thesis are derived using the statistical framework. Classical statistical modeling usually implies using a model of a specific mathematical form and a number of unknown parameters to be estimated. The goal of learning is to infer the values of the unknown parameters based on the observed data. This can be a difficult problem especially for complex models with a great number of parameters, noisy measurements or limited amount of data.

Latent variable models (LVMs) can be useful for capturing important data regularities using a smaller number of parameters. They can also provide a meaningful data representation which may give an insight on the processes reflected in the data. The latter task is solved by so-called *source separation* methods which are the main focus of this thesis. The basic modeling assumption made by the source separation methods is that the observed measurements are combinations of some hidden signals and the goal of the analysis is to estimate these unknown signals from the data. This task cannot be solved without additional assumptions or prior knowledge. A typical assumption used in this problem is *independence* of the processes represented by the hidden signals.

This thesis considers different types of LVMs and different approaches to their estimation. The first half of the work considers so-called *Bayesian* estimation methods which describe unknown parameters using probability distributions. The advantages of the Bayesian theory include its universality for expressing modeling assumptions, its natural treatment of noise and elegant model selection. The research results reported in this part include a study of the properties of variational Bayesian methods and a novel approach designed for a specific type of source separation problems.

The second half of this thesis considers methods which compute point estimates for the unknown quantities. The main advantages of such methods are that they are fast and suit well for large-scale problems. Several new algorithms presented in this part of the work solve the source separation problem by analyzing spectral, dynamic or variance characteristics of the hidden signals.

The methods considered in this thesis can be used in many fields such as biomedical signal processing, speech signal processing or telecommunications. This thesis contains some examples of using the proposed methods in real-world problems. One of the presented applications is the process monitoring task when the model estimated from training data is used for detecting changes in a complex dynamic process. Another interesting application is exploratory analysis of climate data which aims to find interesting phenomena in the climate system using a vast collection of global weather measurements.

1.2 Contributions of the thesis

The most important scientific contributions of this thesis are summarized in the following:

- The properties of variational Bayesian methods are investigated both theoretically and experimentally using linear source separation models.
- A new nonlinear factor analysis model which restricts the generative mapping to the practically important case of post-nonlinear mixtures is presented.
- The variational Bayesian method for learning nonlinear state-space models is applied to the practical problem of change detection in complex dynamic processes.
- Two approaches for source separation based on the frequency contents are presented.
- A computationally efficient algorithm which separates groups of sources by decoupling their dynamic models is proposed.
- An algorithm which extracts components with the most prominent variance structures in the timescale of interest is introduced.
- Several proposed algorithms are applied to spatio-temporal datasets containing global climate measurements from a long period of time.

1.3 Contents of the publications and contributions of the author

This thesis consists of nine publications and an introductory part. The present introductory part aims to provide a general description of the research goals, to give an overview of existing works and to link together different publications of this thesis. The introduction can be read as a separate article but it avoids thorough derivations, for which the reader is addressed to the publications. In any case, the publications should be considered in order to get the full view of the thesis contributions.

The presented work was done in the Laboratory of Computer and Information Science, Helsinki University of Technology. Most of the publications are joint work or done in collaboration with Dr. Harri Valpola, who was the instructor of this thesis. A large portion of the publications is joint work with Prof. Erkki Oja, who was supervising all the work throughout my doctoral studies. In many

cases, the work was done in collaboration or discussed with the members of the research group Bayesian Algorithms for Latent Variable Models led by Prof. Juha Karhunen. Part of the work was done during the author's visit to the Laboratory of Images and Signals (LIS), Institut National Polytechnique de Grenoble, led by Prof. Christian Jutten.

The publications of this thesis can be divided into two parts. The first part (Publications 1–4) deals with variational Bayesian methods applied to different latent variable models. These publications are joint work with Dr. Valpola, Prof. Oja, Dr. Honkela, Dr. Achard and Prof. Jutten, depending on the publications. The second part (Publications 5–9) presents research results on applying source separation methods to exploratory analysis of large-scale climate datasets. This is joint work with Dr. Valpola and Prof. Oja.

The content of the publications and the contributions of the present author are listed in the following. Note that in all cases, writing was a joint effort of the co-authors of the publications.

In **Publication 1**, the properties of the methods based on variational Bayesian learning are studied both theoretically and experimentally. It is shown how the form of the posterior approximation affects the solution found in linear source separation models. In particular, assuming the sources to be independent a posteriori introduces a bias in favor of a solution which has orthogonal mixing vectors. The author ran the experiments in which the effect was detected, derived parts of the considered algorithms and implemented the models with improved posterior approximations.

Publication 2 presents how the variational Bayesian method for nonlinear dynamic factor analysis (NDFA) can be used for detecting abrupt changes in the process dynamics. The changes are detected by monitoring the process entropy rate whose reference value is estimated from training data. It is also possible to analyze the cause of the change by tracking the state of the observed system. The author proposed to use the NDFA algorithm in the change detection problem, participated in the derivations of the test statistic, implemented the change detection algorithm and performed the experiments.

In **Publication 3**, the performance of the variational Bayesian approach to nonlinear independent component analysis (ICA) problem is studied on post-nonlinear test problems. The algorithm is experimentally compared with another popular method of post-nonlinear ICA developed by Taleb and Jutten (1999b). The comparison shows which method is preferred in particular types of problems. This work was done in the LIS laboratory within the framework of the European joint project BLISS on blind source separation and its applications. The author participated in the discussion of the goals of the experimental study and ran the experiments.

Publication 4 presents a new approach for solving the post-nonlinear ICA

problem. It is based on variational Bayesian learning and overcomes some of the limitations of the alternative methods. The author participated in deriving the model, implemented the model and performed the experiments.

In **Publication 5**, it is shown that the well-known El Niño–Southern Oscillation (ENSO) phenomenon can be captured by semiblind source separation methods tuned to extract components exhibiting prominent variability in the interannual time scale. Other interesting components like a component resembling differential ENSO are extracted as well. The author preprocessed the climate dataset, implemented the algorithm and performed the experiments. The original idea of the methodology for exploratory analysis of climate data was due to Dr. Valpola.

Publication 6 proposes a method for rotating components based on their frequency contents. The experimental part shows that the proposed algorithm can give a meaningful representation of slow climate variability as a combination of trends, interannual quasi-periodical signals, the annual cycle and components slowly changing the seasonal variations. The idea and the algorithm were developed together by the authors. The present author implemented the algorithm and performed the experiments.

Publication 7 presents different examples of exploratory analysis of climate data using methods developed in the framework of denoising source separation. The article combines the ideas and results reported in Publication 5 and Publication 6. The additional experiments included in this article were performed by the present author.

Publication 8 presents a method which identifies the independent subspace analysis model by decoupling the dynamics of different subspaces. The method can be used to extract groups of dynamically coupled components which have the most predictable time course.

Publication 9 proposes an algorithm which seeks components whose variances exhibit prominent slow behavior with specific temporal structure. The algorithm is applied to the global surface temperature measurements and several fast changing components whose variances have prominent annual and decadal structures are extracted. The idea and the algorithm were developed together by the authors. The present author implemented the algorithm and performed the experiments.

Chapter 2

Introduction to latent variable models

2.1 Basic latent variable models

The structure of measurement data typically depends on the specific problem domain in which the information is gathered. In some applications, different parts of the data can have certain relations, for example, raw sensor data in image processing applications can be accompanied by object representations with certain properties and relations to each other. This thesis, however, considers only *flat* representations in which data are collected in the form of multivariate measurement vectors $\mathbf{x}(t)$. Each element $x_i(t)$ of the vector $\mathbf{x}(t)$ represents one measurement of the variable x_i and t is the sampling index (e.g., the time instance of the measurement). Such datasets may include various types of time series, for example, sensor data registering video or audio information, weather conditions, electrical activity etc.

The present thesis mostly considers spatio-temporal datasets in which elements of $\mathbf{x}$ correspond to sensors measuring continuous-valued variables in different spatial locations and the index t runs over all time instances in which the measurements are taken. The full set of measurements is often denoted using a matrix $\mathbf{X}$ in which the rows and columns correspond to different sensors and time instances respectively:

$$\mathbf{X} = \left[\mathbf{x}(1), \ldots, \mathbf{x}(t), \ldots, \mathbf{x}(T)\right] . \tag{2.1}$$

An illustration of such a dataset is presented in Fig. 2.1. Only the deviations of the observed variables from their mean values are usually interesting and therefore a usual preprocessing step is centering observations $\mathbf{x}(t)$. It can be done

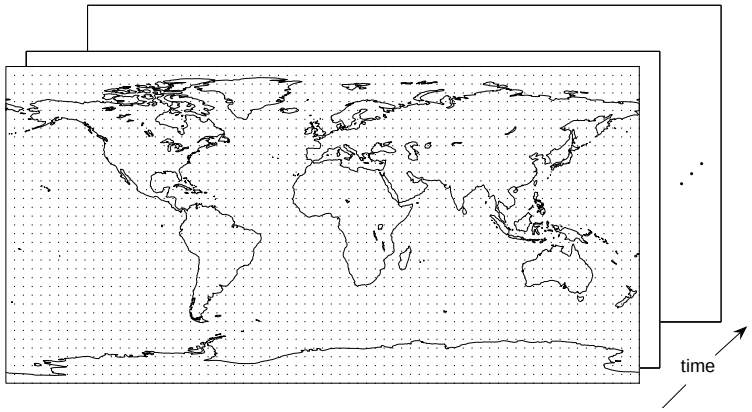


Figure 2.1: An illustration of a spatio-temporal dataset containing global weather measurements. The dots correspond to a $5^\circ \times 5^\circ$ grid (spatial locations) in which the measurements are taken. The measurements made at the same time t all over the globe are collected in one vector $\mathbf{x}(t)$.

by subtracting the estimated mean from each row of the data matrix $\mathbf{X}$. The observations are assumed to be centered everywhere throughout this thesis.

The measured data can be analyzed in many different ways depending on the goals of research. One typical task is to estimate a probabilistic model which covers the regularities in the data, among which the simplest problem is to estimate the probability distribution of the data. The estimated probabilistic model can be used, for example, to make predictions of the future, to detect changes in the process behavior or simply to visualize the data.

The dimensionality of the data matrix $\mathbf{X}$ can be very high in many applications such as exploratory analysis of climate data, image processing and others. However, the processes underlying the data often have limited complexity (Hinton and Sejnowski, 1999; Oja, 2002) and can be described by another set of variables which may have a smaller dimensionality or a simpler (or more interpretable) structure. This is the main modeling assumption used in so-called *latent variable models* (LVMs).

The general property of LVMs is supplementing the set of observed variables with additional latent (hidden) variables (see, e.g., Bishop, 1999a). The relation between the two sets is generally expressed as

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_\mathbf{f}) + \mathbf{n}(t)\,, \tag{2.2}$$

where $\mathbf{x}(t)$ are the observed (measured) variables, $\mathbf{s}(t)$ are the latent variables,

$\mathbf{f}$ is a nonlinear mapping parameterized with vector $\boldsymbol{\theta}_\mathbf{f}$ and $\mathbf{n}(t)$ is a noise term. Different names can be used for the latent variables $\mathbf{s}(t)$ depending on the type of a model; typical terms are *factors*, *components*, *sources* or *states*. The general nonlinear model in Eq. (2.2) can be very difficult to estimate and therefore linear LVMs have gained popularity:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t)\,. \tag{2.3}$$

The matrix $\mathbf{A}$ is usually called a loading matrix or a mixing matrix depending on the context.

The models in Eqs. (2.2)–(2.3) are called *generative* models as they explain the way the data are "generated" from the underlying processes. In unsupervised learning, the parameters of the models such as the hidden variables $\mathbf{s}(t)$, parameters $\boldsymbol{\theta}_\mathbf{f}$, $\mathbf{A}$ of the generative mappings or parameters of the noise distributions are not known and have to be estimated from the observations $\mathbf{x}(t)$.

The remaining sections of this chapter briefly review some of the basic latent variable models and give short descriptions of popular methods for their estimation. We start with some classical tools for dimensionality reduction or data visualization, in which the models in Eqs. (2.2)–(2.3) are sometimes assumed only implicitly. Then, several popular *probabilistic* models are presented. The characteristic of these techniques is describing the hidden sources $\mathbf{s}(t)$ using probability distributions. Finally, models with a meaningful interpretation of the latent variables are discussed. Interpretable solutions are generally found by taking into account some prior information about the hidden signals.

2.1.1 Dimensionality reduction tools

Principal component analysis

Principal component analysis (PCA) is a standard technique for feature extraction, dimensionality reduction or data compression (Diamantaras and Kung, 1996; Oja, 1983). PCA implicitly assumes the linear model in Eq. (2.3) where the dimensionality of the source vector $\mathbf{s}$ is smaller than the dimensionality of the observation vector $\mathbf{x}$. The goal of PCA is to find variables $\mathbf{s}$ such that they would capture most of the data variations and would have less redundancy caused by correlations between variables in $\mathbf{x}$.

The most common derivation of PCA is to maximize the variance of the projected data. The sources are estimated from data using an *orthogonal* matrix $\mathbf{W}$:

$$\mathbf{s}(t) = \mathbf{W}\mathbf{x}(t)\,, \tag{2.4}$$

and the j-th row of $\mathbf{W}$, denoted here by $\mathbf{w}_j^\mathrm{T}$, is found by maximizing the variance of the j-th source $s_j = \mathbf{w}_j^\mathrm{T}\mathbf{x}$ with the constraint that it is orthogonal to the

previously found vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}$. Thus, the maximized criterion is

$$\mathcal{F}_{\text{pca}} = \frac{1}{T}\sum_{t=1}^{T} s_j^2(t) = \frac{1}{T}\sum_{t=1}^{T} \left(\mathbf{w}_j^{\mathrm{T}}\mathbf{x}(t)\right)^2 = \mathbf{w}_j^{\mathrm{T}}\mathbf{C}\mathbf{w}_j \tag{2.5}$$

with $\mathbf{C}$ the sample covariance matrix:

$$\mathbf{C} = \frac{1}{T}\mathbf{X}\mathbf{X}^{\mathrm{T}}\,. \tag{2.6}$$

It follows from the basic linear algebra (see, e.g., Diamantaras and Kung, 1996; Oja, 1983) that the rows of $\mathbf{W}$ are given by the dominant eigenvectors of the matrix $\mathbf{C}$. It can be shown (see, e.g., Diamantaras and Kung, 1996; Hyvärinen et al., 2001) that the principal components are uncorrelated (i.e. their covariance matrix is diagonal) and that the PCA projection minimizes the squared error of the linear reconstruction of $\mathbf{x}(t)$ from the latent variables $\mathbf{s}(t)$.

Nonlinear methods

In case the dimensionality of $\mathbf{s}$ is smaller than the dimensionality of $\mathbf{x}$, the geometrical interpretation of Eq. (2.2) is that data are constrained to a low-dimensional manifold defined by the function $\mathbf{f}(\mathbf{s}, \boldsymbol{\theta}_\mathbf{f})$. This is illustrated in Fig. 2.2. The latent variables $\mathbf{s}(t)$ are then the data coordinates on the manifold. The assumption made by linear methods like PCA is that data lie on a hyperplane. In many cases, however, the structure of the data cloud is more complex and linear methods cannot find its proper representation. With nonlinear models, curved data manifolds, such as the one shown in Fig. 2.2, can be learned and therefore the data variations can be captured by a smaller number of hidden variables. Thus, nonlinear LVMs can be practical tools for dimensionality reduction or feature extraction, and they can efficiently be used in the problems of data visualization, classification or regression.

However, nonlinear models are much more flexible and finding a good nonlinear representation is generally a difficult problem. When flexible models are learned, a serious problem is overfitting when complex models fit perfectly the training data but do not generalize for new data (see, e.g., Bishop, 1995). Practical obstacles for the learning process include multiple local minima and high computational complexity.

Many nonlinear methods for dimensionality reduction find $\mathbf{s}(t)$ so as to preserve the structure of the data when projecting it to the manifold. This is practically implemented by preserving some measure of similarity between data points where typical measures are distance, ordering of the distance, geodesic distance, distance on a graph and others. There is a large number of methods developed in this framework (see, e.g., Lee, 2003; Tipping, 1996; Mardia et al., 1979). The

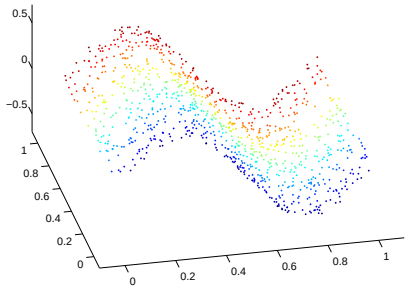


Figure 2.2: Data lying on a two-dimensional manifold embedded in the three-dimensional space.

method called multidimensional scaling (MDS) is the classical technique among such methods (Torgerson, 1952). Other techniques include, for example, Sammon mapping (Sammon, 1969), Isomap (Tenenbaum et al., 2000) or local linear embedding (Roweis and Saul, 2000).

Kernel PCA (Schölkopf et al., 1998) is a method closely connected to MDS (Williams, 1995). The idea of Kernel PCA is to transform the data to a higher-dimension space using an implicitly chosen nonlinear mapping. The sources are then estimated as principal components in the new space. An implicit choice of a suitable transformation makes it possible to do the calculations using a kernel matrix whose dimensionality is restricted to the number of data samples. The method can be used as a feature extraction tool but it is not specifically designed for estimation of nonlinear data manifolds.

Some tools for multivariate data analysis have been implemented in neural network architectures. For example, nonlinear autoassociators (Kramer, 1991; Oja, 1991) use a feedforward neural network with an internal "bottleneck" layer which forces the network to develop a compact representation of the data. A popular data visualization tool is the self-organizing map (Kohonen, 1995) in which the sources are placed on a regular grid and the compact representation is learned by competitive learning. The generative topographic mapping (Bishop et al., 1998) is a probabilistic version of the self-organizing map. A neural network approach for nonlinear data representation and topographic mapping was also developed by Ghahramani and Hinton (1998).

2.1.2 Probabilistic models for dimensionality reduction

There are several probabilistic models which can be used for finding a lower-dimensional representation of data. The simplest ones are based on the linear

generative model in Eq. (2.3) with the Gaussian assumption for the latent variables $\mathbf{s}$. This approach was used, for example, by Tipping and Bishop (1999) in a technique called *probabilistic PCA* and by Roweis (1998) in a similar model called *sensible PCA*. This type of models can be used for the problem of density estimation as it usually requires less parameters than modeling the data $\mathbf{x}(t)$ with the Gaussian distribution. The number of parameters in these models grows linearly with the dimension of $\mathbf{x}$, and yet the model can capture the dominant correlations (Bishop, 1999a; Roweis and Ghahramani, 1999).

Nonlinear probabilistic extensions of PCA assume the general generative model in Eq. (2.2). MacKay (1995a) introduced a probabilistic model called *density networks* in which he showed how to train a multi-layer perceptron (MLP) network without knowing its inputs. Though the assumed model is rather general and resembles Eq. (2.2), the emphasis in the experiments was on predicting binary observations. The proposed training method is based on approximate Bayesian inference and the required integrals are approximated using sampling methods.

A similar approach was used by Bishop et al. (1995) who focused on continuous data and proposed to use radial basis function (RBF) networks for modeling $\mathbf{f}$ in order to reduce the computational complexity of the method. Later, this method evolved into the generative topographic mapping. A good description of the MLP and RBF networks used in the mentioned models can be found, for instance, in the books by Haykin (1999) and by Bishop (1995).

A probabilistic model considered by Valpola and Honkela (Lappalainen and Honkela, 2000; Honkela and Valpola, 2005) is a nonlinear extension of the factor analysis model (see Section 2.2.1). The authors present a *nonlinear factor analysis* (NFA) model in which the mapping $\mathbf{f}$ is modeled by an MLP network, the latent variables $\mathbf{s}$ are assumed uncorrelated and they are described by Gaussian probability distributions. The inference of the unknown parameters is done by variational Bayesian learning.

Recently, Lawrence (2005) has introduced a *Gaussian process LVM* in which the mapping $\mathbf{f}$ is modeled by a Gaussian process (see, e.g., MacKay, 2003). The covariance of the Gaussian process is parameterized with the unknown source values. The sources are described by Gaussian distributions and their values are found as maximum *a posteriori* (MAP) estimates. This model can be seen as a nonlinear extension of probabilistic PCA.

Publication 4 of this thesis presents a model which is close to NFA and which can be used for learning data manifolds of a specific type. The presented model is called *post-nonlinear factor analysis* (PNFA) and it can be seen as the special case of NFA when the general mapping $\mathbf{f}$ is restricted to the practically important case of post-nonlinear mixing structures. The hidden variables are also described by Gaussian distributions and the model is learned using the variational Bayesian principles. The PNFA model is useful for a certain type of source separation

problems and it can overcome some of the limitations of the alternative methods, as explained later in Section 3.3.

2.1.3 Dynamic models

In many situations, the elements $x_i(t)$ of the observed data $\mathbf{x}(t)$ are time series which have a certain temporal structure. For example, successive observations in weather measurements or EEG recordings at a given sensor are correlated. Such correlations can be captured by dynamic models.

Autoregressive processes are classical tools to model temporally structured signals. The basic modeling assumption is that the current observation vector can be roughly estimated from past observations using a linear or nonlinear mapping $\mathbf{g}$:

$$\mathbf{x}(t) = \mathbf{g}(\mathbf{x}(t-1), \ldots, \mathbf{x}(t-D)) + \mathbf{m}(t)\,, \tag{2.7}$$

where the term $\mathbf{m}(t)$ accounts for prediction errors and noise.

Linear autoregressive processes have been studied extensively (see, e.g., Lütkepohl, 1993). The nonlinear autoregressive model is, however, a much more powerful tool. Taken's delay-embedding theorem (Takens, 1981) says that under suitable conditions, the model in Eq. (2.7) can reconstruct the dynamics of a complex nonlinear process provided that the number of delays D is large enough (Haykin and Principe, 1998). In practice, however, the required number of time delays may be too large, which would lead to a great number of parameters in the model and consequently such problems as overfitting.

Latent variable models for dynamical systems are traditionally called *state-space models* and the hidden variables are termed *states*. Dynamic LVMs may decrease the number of parameters required to capture process dynamics (Roweis and Ghahramani, 2001). Another important advantage of these models is the possibility to design an appropriate model from the first principles.

Linear state-space models

Linear state-space models are described by the following equations:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{2.8}$$

$$\mathbf{s}(t) = \mathbf{B}\mathbf{s}(t-1) + \mathbf{m}(t)\,. \tag{2.9}$$

The states $\mathbf{s}(t)$ are usually assumed Gaussian and they follow a first-order autoregressive model. Using only one time delay in Eq. (2.9) does not restrict the generality of the model as any dynamic model with more time delays can be transformed to the model in Eqs. (2.8)–(2.9) by, for example, using an augmented state vector. The observation vectors $\mathbf{x}(t)$ are connected to the states through

a linear mapping $\mathbf{A}$ similarly to Eq. (2.3). The state noise $\mathbf{m}(t)$ and observation noise $\mathbf{n}(t)$ are also assumed Gaussian. Note that many central unsupervised learning techniques can be unified as variations of the basic generative model in Eqs. (2.8)–(2.9) (Roweis and Ghahramani, 1999).

Linear state-space models have been extensively studied in control theory. There, the model usually contains external inputs which affect the observation generation process in Eq. (2.8) and the state evolution in Eq. (2.9). The case of external inputs is not considered in the models presented in this thesis but it is important in many real process monitoring applications. A classical task for state-space models with *known* parameters is estimation of the hidden states $\mathbf{s}(t)$ corresponding to observed vectors $\mathbf{x}(t)$. The standard techniques for solving this problem are Kalman filtering and smoothing (see, e.g., Grewal and Andrews, 1993). Learning a model with *unknown* parameters is termed *system identification* and several approaches exist for the identification of linear state-space models (see, e.g., Ljung, 1987).

In the machine learning community, learning the parameters of linear Gaussian dynamical systems is traditionally done by the expectation-maximization (EM) algorithm (see Section 3.2.2). It was originally derived for linear state-space models with a known matrix $\mathbf{A}$ by Shumway and Stoffer (1982) and reintroduced later for a more general case by Ghahramani and Hinton (1996). The focus of these algorithms is on finding the most likely values for matrices $\mathbf{A}$, $\mathbf{B}$ and the noise covariance matrices.

Nonlinear state-space models

The *nonlinear state-space model* (NSSM) is a much more flexible tool for modeling multivariate time series data. The observation vectors $\mathbf{x}(t)$ are assumed to be generated from the hidden states of a dynamical system through a nonlinear mapping $\mathbf{f}$, and the states follow nonlinear dynamics $\mathbf{g}$:

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{n}(t) \tag{2.10}$$

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1)) + \mathbf{m}(t)\,. \tag{2.11}$$

The terms $\mathbf{n}(t)$ and $\mathbf{m}(t)$ account for modeling errors and noise. The Gaussian distribution is often used to model the states $\mathbf{s}(t)$ and the noise terms.

The geometrical intuition for NSSM is that the dynamic process (flow) described by $\mathbf{s}(t)$ has been either projected or embedded into a manifold using the function $\mathbf{f}$ (Roweis and Ghahramani, 2001). If the dimensionality of $\mathbf{x}$ is larger than the dimensionality of $\mathbf{s}$, the observed sequence forms a flow inside an embedded manifold. Otherwise, some projection of the hidden flow is observed, as shown in Fig. 2.3.

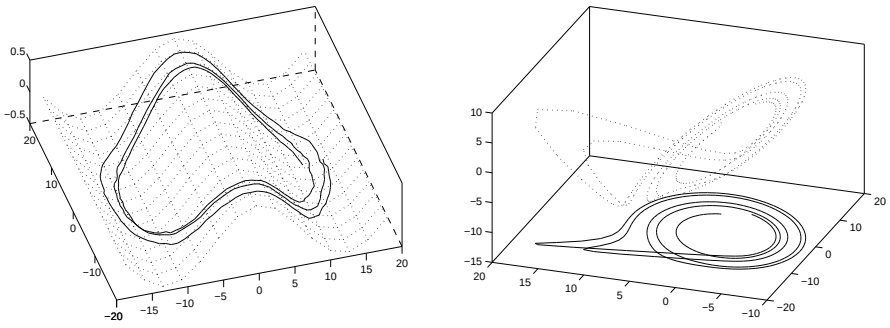


Figure 2.3: An illustration of the nonlinear state-space model assumptions. Left: The observation sequence forms a flow inside an embedded nonlinear manifold. Right: The hidden states describe a three-dimensional dynamic process (dotted line) but the observed flow (solid line) is a two-dimensional projection of the hidden process.

Although nonlinear state-space models are often able to capture the essential properties of a complex dynamical system, they are not in extensive use as it is usually difficult to find a sufficiently accurate model. Even using a NSSM with *known* nonlinearities is not trivial. For example, estimation of the hidden states for a known NSSM is difficult as the nonlinear transformations $\mathbf{f}$ and $\mathbf{g}$ break down the Gaussianity of the state posterior distributions. Several techniques such as the extended Kalman filtering (Grewal and Andrews, 1993), particle filters (Doucet et al., 2001) or unscented transform (Julier and Uhlmann, 1996; Wan and van der Merwe, 2001) have been proposed to do approximate inference.

Estimating an NSSM with *unknown* nonlinearities $\mathbf{f}, \mathbf{g}$ from the observations is much more difficult than learning a linear state-space model. First, the model is very flexible and it contains many unknown parameters including the hidden states. Thus, the main obstacle is overfitting, and some regularization is necessary. Second, there are infinitely many solutions. Any invertible nonlinear transformation of the state-space can be compensated by a suitable transformation of the dynamics and the observation mapping.

Recently, Bayesian techniques have been introduced for identification of nonlinear state-space models. Roweis and Ghahramani (2001) estimate the nonlinearities using RBF networks whose parameters are learned by the EM algorithm. Briegel and Tresp (1999) model the nonlinearities using MLP networks with sampling. Valpola and Karhunen (2002) also used MLP networks for approximating the nonlinear mappings in the model called *nonlinear dynamic factor analysis* (NDFA). All the unknown quantities including the hidden states and the parameters of the MLPs are described by probability distributions inferred with

variational Bayesian learning.

Nonlinear state-space models learned by the NDFA approach are considered in **Publication 2** of this thesis. In particular, it is shown how the model learned with the NDFA algorithm can be used in the problem of detecting changes in process dynamics. The proposed approach to change detection makes use of the cost function provided by the NDFA algorithm in order to monitor the differential entropy rate of the observed process. This quantity is taken as the indicator of change. It is also shown how analyzing the structure of the cost function helps localize a possible reason of change. The important results reported in Publication 2 are outlined in Section 3.5.2.

2.2 Blind source separation

The basic LVMs considered in the previous section are powerful tools for data compression, data visualization or feature extraction. They are able to provide components which capture most of the data variablity and explain correlations present in the data. However, they often provide components of a very limited interpretation, as they are based on very vague prior assumptions. An indicator of this fact is the existence of multiple solutions which satisfy the estimation criteria. For example, a PCA solution derived from probabilistic principles is given by *any orthogonal rotation* of the leading eigenvectors of the sample covariance. This rotation degeneracy is usually fixed by the maximum variance criterion which does not guarantee the interpretability of the results. In nonlinear methods, there is even more ambiguity about the solution as the components are often estimated only up to a nonlinear transformation.

In many applications, the goal is to find components that would have a *meaningful interpretation*. For example, one of the goals of statistical climate data analysis is to find components which would correspond to physically meaningful modes of the weather conditions. Meaningful data representations are typically obtained when some prior knowledge or assumptions (e.g., about the data generation process or the signals underlying the data) are used in the estimation procedure. In this case, the obtained solutions are likely to be explained by domain experts. However, using this type of prior information can be a very difficult task as it requires formalization of the knowledge of the domain experts.

The methods considered in this section are meant for finding interpretable solutions for LVMs. They typically use some prior assumptions that proved plausible and useful in many applications. This may correspond to *exploratory data analysis* when the goal is to find components with distinct and interesting features (a relevant method is *projection pursuit*, see, e.g., Jones and Sibson, 1987). Another possible goal is to solve the *source separation* problem, that is to

extract the signals that would reflect the sources actually generating the data.

2.2.1 Factor analysis

Factor analysis (FA) is a classical statistical tool which was originally used in social sciences and psychology in order to find relevant and meaningful components explaining variability of observed variables (see, e.g., Harman, 1960, for introduction). In FA, the observed variables $\mathbf{x}$ are modeled as linear combinations of some hidden *factors* $\mathbf{s}$ as in Eq. (2.3). The elements of the matrix $\mathbf{A}$ are called factor loadings and $\mathbf{n}(t)$ is an additive term whose elements are called specific factors. The factors are assumed to be mutually uncorrelated Gaussian variables with unit variances. The observation noise $\mathbf{n}$ is assumed Gaussian with a diagonal covariance matrix $\mathbf{\Sigma_n}$.

The unknown parameters of the model including the loading matrix $\mathbf{A}$ and the noise covariance $\mathbf{\Sigma_n}$ should be estimated from the data. There is no closed-form analytic solution for them. Moreover, the FA solution is not unique without some additional constraints. In order to obtain interpretable results, several FA techniques search for such $\mathbf{A}$ that would have only a small number of high loadings and low loadings otherwise. This is implemented in iterative procedures called Varimax, Quartimax or Oblimin rotations (Harman, 1960). Similar approaches have been used in climatology to rotate principal components using general ideas of simple structures in order to obtain components localized either in space or in time (see, e.g., Richman, 1986).

2.2.2 Linear source separation problem

The basic modeling assumption of *linear* source separation methods is very similar to the one of FA. There are some hidden component signals or time series $s_j(t)$, often called *sources*, which are linearly mixed into the multivariate measurements $x_i(t)$:

$$x_i(t) = \sum_{j=1}^{M} a_{ij} s_j(t) + n_i(t), \quad i = 1, ..., N\,, \tag{2.12}$$

where the observation noise term $n_i(t)$ is typically omitted. The index i runs over the measurement sensors (typically spatial locations), and discretized time t runs over the observation period $t = 1, ..., T$. Often, it is assumed that the number N of the observed signals is equal to the number M of the hidden sources.

In the matrix notation, Eq. (2.12) is equivalent to the linear LVM in Eq. (2.3). It is convenient to rewrite this model as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) = \sum_{j=1}^{M} \mathbf{a}_j s_j(t) + \mathbf{n}(t)\,. \tag{2.13}$$

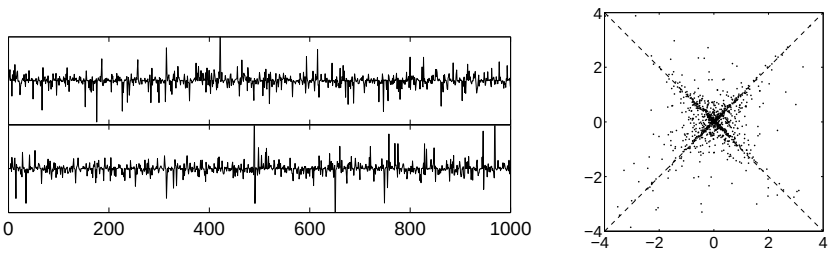


Figure 2.4: Left: Two independent components with non-Gaussian distributions. Right: The joint distribution of two mixtures of these components. The mixing directions are shown with the dashed lines.

The mapping $\mathbf{A}$ is called a *mixing matrix* and it is made up from the coefficients a_{ij} in Eq. (2.12). The columns of matrix $\mathbf{A}$ are denoted here by $\mathbf{a}_j$ and they are called *mixing vectors*.

The goal of the analysis is to estimate the unknown components $s_j(t)$ and the corresponding loading vectors $\mathbf{a}_j$ from the observed data $\mathbf{x}(t)$. With minimum a priori assumptions about the sources, the problem is called *blind source separation* (BSS). A classical example of the source separation problem is the cocktail party problem where several microphones pick up speeches of several people speaking simultaneously and the goal is to separate individual voices from the microphone recordings.

The BSS problem is typically solved by assuming *independence* of the mixed signals s_j. Methods that achieve source separation using some prior information about the unknown parameters are often called *semiblind*. Several existing source separation approaches (see, e.g., Hyvärinen et al., 2001; Cichocki and Amari, 2002, for introduction) are overviewed in the next sections.

2.2.3 Independent component analysis

Independent component analysis (ICA) is a popular method for solving the BSS problem. ICA algorithms identify the model in Eq. (2.13) using only the assumption that the sources are statistically independent. Each $s_j(t)$ is regarded as a sample from a random variable s_j and these variables are assumed mutually independent. Statistical independence will be defined more rigorously in Section 3.1.1. In simple terms, two variables are independent if knowing the value of one variable does not give any information about the value of the other. The statistical independence of the sources implies that these signals are produced by physically independent processes and the goal of the analysis is to separate such processes.

ICA is based on the fundamental result about the separability of linear mixtures (see, e.g., Comon, 1994), which says that using the independence criterion it is possible to estimate sources among which there is at most one Gaussian source. Fig. 2.4 illustrates that linear mixtures of non-Gaussian sources are structured, and therefore the reconstruction of the original sources can be achieved. However, there are well-known ambiguities about the ICA solution. First, the scale (or variance) of the components cannot be determined and therefore the variances of the sources are usually normalized to unity. This still leaves the ambiguity of the sign. Second, the order of the independent components cannot be determined. These ambiguities are known as the scaling and permutation indeterminacies of ICA. They can be solved only with some additional information.

There exist several approaches to solve the ICA problem. Many classical methods consider the noiseless case in which the noise term is omitted from Eqs. (2.12)–(2.13). They typically estimate the sources using a demixing matrix $\mathbf{W}$:

$$\mathbf{s}(t) = \mathbf{W}\mathbf{x}(t)\,. \tag{2.14}$$

Perhaps the most rigorously justified approach to ICA is minimizing the *mutual information* (see, e.g., Cover and Thomas, 1991, for definition) as a measure of dependence between the sources. There are several algorithms based on different approximations of the mutual information, for example, using cumulants (Comon, 1994) or order statistics (Pham, 2000).

It can be shown, however, that the minimization of mutual information is essentially equivalent to maximizing non-Gaussianity of the estimated sources (Hyvärinen et al., 2001). This is a natural result which can be understood from the central limit theorem saying that under certain conditions a linear combination of independent random variables tends toward a Gaussian distribution. Thus, the distributions of the observations x_i should be closer to Gaussian compared to the original sources s_j and the goal of ICA is intuitively to find maximally non-Gaussian components.

FastICA (Hyvärinen and Oja, 1997; Hyvärinen, 1999a) is a popular algorithm based on optimizing different measures of non-Gaussianity. *Kurtosis* (or the fourth-order cumulant) is perhaps the simplest statistical quantity that can be used for indicating non-Gaussianity. It is defined as

$$\mathrm{kurt}(s) = \mathrm{E}\{s^4\} - 3(\mathrm{E}\{s^2\})^2\,. \tag{2.15}$$

where E denotes expectation. The kurtosis is zero for a Gaussian s and is non-zero for many other distributions. However, kurtosis is very sensitive to outliers and therefore other measures are often used. Efficient algorithms can be derived by optimizing some approximations of the quantity called *negentropy*. It is rigorously defined as

$$J(s) = H(s_{\mathrm{gauss}}) - H(s)\,, \tag{2.16}$$

where H denotes the differential entropy of s (Cover and Thomas, 1991) and s_{gauss} is a Gaussian random variable with the same variance as s. The Gaussian variable s_{gauss} has the maximum entropy among all random variables with the same variance and therefore negentropy is always nonnegative and attains zero if and only if s has a Gaussian distribution. Estimating negentropy is very difficult and it is usually approximated using higher-order moments or some appropriately chosen functions (Hyvärinen et al., 2001).

Another popular approach is the maximum likelihood estimation of the demixing matrix $\mathbf{W}$ in Eq. (2.14). In case the dimensionality N of $\mathbf{x}$ equals the dimensionality M of $\mathbf{s}$, the corresponding log-likelihood (see, e.g., Pham et al., 1992) is given by

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{j=1}^{N} \log p_j(\mathbf{w}_j^{\mathrm{T}} \mathbf{x}(t)) + T \log |\det \mathbf{W}| \,, \tag{2.17}$$

where T is the number of samples, $\mathbf{w}_j^{\mathrm{T}}$ denotes the j-th row of matrix $\mathbf{W}$ and the functions p_j are the probability density functions of the sources s_j. The density functions p_j are not known and have to be estimated somehow. It can be shown (see, e.g., Cardoso, 1997) that the maximum likelihood approach is closely related to the Infomax algorithm derived by Bell and Sejnowski (1995) from the principle of maximizing the output entropy of a neural network. In practice, the maximization of the likelihood is considerably simplified using the concept of *natural gradient*, as introduced by Amari et al. (1996).

Another way to achieve independence is based on the theorem saying that two random variables s_1 and s_2 are independent if and only if any of their functions $f(s_1)$ and $g(s_2)$ are uncorrelated (see, e.g., Feller, 1968). Thus, ICA can be performed by *nonlinear decorrelation*, that is by decorrelating some nonlinear transformations of the sources. This approach includes the early algorithm developed by Jutten and Herault (1991), the Cichocki-Unbehauen algorithm (Cichocki and Unbehauen, 1996) and the EASI algorithm by Cardoso and Laheld (1996). The estimating function approach (Amari and Cardoso, 1997) gives a disciplined basis for this. A related approach is Kernel ICA introduced by Bach and Jordan (2002).

Note that independence of random variables is a stronger assumption than uncorrelatedness as it implies uncorrelatedness of any nonlinear transformations of variables. Independence is equivalent to uncorrelatedness only for Gaussian variables, but since there are always infinitely many linear transformations providing uncorrelated sources, ICA is not possible for Gaussian variables. In practice, the preprocessing step called whitening is used in many ICA algorithms in order to remove second-order correlations, and after that, higher-order statistics are considered. However, a problem with higher-order statistics is that their estimates are very sensitive to outliers, which may cause overfitting (Hyvärinen et al., 1999).

Among other ICA approaches, one should mention tensorial methods such as the algorithms called FOBI and JADE (Cardoso, 1989, 1999), methods based on minimizing the mean-square reconstruction error (Karhunen and Joutsensalo, 1994) and variational algorithms (Attias, 1999; Lappalainen, 1999; Miskin and MacKay, 2001; Højen-Sørensen et al., 2002, see also references in Section 3.2.5).

Publication 1 of this thesis presents a theoretical and experimental study of the properties of variational methods in their application to linear ICA models. Two ICA models with non-Gaussian source models are investigated. The presented study shows how the form of the posterior approximation affects the solution found by the variational methods in linear ICA models. In particular, assuming the sources to be independent a posteriori introduces a bias in favor of solutions which have orthogonal mixing vectors. This result suggests that for sources with weak non-Gaussian structure, posterior correlations of the sources should be taken into account in order to achieve good separation performance. This is explained in more details in Section 3.4.

Independent subspace analysis

Multidimensional ICA or *independent subspace analysis* (ISA) is a natural extension of ICA. In this model, the source vector $\mathbf{s}$ in Eq. (2.13) is decomposed into several groups (or linear subspaces):

$$\mathbf{s} = \begin{bmatrix} \mathbf{s}_1^{\mathrm{T}} & \dots & \mathbf{s}_k^{\mathrm{T}} & \dots & \mathbf{s}_K^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} . \tag{2.18}$$

The sources within one group $\mathbf{s}_k$ are generally assumed dependent while components from different groups are mutually independent. Multidimensional ICA has more ambiguity of the solution compared to classical ICA as the sources can be estimated only up to a linear rotation within the subspaces. The problem of estimating such a model was first addressed by Cardoso (1998) and later by Hyvärinen and Hoyer (2000).

2.2.4 Separation using dynamic structure

The basic ICA model considered in the previous section assumes a mixture of random variables, and their statistical independence is used as the only criterion for source separation. No assumption is made there about the order of the data samples $\mathbf{x}(t)$ and therefore the samples can be shuffled in any way without affecting the separation results. In many applications, however, observed signals are time series and their temporal structure can provide additional information which can be used for source separation. An example of temporally structured signals is presented in Fig. 2.5.

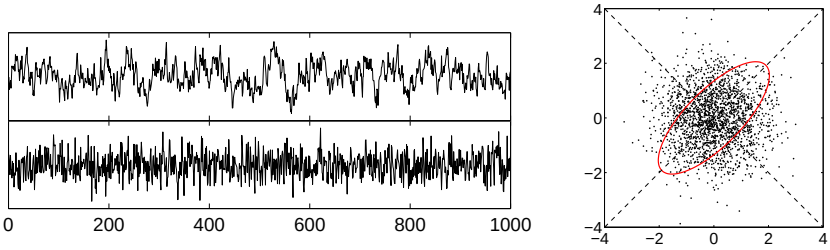


Figure 2.5: Left: Two independent components with distinct temporal structures. Right: The joint distribution of two mixtures of these components. The mixtures are uncorrelated and have unit variances. The mixing directions are shown with the dashed lines. The ellipsoid represents a symmetric lagged covariance matrix $\frac{1}{2}(C_\tau + C_\tau^{\mathrm{T}})$ calculated for $\tau = 1$.

One alternative way to solve the BSS problem is to exploit distinct dynamic structures of the mixed signals. The independence assumption in this case implies that the sources are produced by independent physical processes and a relevant criterion for separation is that the sources should have as little dynamic couplings as possible (cf. this physical independence with statistical independence criterion in basic ICA). In practice, source separation can be performed by decoupling the temporal correlations present in the sources or by explicitly modeling the source dynamics using decoupled predictors. A related approach is based on separating the frequency contents of the sources. The advantage of such methods is that they are typically based on second-order statistics and they can separate sources with Gaussian distributions provided that the sources have different time structures.

Using autocorrelation and frequency structures

The first approach is motivated by the fact that the independent components should have zero cross-covariances calculated for different time lags τ:

$$\mathrm{E}\{s_j(t)s_l(t-\tau)\} = 0\,, \quad j \neq l\,. \tag{2.19}$$

Therefore, BSS can be achieved by joint diagonalization of the covariance matrix $\mathbf{C}$ and the estimate $\mathbf{C}_\tau$ of the time-lagged covariance matrix $\mathrm{E}\{\mathbf{x}(t)\mathbf{x}(t-\tau)^{\mathrm{T}}\}$. The example shown in Fig. 2.5 indicates that the mixing structure can be revealed by analyzing the structure of a lagged covariance matrix. This idea was exploited by several researchers (Tong et al., 1991; Molgedey and Schuster, 1994). Joint diagonalization of several covariance matrices calculated for different time lags usually improves the quality of separation. These principles are used in the

algorithms called SOBI (Belouchrani et al., 1997), TDSEP (Ziehe and Müller, 1998) and in the algorithm proposed by Kawamoto et al. (1997).

Separation of sources can also be achieved by analyzing spectral structures of signals. This is essentially equivalent to using cross-covariances but it is sometimes more natural to formulate the separation criterion in terms of frequencies rather than time lags τ. Different spectral components of independent sources can naturally be assumed uncorrelated and it is therefore possible to separate the sources by joint diagonalization of the data covariance matrix $\mathbf{C}$ and the covariance matrix of the filtered data $\mathbf{x}_f(t)$:

$$\mathbf{C}_f = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_f(t)\mathbf{x}_f(t)^{\mathrm{T}}\,. \tag{2.20}$$

This approach was discussed, for example, by Cichocki and Amari (2002), and a related BSS method was proposed by Stone (2001) where the slowest frequencies are implicitly used for separation. The present approach requires the knowledge of the frequency band in which the separation should be performed. Therefore, the method can be regarded as *semiblind*. In some cases, the choice of the separation frequency band is very natural and follows from the evident signal properties.

If there is no prior on the periodic structure of signals, frequency-based separation can be performed in blinder settings. Gharieb and Cichocki (2003) propose to diagonalize jointly several covariance matrices like in Eq. (2.20) calculated for different frequency bands. This enables separation of signals with distinct spectral contents. Cichocki and Belouchrani (2001) use a bank of adaptive band-pass filters in order to separate sources with prominent dominant frequencies (see also Cichocki and Amari, 2002; Cichocki et al., 2002).

Note that different separation approaches based on analyzing the dynamic structures of the sources are connected. For example, choosing a proper time lag τ for calculating $\mathbf{C}_\tau$ is roughly equivalent to using a specific filter for producing a covariance matrix in Eq. (2.20). Joint diagonalization can therefore include both type of matrices (Gharieb and Cichocki, 2003). Therefore, the results produced by different temporal methods can be quite similar in practice. Note also that the temporal structure of the source signals can vary in time and this information can be taken into account in order to achieve better separation quality. For example, it is possible to use the non-stationarity of the spectral contents in order to separate sources with the same overall frequency contents (see, e.g., Särelä and Valpola, 2005).

Publication 5 of this thesis reports a practical application of the simple semiblind approach based on the joint diagonalization of the data covariance matrix $\mathbf{C}$ and the covariance matrix $\mathbf{C}_f$ of the filtered data defined in (2.20). This frequency-based analysis is implemented following the algorithmic frame-

work of denoising source separation (Särelä and Valpola, 2005) and it is used for exploratory analysis of climate data. An interesting practical result of this analysis is the extraction of the well-known climate phenomenon El Niño–Southern Oscillation as the component with the most prominent variability in the interannual timescale. The practical details of the used algorithm is explained in more details in Section 4.3.1 and the results for the climate data analysis are discussed in Section 4.4.3.

Publication 6 of this thesis presents a more general (blinder) frequency-based separation algorithm. Its aim is to separate the sources by making their spectral contents as distinctive as possible. The algorithm is also implemented in the algorithmic framework of denoising source separation and the separation is achieved by using a competition mechanism between the power spectra of the source estimates. This frequency-based approach is applied to exploratory analysis of global climate measurements and it provides a meaningful representation of the slow climate variability as a combination of trends, interannual quasi-periodical signals, the annual cycle and slowly changing seasonal variations. The proposed algorithm is described in more details in Section 4.4.3 and the results of the climate data application are discussed in Section 4.4.4.

Publication 7 presents a somewhat more detailed exposition of the frequency-based separation approaches considered in this thesis with their application to climate data analysis.

Separation by decoupling dynamic models

An alternative approach is to separate sources by using explicit dynamic models for the sources. The dynamic models are decoupled, which means that the development of each source is explained only from previous measurements of the same source:

$$s_j(t) = g_j(s_j(t-1), \ldots, s_j(t-D)) + m_j(t)\,. \tag{2.21}$$

Together with Eq. (2.13), this equation defines a latent variable model with a linear observation equation and decoupled source dynamics (see Fig. 2.6). Cichocki and Thawonmas (2000) proposed to use linear predictors to model the dynamics g_j and the sources are extracted so as to minimize the prediction errors given by the fitted linear autoregressive models. It is also possible to use a nonlinear predictor g_j modeled by, for example, a multi-layer perceptron or radial basis function network (Cichocki and Amari, 2002). Särelä et al. (2001) used a similar principle in the model called *dynamic factor analysis* (DFA). There, the sources are combined into groups and each group is assumed to follow a separate nonlinear dynamic model. The focus of the experiments was on finding coupled oscillators in MEG data and therefore the sources appeared in pairs.

Two publications of the present thesis deal with similar separation models.

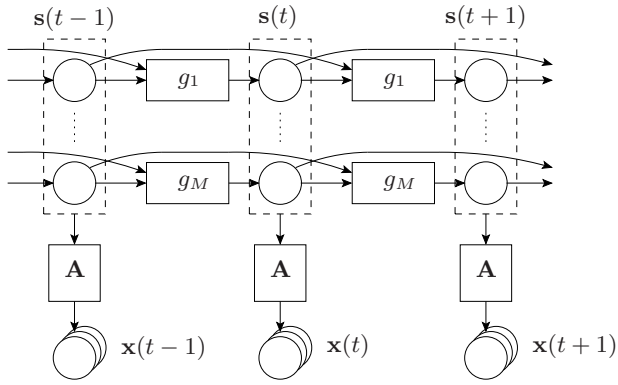


Figure 2.6: An illustration of the linear LVM with decoupled dynamics of the hidden variables.

Publication 1 considers a Bayesian model based on first-order linear predictors as a test problem for studying general properties of variational Bayesian learning in ICA problems. The presented study emphasizes the importance of modeling posterior correlations of the sources in order to achieve good separation quality, as explained later in Section 3.4.

Publication 8 of this thesis presents a method called *independent dynamics subspace analysis* which combines several ideas discussed in this section. The sources are combined into groups (like in ISA) and the independent subspaces are separated by decoupling the dynamic models of the groups. First-order nonlinear predictors are used to model the dynamics of each subspace and the subspaces are extracted so as to minimize the prediction error given by a fitted dynamic model. The model used in this approach is close to DFA but the proposed algorithm is computationally more efficient. The algorithm is described in more details in Section 4.3.3.

2.2.5 Separation using variance structure

The third popular criterion to achieve source separation is to use distinct non-stationary structures of the source variances (activations). The assumption used in this approach is that the variances of independent sources vary independently in time. An example of such signals is presented in Fig. 2.7.

Separation of non-stationary signals was first considered by Matsuoka et al. (1995). They proposed a neural network separation algorithm whose simplified version can be derived from the requirement that the sources are uncorrelated *at any time instant* (Hyvärinen et al., 2001). Then, the sources are estimated

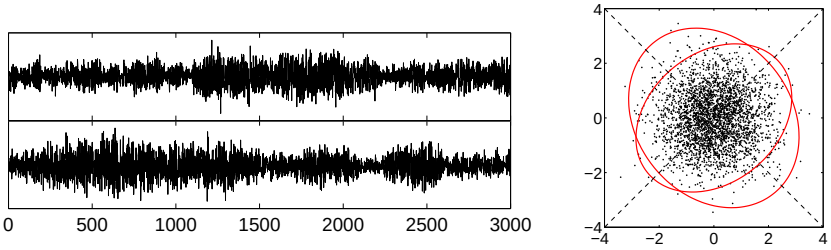


Figure 2.7: Left: Two independent components with temporally structured variances. Right: The joint distribution of two mixtures of these components. The mixtures are uncorrelated and have unit variances. The mixing directions are shown with the dashed lines. The two ellipsoids represent the covariance matrices calculated on two subintervals $\mathcal{T}_1 = [\,1,\,1500\,]$ and $\mathcal{T}_2 = [\,1501,\,3000\,]$.

simultaneously, as in Eq. (2.14), so as to minimize the following measure:

$$\mathcal{C} = \sum_t \sum_j \log v_j(t) - \log |\det \mathrm{E}\{\mathbf{s}_t \mathbf{s}_t^{\mathrm{T}}\}| \,, \tag{2.22}$$

where the source values $\mathbf{s}(t)$ are regarded as samples from random variables $\mathbf{s}_t$ and $v_j(t)$ denotes the variance of the j-th source at time t (a somewhat more detailed explanation of the notation is presented in Section 4.3.4).

There are other separation approaches based on the non-stationary of the sources. Pham and Cardoso (2001) derived a maximum likelihood approach and an algorithm minimizing the Gaussian mutual information. They argue that both approaches can be reduced to joint diagonalization of a set of the data covariance matrices calculated on several subintervals $\mathcal{T}_l$:

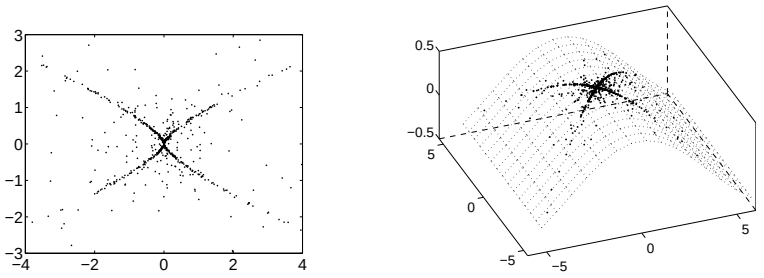
$$\mathbf{C}_{T_l} = \frac{1}{\#\mathcal{T}_l} \sum_{t \in \mathcal{T}_l} \mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}} \,, \tag{2.23}$$

where $\#\mathcal{T}_l$ denotes the number of time instants in $\mathcal{T}_l$. This approach is illustrated in the example shown in Fig. 2.7, where the mixing structure is visible from the structures of two covariance matrices calculated on subintervals.

Hyvärinen (2001) gives an interpretation of non-stationary sources in terms of higher order cross-cumulants

$$\mathrm{E}\{s^2(t)s^2(t-\tau)\} - \mathrm{E}\{s^2(t)\}\,\mathrm{E}\{s^2(t-\tau)\} - 2\,\mathrm{E}\{s(t)s(t-\tau)\}^2 \tag{2.24}$$

and proposes an algorithm maximizing the absolute value of the quantity in Eq. (2.24). Models combining non-stationarity of sources with other separation criteria have also been proposed (see, e.g., Hyvärinen, 2005; Choi et al., 2002).



(a) (b)

Figure 2.8: Examples of nonlinear mixtures of independent sources. Joint distributions of mixtures are shown in two cases: when the dimensionality N of $\mathbf{x}$ equals the dimensionality M of $\mathbf{s}$ (a) and when $N > M$ (b).

Publication 9 of this thesis presents a separation algorithm also based on analyzing the source variance structure. In order to facilitate the analysis of high-dimensional data, we propose to extract components one by one by maximizing a quantity related to the entropy rate and negentropy. This yields an algorithm similar to the one proposed by Matsuoka et al. (1995). We emphasize the possibility to analyze distinct variance structure in different frequency ranges. The proposed algorithm is applied to global climate measurements over a long period of time. A more detailed exposition of the method is presented in Section 4.3.4 and the results of the climate data analysis are discussed in Section 4.4.5.

2.2.6 Nonlinear mixtures

A natural extension of the linear mixing model in Eq. (2.13) is to assume a *nonlinear* mixture model for the observations:

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{n}(t)\,. \tag{2.25}$$

This may be required if the linear model is too simple to describe the mixing process (see, e.g., Almeida, 2005, for a practical example of such mixtures). A nonlinear mixing structure can be quite prominent in the observations, as presented in the examples in Fig. 2.8.

A nonlinear BSS problem is much more difficult compared to the linear case. As pointed out by many researchers (see, e.g., Hyvärinen and Pajunen, 1999; Jutten and Karhunen, 2004), the independence assumption is not sufficient as there exist infinitely many solutions to the *nonlinear ICA* problem. For example, Hyvärinen and Pajunen (1999), generalizing the results of Darmois (1951), describe a procedure that provides a family of nonlinear ICA solutions. Also, the

fact that any nonlinear functions of two independent random variables are also independent shows that the *original* sources can hopefully be estimated only up to nonlinear scaling, if the independence assumption is used alone.

Thus, ICA as a method of finding as independent components as possible does not make much sense in the general nonlinear case. ICA is possible only for some special cases in which structural constraints are imposed on the nonlinear mapping $\mathbf{f}$ (Jutten and Karhunen, 2004). Therefore, the term *nonlinear BSS* is more often used in the context of nonlinear mixtures as it emphasizes that the estimated components should be close to the original sources generating the data. A good introduction to the existing methods of nonlinear BSS can be found in the book by Almeida (2006).

Post-nonlinear mixtures

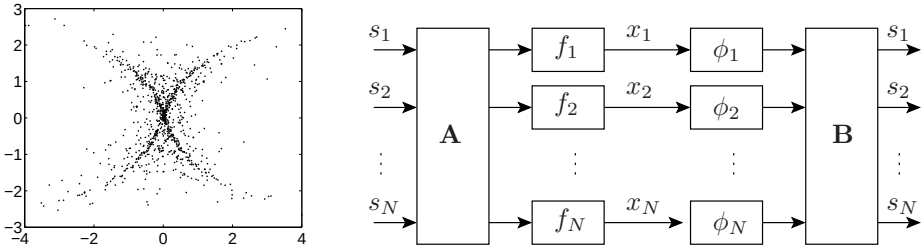
An important special case of the structural constraints are so-called post-nonlinear (PNL) mixtures. These mixtures were first studied by Taleb and Jutten (1999b) and they have the following form:

$$x_i(t) = f_i\left(\sum_{j=1}^{M} a_{ij} s_j(t)\right), \qquad i = 1, \ldots, N\,. \tag{2.26}$$

Thus, the sources are first mixed according to the basic linear model but after that a component-wise nonlinearity f_i is applied to each measuring channel. The post-nonlinearities f_i could correspond, for instance, to sensor nonlinear distortions. The PNL mixing structure and an example of such a mixture is presented in Fig. 2.9.

In the classical post-nonlinear ICA problem, it is typically assumed that the dimensionality N of the observation vector is equal to the number M of the sources and that all the nonlinearities f_i are invertible. Then, the BSS problem can be solved based on the assumption that the sources are statistically independent. Taleb and Jutten (1999b) have shown that if there is at most one Gaussian source in the mixture and the mixing matrix $\mathbf{A}$ (which is made up from the elements a_{ij}) has at least two nonzero entries on each row or column, PNL mixtures are separable with the same scaling and permutation indeterminacies as for linear mixtures.

The classical approach for separating post-nonlinear mixtures is based on minimizing the mutual information (Taleb and Jutten, 1999b,a). The separating structure, as shown in Fig. 2.9b, contains two subsequent stages: a nonlinear stage which cancels the nonlinear distortions by estimating their inverse functions, and a linear stage that solves the standard linear ICA problem. The parameters of the separating systems are estimated by a gradient-based optimization process. The optimization of the mutual information is in practice implemented using a



(a) (b)

Figure 2.9: (a): The distribution of a post-nonlinear mixture of two independent components. The mixtures are generated by applying a nonlinearity to one of the linear mixtures shown in Fig. 2.4. (b): The post-nonlinear mixing structure which produces the outputs x_i, and the separating structure used by Taleb and Jutten.

cumulant-based approximation (Taleb and Jutten, 1999b) or a Gaussian kernel density estimator of the score functions (Taleb and Jutten, 1999a).

Publication 4 of this thesis proposes a new approach for solving the post-nonlinear BSS problem. The proposed algorithm is based on the *post-nonlinear factor analysis* (PNFA) model and can be used in *noisy* PNL mixtures where the number of the measurements is larger than the number of the hidden sources (i.e. $N > M$). Then, as discussed in Section 2.1.1, the data lie on a smaller-dimensional manifold which can be estimated using a probabilistic model. In PNFA, the structure of the manifold is restricted to the post-nonlinear mixing structure for the generative mapping $\mathbf{f}$, as in Eq. (2.26). All the unknown quantities are estimated using variational Bayesian learning. The proposed PNFA algorithm can estimate the original sources only up to a rotation and therefore a standard linear ICA algorithm is applied on the second stage. The advantage of the proposed method is its ability to separate PNL mixture with non-invertible nonlinear distortions f_i provided that the full generative mapping is invertible. The proposed PNFA algorithm is presented in Section 3.3.

General mixtures

General nonlinear BSS is an ill-posed problem and therefore it is necessary to put some additional constraints or to use some sort of regularization in order to find a meaningful solution. One of the earliest algorithms for nonlinear BSS was proposed by Pajunen et al. (1996). They used a somewhat heuristic idea to learn the inverse of the mixing function $\mathbf{f}$ using the self-organizing map. The self-organizing map tries to preserve the structure of the data and therefore the

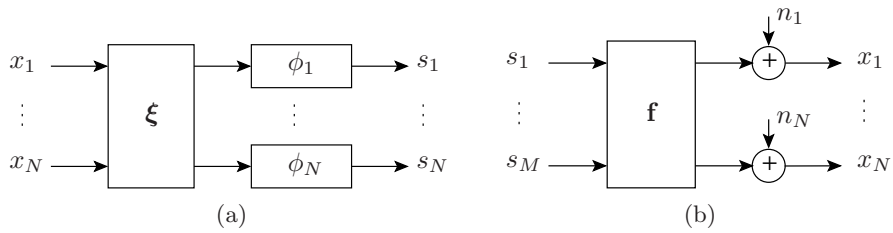


Figure 2.10: (a): The separating structure used by MISEP. (b): The structure of the model learned by NFA and NIFA. For both algorithms, the nonlinearities $\boldsymbol{\xi}$, ϕ_1, ..., ϕ_N, and $\mathbf{f}$ are modeled using MLP networks.

implicit assumption is that the generative mapping should be as simple as possible. Later, Yang et al. (1998) introduced an MLP-based approach where the inverse of the nonlinear mapping $\mathbf{f}$ is restricted to the class of functions approximated by MLPs with the same number of hidden neurons as the number of observations and the number of sources. Tan et al. (2001) use the constraint that the moments of the sources are known. They use an RBF network to learn the inverse of the generative mapping $\mathbf{f}$ and their learning algorithm minimizes the mutual information.

Recently, Almeida (2003) has introduced a nonlinear BSS method called MISEP. He proposed to use an MLP network for learning the inverse of $\mathbf{f}$ and to estimate the parameters of the MLP such that the mutual information between its outputs is minimized. The network is followed by component-wise output nonlinearities modeled by a set of MLPs with bounded outputs, as illustrated in Fig. 2.10a. Almeida uses the idea that minimizing the mutual information between the estimated sources is essentially equivalent to maximizing the output entropy of the separating system. A properly constructed backpropagation procedure is used to learn the parameters of the separating system. Even though MISEP uses rather a general demixing model, the implicit idea of the method is to find as smooth nonlinear transformation as possible, such that the provided components are independent. The smoothness of the mapping can be achieved by any standard regularization used for MLPs (see, e.g., Haykin, 1999). Although the smoothness of the mapping does not guarantee the separation of nonlinear mixtures (Jutten and Karhunen, 2004), MISEP is an elegant solution for the nonlinear BSS problem.

All the methods mentioned so far aim to find the inverse of the nonlinear function $\mathbf{f}$. The alternative approach is to learn the *generative* mapping $\mathbf{f}$ using the model in Eq. (2.25) (see also Fig. 2.10b). This was done by Valpola and Honkela (Lappalainen and Honkela, 2000) in a two-stage separation procedure which is referred as *NFA+FastICA* in this thesis. In the case when the dimensionality of

$\mathbf{x}$ is greater than the dimensionality of $\mathbf{s}$, the data lie on a smaller-dimensional manifold in the observation space, as shown in Fig. 2.8b. Then, this manifold can be learned using the NFA model in which the latent variables are described by Gaussian probability distributions. Based on the central limit theorem, one can assume that the factors found by NFA are some linear combinations of the original independent sources. These factors can be rotated using any algorithm for linear ICA (e.g., FastICA) in order to achieve independence. A similar two-stage approach was later used by Lee et al. (2004).

Valpola and Honkela also developed a modification of the NFA model which takes into account the independence assumption for the sources. The resulting model is called *nonlinear independent factor analysis* (NIFA). Similarly to the linear independent factor analysis technique (Attias, 1999), the sources are described by mixtures of Gaussians. Again, the authors apply the variational Bayesian approach for learning. The proposed NIFA algorithm obtains somewhat better separation quality compared to the NFA+FastICA approach. However, NFA+FastICA is faster and more practical.

There are several approaches to nonlinear ICA which use the temporal structure of the sources to achieve separation. Harmeling et al. (2003) derive kernel-based algorithms and Blaschke and Wiskott (2004) combine nonlinear slow feature analysis (Wiskott and Sejnowski, 2002) with ICA based on temporal decorrelation.

The algorithm for nonlinear dynamic factor analysis (NDFA) developed by Valpola and Karhunen (2002) can also be seen as a method for nonlinear BSS based on temporal structure. The generative model of NDFA follows the standard NSSM equations (2.10)–(2.11). However, the type of posterior approximation used in the proposed learning algorithm favors solutions in which the sources have as little dynamic couplings as possible. This is explained in Section 3.5.1 of this thesis. Thus, the NDFA algorithm favors solutions with dynamically independent sources or subspaces.

Publication 3 of this thesis studies the performance of the NFA+FastICA approach on test problems with post-nonlinear mixtures and experimentally compares it with Taleb and Jutten's algorithm for post-nonlinear mixtures. The study shows the limitations of the two compared methods and the domains of their preferable use. A new interesting result of the presented experiments is that globally invertible PNL mixtures, but with non-invertible component-wise nonlinearities, can be identified and the sources can be separated. This shows the relevance of exploiting more observations than sources. Some results of this study are presented in Section 3.3.1.

Publication 4 presents a PNFA model which can be used to extend the NFA+FastICA approach to the case of post-nonlinear mixtures. These studies are presented in Section 3.3.

2.3 Conclusions

In this chapter, basic latent variable models related to the publications of this thesis have been introduced. Both classical and some recent approaches to learning these models have been outlined. We started from introducing some tools for lower-dimensional data representation, among which PCA is the classical technique. The principles used in some nonlinear tools for dimensionality reduction have been discussed. We also reviewed probabilistic models which either give a probabilistic interpretation for the classical dimensionality reduction techniques or provide some novel nonlinear approaches. In these models, the Gaussian probability distribution is typically used to describe the hidden variables.

Standard probabilistic tools for modeling time series have also been introduced. Linear state-space representation has become a classical modeling technique in this task. It is also a probabilistic LVM with the Gaussian probability model used for the latent variables. Nonlinear state-space models have been studied less extensively because even using a known NSSM is not trivial. Learning an accurate NSSM is a difficult task and there is no classical tool in this problem. We outlined several recent approaches based on approximate Bayesian methods.

Finding a compact data representation can be useful for different tasks such as data compression, information visualization and others. In many applications, it is also desirable that the estimated model would have a meaningful interpretation. For example, individual hidden variables may correspond to independent physical processes underlying the data, and this would provide an insight into the data generation process. The methods considered in the first part of this chapter do not generally provide models with a meaningful interpretation. They can often be used as a first, preprocessing step followed by other techniques which rotate the found components.

Several tools for finding meaningful data representations have been discussed. The classical linear technique is factor analysis which is based on a probabilistic LVM with Gaussian sources. The meaningful representation is achieved by optimizing some measures of structure which are often rather heuristic.

Source separation methods can be seen as extension of factor analysis. These methods typically assume *independence* of the individual hidden variables (called sources), which implies independence of the physical processes represented by individual sources. Separation is done by making the estimated components as independent as possible. In this chapter, three standard ways to achieve source separation have been discussed. The classical approach is ICA when the sources are assumed to have non-Gaussian distributions. The second approach is based on decoupling dynamic structures of the sources, and the third approach uses the non-stationarity of the source variances. Several popular methods for solving the source separation problem have been outlined.

Nonlinear source separation is a much more difficult problem as the independence assumption alone is not enough to find a meaningful nonlinear representation of data. Some additional assumptions have to be used in order to make separation possible. Restricting the generative mapping to the post-nonlinear mixing structure is an important case of such constraints. The general case of nonlinear mixtures can be solved by finding an optimal compromise between the accuracy of the model and its complexity, where simpler models typically imply smoother mappings. Several methods for the general case of nonlinear BSS have been outlined.

During the presentation, the connections between the discussed LVMs and the models considered in the publications of this thesis have been emphasized. Thus, this chapter links together different research results presented in this thesis. It should be noted that the variety of LVMs is not constrained to the models introduced in this chapter. For example, the discussion of so-called *mixture models* or source separation methods for *convolutive mixtures* have been omitted.

Chapter 3

Variational Bayesian methods

3.1 Introduction to Bayesian methods

Bayesian estimation is a principled framework to do inference about unknown parts of a model. The characteristic feature of Bayesian methods is representing all unknown quantities with probability distributions. The unknown parameters of the model (as well as the observed variables) are always assumed to be random variables rather than some deterministic constants. In the Bayesian viewpoint, probability is seen as a measure of our *uncertainty* about the values of a random variable. The solution provided by pure Bayesian methods is always *probabilistic*, that is it contains several possible explanations for the data, accompanied with the probabilities of different explanations. Therefore, Bayesian estimation provides a natural way to overcome the well-known *overfitting* problem when complex solutions explain the training data very well but do not generalize for new data. Other advantages of Bayesian methods include their principled way to do comparison between possible explanations for the data (which is called *model selection*) and the natural treatment of noise.

3.1.1 Basics of probability theory

Let us recall some basic concepts from probability theory (Papoulis, 1991). A popular way to characterize the probability distribution of a continuous variable X is *probability density function* (pdf) $p(x)$ from which the probability that the

variable X takes on a value x on an interval $[a\ \ b]$ is calculated as

$$P(a \leq X \leq b) = \int_a^b p(x)\, dx\,. \tag{3.1}$$

In analogy to physical mass, P is often called *probability mass*. The *joint* density function $p(x, y)$[1] of two random variables X, Y is a function from which the probability that the value of a pair (X, Y) lies in a region A is calculated as

$$P((X, Y) \in A) = \iint_A p(x, y)\, dx\, dy\,. \tag{3.2}$$

This can be easily generalized to the case of multiple variables.

The *marginal* pdfs of the individual variables X or Y can be calculated from the joint pdf $p(x, y)$ using the marginalization principle:

$$p(y) = \int p(x, y)\, dx\,, \qquad p(x) = \int p(x, y)\, dy\,. \tag{3.3}$$

The ratio of the joint pdf and the marginal pdf is called the *conditional* probability density:

$$p(x \,|\, y) = \frac{p(x, y)}{p(y)}\,, \qquad p(y \,|\, x) = \frac{p(x, y)}{p(x)}\,. \tag{3.4}$$

The conditional pdf $p(x \,|\, y)$ can be understood as the uncertainty about the value of X if the value of Y is known.

Two random variables are said to be *independent* if their joint pdf is the product of the two marginals:

$$p(x, y) = p(x)\, p(y)\,. \tag{3.5}$$

It follows from Eq. (3.4) that two random variables are independent if the conditional density of one of the variables does not depend on the value of the other variable, that is

$$p(x \,|\, y) = p(x)\,, \qquad p(y \,|\, x) = p(y)\,. \tag{3.6}$$

In simple terms, two random variables are independent if knowing the value of one variable does not give any information about the value of the other.

The basic principle used by Bayesian methods is the direct consequence of Eq. (3.4). The conditional probability of the unknown variable Y given the value x of the observed variable X can be calculated as

$$p(y \,|\, x) = \frac{p(x, y)}{p(x)} = \frac{p(x \,|\, y) p(y)}{p(x)}\,. \tag{3.7}$$

[1] Following the common practice, $p(\cdot)$ is used as a generic symbol for a pdf, although rigorously subscripts like $p_x(x)$, $p_{x,y}(x, y)$ should be used.

This equation is known as the Bayes rule.

All the above definitions generalize to random vectors (see, e.g., Papoulis, 1991; Hyvärinen et al., 2001).

3.1.2 Density function of latent variable models

In Bayesian methods, all our assumptions about the data structure are expressed in the form of a joint pdf over all the known and unknown variables. This thesis considers latent variable models

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_\mathbf{f}) + \mathbf{n}(t)\,, \tag{3.8}$$

for which the joint pdf always includes the observed variables $\mathbf{X}$, the hidden variables $\mathbf{S}$ and the rest of the parameters $\boldsymbol{\theta}$ (e.g., the parameters $\boldsymbol{\theta}_\mathbf{f}$ of the generative mapping $\mathbf{f}$). Here, we can assume that the matrix $\mathbf{S}$ of source values is defined similarly to Eq. (2.1).

The joint pdf for all the probabilistic LVMs considered in this thesis is expressed in the following form:

$$p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta}) = p(\mathbf{X} \,|\, \mathbf{S}, \boldsymbol{\theta})\, p(\mathbf{S}, \boldsymbol{\theta})\,. \tag{3.9}$$

The term $p(\mathbf{X} \,|\, \mathbf{S}, \boldsymbol{\theta})$ is called the *likelihood* of $\mathbf{S}$ and $\boldsymbol{\theta}$ and it reflects our assumptions on the way the data $\mathbf{X}$ are generated from the hidden variables $\mathbf{S}$. As an example, consider a linear model

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{3.10}$$

with the Gaussian assumption for the observation noise $\mathbf{n}(t)$. The corresponding likelihood factor is given by:

$$\prod_{t=1}^{T} p(\mathbf{x}(t) \,|\, \mathbf{s}(t), \boldsymbol{\theta}) = \prod_{t=1}^{T} N\left(\, \mathbf{x}(t) \,|\, \mathbf{A}\mathbf{s}(t),\, \boldsymbol{\Sigma}_\mathbf{n} \,\right)\,. \tag{3.11}$$

Here and throughout this thesis, $N\left(\, \mathbf{x} \,|\, \boldsymbol{\mu},\, \boldsymbol{\Sigma} \,\right)$ denotes the Gaussian (or normal) distribution over $\mathbf{x}$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

The terms $p(\mathbf{S}, \boldsymbol{\theta})$ in the density model in Eq. (3.9) define our *prior* uncertainty (prior expectations) about the values of the unknown parameters $\mathbf{S}$, $\boldsymbol{\theta}$. For example, the simple factor analysis model specifies the same prior distribution for each $\mathbf{s}(t)$:

$$p(\mathbf{s}(t) \,|\, \boldsymbol{\theta}) = N\left(\, \mathbf{s}(t) \,|\, \mathbf{0},\, \mathbf{I} \,\right)\,, \tag{3.12}$$

where $\mathbf{0}$ is a vector containing all zeros and $\mathbf{I}$ denotes the identity matrix. In dynamic models, the prior source distribution is more complex and it takes into account the source dynamics defined, for example, by Eq. (2.9):

$$p(\mathbf{s}(t) \,|\, \mathbf{s}(t-1), \boldsymbol{\theta}) = N\left(\, \mathbf{s}(t) \,|\, \mathbf{B}\mathbf{s}(t-1),\, \boldsymbol{\Sigma}_\mathbf{m} \,\right)\,. \tag{3.13}$$

Assigning priors for the rest of the parameters $\boldsymbol{\theta}$ can be nontrivial. When it is desirable to introduce minimum information in the prior so that the solution would be maximally defined by the likelihood, *noninformative priors* are used (Gelman et al., 1995). In many cases, however, suitably chosen priors bias the model in favor of specific types of solutions. For example, when the generative mapping $\mathbf{f}$ is modeled by an MLP network, using so-called weight decay priors for the parameters of the MLP can penalize non-smooth solutions for $\mathbf{f}$ (see, e.g., Haykin, 1999).

It should be noted that selecting priors is the most subjective part of Bayesian methods. One should generally specify all plausible values for the unknown quantities and express the prior expectations in the form of pdf. A specific form of pdf can be chosen in order to enable mathematical tractability of further inference.

3.1.3 Bayesian inference

The density model $p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta})$ expresses all our assumptions about the modeled process. Once the density model is defined, all one has to do is to infer the probabilistic solution for the unknown parts of the model. This is generally done by applying the Bayes rule in Eq. (3.7) in order to find the conditional distributions of the unknown parameters given the observations:

$$p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{X})} = \frac{p(\mathbf{X} \,|\, \mathbf{S}, \boldsymbol{\theta})\, p(\mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{X})} \,. \tag{3.14}$$

Here, the numerator is the full joint pdf and the denominator is the marginal pdf of the observed variables. The pdf in Eq. (3.14) is called the *posterior* pdf as it expresses our uncertainty about the values of the unknown variables after the measurements $\mathbf{X}$ have been obtained. The posterior pdf is always a compromise between the prior $p(\mathbf{S}, \boldsymbol{\theta})$ and the likelihood $p(\mathbf{X} \,|\, \mathbf{S}, \boldsymbol{\theta})$.

Computing the posterior distribution of the unknown quantities is a central problem of Bayesian methods. Evaluation of the posterior is relatively easy for simple models with so-called *conjugate* priors (see, e.g., Gelman et al., 1995) when the parametric form of $p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})$ in Eq. (3.14) is known. However, computing the posterior is generally a difficult task and, in most cases, the posterior has to be approximated somehow.

The evaluated posterior pdf is usually used for further inference or decision making. For example, one may want to compute the probability distribution for a future measurement $\mathbf{x}(t)$ given the observed data $\mathbf{X}$. Such a density function $p(\mathbf{x}(t) \,|\, \mathbf{X})$ is called a *predictive* pdf. As an example, let us consider the predictive

pdf for the model described in Eqs. (3.11)–(3.12):

$$\begin{aligned} p(\mathbf{x}(t) \,|\, \mathbf{X}) &= \iint p(\mathbf{x}(t), \mathbf{s}(t), \boldsymbol{\theta} \,|\, \mathbf{X}) \, d\mathbf{s}(t) \, d\boldsymbol{\theta} \\ &= \iint p(\mathbf{x}(t), \mathbf{s}(t) \,|\, \boldsymbol{\theta}, \mathbf{X}) \, p(\boldsymbol{\theta} \,|\, \mathbf{X}) \, d\mathbf{s}(t) \, d\boldsymbol{\theta} \,. \end{aligned} \tag{3.15}$$

Now we note that the likelihood $p(\mathbf{x}(t), \mathbf{s}(t) \,|\, \boldsymbol{\theta}, \mathbf{X})$ does not depend on $\mathbf{X}$ and integrating out the source $\mathbf{s}(t)$ yields

$$\int p(\mathbf{x}(t), \mathbf{s}(t) \,|\, \boldsymbol{\theta}, \mathbf{X}) \, d\mathbf{s}(t) = p(\mathbf{x}(t) \,|\, \boldsymbol{\theta}) \,. \tag{3.16}$$

This gives the predictive pdf in the following form:

$$p(\mathbf{x}(t) \,|\, \mathbf{X}) = \int p(\mathbf{x}(t) \,|\, \boldsymbol{\theta}) \, p(\boldsymbol{\theta} \,|\, \mathbf{X}) \, d\boldsymbol{\theta} \,. \tag{3.17}$$

The predictive probability in Eq. (3.17) can be understood as a sum of separate probabilistic models $p(\mathbf{x}(t) \,|\, \boldsymbol{\theta})$ weighted by their posterior probabilities $p(\boldsymbol{\theta} \,|\, \mathbf{X})$. Thus, the pure Bayesian approach takes into account a set of possible models, which offers a good compromise between under- and overfitting, that is using too simple or complex models in light of the available data.

The same averaging principle should also be used in case of a discrete set of possible models $\mathcal{M}_i$. In this context, each possible density model is often written as $p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta} \,|\, \mathcal{M}_i)$ as it expresses some structural assumptions $\mathcal{M}_i$. One may assign a prior distribution over model structures $p(\mathcal{M}_i)$ and then average similarly to Eq. (3.17) using as the weights the posterior probablities of the models

$$p(\mathcal{M}_i \,|\, \mathbf{X}) = \frac{p(\mathcal{M}_i) \, p(\mathbf{X} \,|\, \mathcal{M}_i)}{\sum_i p(\mathcal{M}_i) \, p(\mathbf{X} \,|\, \mathcal{M}_i)} \,. \tag{3.18}$$

The term $p(\mathbf{X} \,|\, \mathcal{M}_i)$ in Eq. (3.18) is called the *evidence* (or marginal likelihood) for the model $\mathcal{M}_i$. It appears as the denominator (normalization constant) in the posterior distribution in Eq. (3.14).

A pragmatic approach, however, is to select one model among the possible ones and use it for future inference. In general, the most suitable model should be chosen depending on the goals and some utility function is needed in order to assess the usefulness of a model. For example, for prediction and decision problems, comparison and selection between Bayesian models can be done by the assessment of the predictive abilities of the models (see, e.g., Vehtari and Lampinen, 2002). Since it is often difficult to find a proper utility function which should include all informal knowledge of domain experts, a practical approach is

to select the most probable model, that is the one that maximizes the posterior in Eq. (3.18).

Computation of the model evidence $p(\mathbf{X} \,|\, \mathcal{M}_i)$ is another central problem of Bayesian methods. It is formally calculated by integrating out the unknown parameters from the joint pdf:

$$P(\mathbf{X} \,|\, \mathcal{M}) = \iint p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta} \,|\, \mathcal{M}) \, d\mathbf{S} \, d\boldsymbol{\theta} \,. \tag{3.19}$$

However, this integral is intractable in most cases and some approximations have to be made.

3.2 Approximate Bayesian methods

This section considers the classical methods for evaluating the posterior distribution of the unknown parameters $p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})$. Computation of the posterior is an important problem as the posterior can be used, for example, to infer the most probable values of the unknown parameters, to calculate the predictive distribution, or to approximate the integral defining the evidence in Eq. (3.19). As was pointed out in the previous section, the posterior can be calculated exactly only for simple models and some sort of approximation is typically required.

Posterior approximations can be useful in practice as, for example, they can reduce the information in the posterior to the neighborhood of one particular solution. Sometimes, they can also regularize the estimation problem. However, the possible negative side effects of posterior approximation is overfitting, as some of the probable models are typically discarded from the posterior.

3.2.1 MAP and sampling methods

Perhaps the simplest way to approximate the posterior distribution is the *maximum a posteriori* (MAP) estimation, in which the posterior is characterized by the values that maximize it:

$$\{\mathbf{S}_{\mathrm{MAP}}, \boldsymbol{\theta}_{\mathrm{MAP}}\} = \arg\max_{\mathbf{S}, \boldsymbol{\theta}} p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X}) \,. \tag{3.20}$$

The MAP estimation is equivalent to the popular *maximum likelihood* (ML) method under the assumption that the prior for the unknown parameters $p(\mathbf{S}, \boldsymbol{\theta})$ is uniform and therefore the posterior is proportional to the likelihood $p(\mathbf{X} \,|\, \mathbf{S}, \boldsymbol{\theta})$.

The main advantage of the MAP estimate is its simplicity because to maximize the posterior can be a relatively easy task. However, its main drawback is reducing the full posterior to only one point. For example, MAP estimation does not generally provide the confidence regions showing the posterior uncertainty

about the MAP estimates. Another possible problem is overfitting. For complex models without proper regularization, it is possible that the MAP estimate corresponds to a narrow peak in the posterior and therefore the estimate can be very sensitive to small changes in the data.

A somewhat improved approach is the *Laplace approximation* which uses a local Gaussian approximation of the posterior around the MAP estimate (MacKay, 1995c). The covariance matrix of the Gaussian approximation is taken as the Hessian matrix of the log-posterior. However, the drawback of this approach is that the approximation can be poor, especially for small datasets, and the calculation of the Hessian can be computationally expensive. Note also that the Laplace approximation can detect overfitting problems but not fix them directly.

Markov chain-Monte Carlo methods (Neal, 1993) approximate the posterior by a collection of samples drawn from it. These methods are typically very slow and computationally very demanding. Also, it is generally difficult to assess the convergence of the sampling procedure. Another problem is that sampling methods require that all the samples drawn from the posterior be stored for future inference, which is usually memory consuming. Despite their drawbacks, sampling methods are very popular because they are easy to implement and to use. These methods are often preferred if they are computationally feasible.

3.2.2 The EM algorithm

The *Expectation-Maximization* (EM) algorithm is the extension of the ML/MAP estimation[2] to the case when some of the unknown parameters are uninteresting or unimportant for future inference (they are called *nuisance* parameters). As an example, consider the predictive distribution in Eq. (3.17) where all relevant information is contained in the marginal posterior $p(\boldsymbol{\theta} \,|\, \mathbf{X})$. The sources $\mathbf{S}$ are not important for future inference and therefore they can be integrated out from the posterior. Thus, the problem addressed by the EM algorithm is to find the MAP estimate for the set of interesting parameters (typically $\boldsymbol{\theta}$) in the presence of nuisance parameters (typically the hidden variables $\mathbf{S}$):

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \,|\, \mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \int p(\boldsymbol{\theta}, \mathbf{S} \,|\, \mathbf{X})\, d\mathbf{S}\,. \tag{3.21}$$

The classical presentation of the algorithm was done by Dempster et al. (1977) but here we follow the view of the EM algorithm presented by Neal and Hinton (1999).

The function that should be maximized is the logarithm of the marginal posterior:

$$\mathcal{L} = \log p(\boldsymbol{\theta} \,|\, \mathbf{X}) = \log \int p(\boldsymbol{\theta}, \mathbf{S} \,|\, \mathbf{X})\, d\mathbf{S}\,. \tag{3.22}$$

[2]The EM algorithm originally extends the ML approach but is applicable to MAP too.

However, it is possible to estimate the lower bound of $\mathcal{L}$ using any distribution over the hidden variables $q(\mathbf{S})$:

$$\mathcal{L} = \log \int q(\mathbf{S}) \frac{p(\boldsymbol{\theta}, \mathbf{S} \mid \mathbf{X})}{q(\mathbf{S})} \, d\mathbf{S} \geq \int q(\mathbf{S}) \log \frac{p(\boldsymbol{\theta}, \mathbf{S} \mid \mathbf{X})}{q(\mathbf{S})} \, d\mathbf{S} = \mathcal{F}(q, \boldsymbol{\theta}) \,, \tag{3.23}$$

which holds due to Jensen's inequality. The lower bound $\mathcal{F}(q, \mathbf{S})$ is the actual functional optimized in the EM algorithm.

The optimization of $\mathcal{F}$ is practically done by alternate updating $q(\mathbf{S})$ and $\boldsymbol{\theta}$ in the steps called *E-step* and *M-step* (Ghahramani and Beal, 2001). In the following, these steps are presented using the notation $\boldsymbol{\theta}^{(k)}$ and $q^{(k)}(\boldsymbol{\theta})$ for the instances computed on the k-th iteration:

1. The E-step maximizes $\mathcal{F}(q, \boldsymbol{\theta})$ w.r.t. the distribution over the latent variables $q(\mathbf{S})$ given the fixed parameters $\boldsymbol{\theta}^{(k-1)}$:

$$q^{(k)}(\mathbf{S}) = \arg\max_{q(\mathbf{S})} \mathcal{F}(q(\mathbf{S}), \boldsymbol{\theta}^{(k-1)}) \,. \tag{3.24}$$

   It can be shown that the optimal $q^{(k)}(\mathbf{S})$ is the posterior distribution of $\mathbf{S}$ given the fixed value $\boldsymbol{\theta}^{(k-1)}$:

$$q^{(k)}(\mathbf{S}) = p(\mathbf{S} \mid \mathbf{X}, \boldsymbol{\theta}^{(k-1)}) \,. \tag{3.25}$$

2. The M-step optimizes $\mathcal{F}(q, \boldsymbol{\theta})$ w.r.t. the parameters $\boldsymbol{\theta}$ given the fixed distribution $q^{(k)}(\mathbf{S})$:

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q^{(k)}(\mathbf{S}), \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \int q^{(k)}(\mathbf{S}) \log p(\boldsymbol{\theta}, \mathbf{S} \mid \mathbf{X}) \, d\mathbf{S} \,. \tag{3.26}$$

   The term $-\int q(\mathbf{S}) \log q(\mathbf{S}) \, d\mathbf{S}$ is removed from $\mathcal{F}$ in Eq. (3.26) as it does not depend on $\boldsymbol{\theta}$.

It can be shown that each iteration of the presented procedure always increases the true posterior $p(\boldsymbol{\theta} \mid \mathbf{X})$ or leaves it unchanged. This algorithm converges to a local maximum of the posterior except in some special cases.

The *generalized EM* algorithm extends the classical EM algorithm by making partial M-steps, when the parameters $\boldsymbol{\theta}$ are updated so as to increase the functional $\mathcal{F}(q, \boldsymbol{\theta})$ but not necessarily maximize it. In the extension proposed by Neal and Hinton (1999), partial E-steps are made as well: the functional $\mathcal{F}(q, \boldsymbol{\theta})$ is increased but not necessarily maximized w.r.t. to the distribution $q(\mathbf{S})$. In practice, this could speed up the convergence of the EM algorithm.

3.2.3 Variational Bayesian learning

Recently, *variational Bayesian* (VB) learning has been widely used in Bayesian latent variable models. Its goal is to approximate the actual posterior probability density of the unknown variables by a function with a restricted form. This approach was introduced in the neural network literature by Hinton and van Camp (1993) and the term *ensemble learning* was also used to describe the method (MacKay, 1995b; Lappalainen and Miskin, 2000). VB learning is closely related to variational *mean-field* methods (Jaakkola, 2000; MacKay, 2003).

In the case of latent variable models, the approximating distribution $q(\mathbf{S}, \boldsymbol{\theta})$ is defined over the sources $\mathbf{S}$ and the other parameters $\boldsymbol{\theta}$. The goodness of fit between the two probability density functions $p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})$ and $q(\mathbf{S}, \boldsymbol{\theta})$ is measured by the Kullback-Leibler divergence:

$$D(q(\mathbf{S}, \boldsymbol{\theta}) \,||\, p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})) = \int q(\mathbf{S}, \boldsymbol{\theta}) \log \frac{q(\mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})} \, d\mathbf{S} \, d\mathbf{X} \,. \tag{3.27}$$

The Kullback-Leibler (KL) divergence is a standard dissimilarity measure for probability densities (see, e.g., Cover and Thomas, 1991). It is always nonnegative and attains the value zero if and only if the two distributions are equal. Therefore, the pdf $q(\mathbf{S}, \boldsymbol{\theta})$ is optimized to get the approximation as close to the true posterior as possible. Interpreted in information-geometric terms (Amari and Nagaoka, 2000), minimizing the KL divergence means finding the projection of the true pdf $p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})$ on the manifold of the approximating densities $q(\mathbf{S}, \boldsymbol{\theta})$.

Unfortunately, the KL divergence is difficult to compute as the posterior $p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})$ in Eq. (3.27) includes the term $p(\mathbf{X})$ which cannot be evaluated. However, as it is constant w.r.t. $q(\mathbf{S}, \boldsymbol{\theta})$, it can be subtracted from Eq. (3.27), and the actually minimized function is

$$\begin{aligned} \mathcal{C}(q) &= D(q(\mathbf{S}, \boldsymbol{\theta}) \,||\, p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})) - \log p(\mathbf{X}) \\ &= \int q(\mathbf{S}, \boldsymbol{\theta}) \log \frac{q(\mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X}) p(\mathbf{X})} \, d\mathbf{S} \, d\mathbf{X} \\ &= \int q(\mathbf{S}, \boldsymbol{\theta}) \log \frac{q(\mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta})} \, d\mathbf{S} \, d\mathbf{X} \,. \end{aligned} \tag{3.28}$$

It follows from the nonnegativity property of the KL divergence that the cost function in Eq. (3.28) gives the lower bound for the model evidence:

$$-\mathcal{C}(q) \leq \log p(\mathbf{X}) \,. \tag{3.29}$$

Therefore, the VB approach is sometimes seen as a way to optimize the lower bound for the marginal likelihood $p(\mathbf{X} \,|\, \mathcal{M})$.

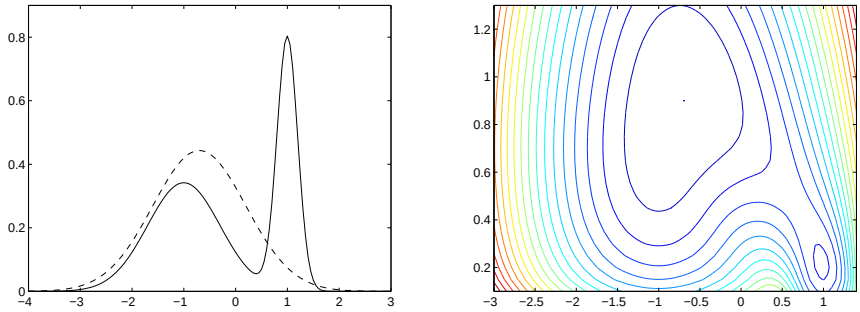


Figure 3.1: Left: A hypothetical posterior distribution p (solid) approximated by a Gaussian distribution q (dashed) so as to minimize the KL divergence $D(q\,||\,p)$. Right: The KL divergence as a function of the mean (abscissa) and the standard deviation (ordinate) of the approximating Gaussian distribution. The wider posterior mode corresponds to the global minimum of the KL divergence.

The posterior approximation $q(\mathbf{S}, \boldsymbol{\theta})$ has to be tractable and therefore it is always chosen to have a suitably factorial form. In latent variable models, at least the sources $\mathbf{S}$ are typically assumed independent a posteriori of the rest of the parameters $\boldsymbol{\theta}$:

$$q(\mathbf{S}, \boldsymbol{\theta}) = q(\mathbf{S})q(\boldsymbol{\theta})\,. \tag{3.30}$$

The optimization of the cost function is done by alternate updating the factors of q. For example, if q is factorized as in Eq. (3.30), $q(\mathbf{S})$ and $q(\boldsymbol{\theta})$ are alternately updated, each while holding the other fixed.

Characteristics of variational Bayesian learning

In flexible models, the true posterior typically has multiple peaks and each peak corresponds to one possible explanation for the data. Let us present an example which shows that the VB approximation usually captures only the neighborhood of one of the posterior modes. Thus, the VB approximation typically underestimates the posterior uncertainty about the unknown parameters.

Fig. 3.1 presents a hypothetical bimodal posterior distribution approximated by a Gaussian distribution so as to minimize the KL divergence in Eq. (3.27). The cost function presented in the right plot of Fig. 3.1 has two local minima, each corresponding to one of the two modes of the posterior. The global minimum, however, corresponds to the wider peak that contains more probability mass. Note also that in practice the wider peak could be more attractive for the optimization procedure.

VB learning has gained popularity because of its attractive characteristics that we summarize in the following:

1. The VB cost function provides the lower bound of the model evidence $p(\mathbf{X} \,|\, \mathcal{M}_i)$, which allows for elegant model selection.

2. The VB approximation is sensitive to high posterior mass in contrast to the MAP estimation which is sensitive to high posterior densities (Lappalainen and Miskin, 2000). Thus, VB learning is less subject to overfitting and provides more robust solutions.

3. Selecting a suitable form for the posterior approximation corresponds to a specific regularization of the solution, which also helps avoid overfitting and sometimes makes the estimation problem well-posed (see Section 3.4 and Publication 1).

4. Using densities for representing the unknown quantities preserves more information about the full posterior, compared to point estimates. For example, if necessary, the mean $\langle\theta\rangle = \int \theta q(\theta)\, d\theta$ can be taken as a point estimate of the parameter θ, and the variance $\int (\theta - \langle\theta\rangle)^2 q(\theta)\, d\theta$ can define a confidence region for the point estimate $\langle\theta\rangle$. Note also that the approximating distribution could be used for sampling from the true posterior (Ghahramani and Beal, 2001).

However, applying VB methods can be difficult in practice because of the following problems:

1. One of the main drawbacks of VB methods is their high computational complexity, which often results in long time until convergence.

2. The cost function usually has multiple local minima and it can be difficult to find the global minimum because of the slow convergence.

3. VB may tend to converge to solutions that correspond to wider posterior modes. Such solutions typically provide simpler explanations of the data. Therefore, VB methods may suffer from the underfitting problem.

4. Too simple posterior approximations usually result in an efficient learning algorithm but they can introduce a bias in favor of some type of solutions (see Section 3.4 and Publication 1).

3.2.4 Other approaches related to VB learning

The view presented by Neal and Hinton (1999) helps understand the relation between the EM algorithm and the VB approach. The generalized EM algorithm

can be seen as the special case of the VB approach when the approximating distribution $q(\mathbf{S}, \boldsymbol{\theta})$ uses point estimates for one set of parameters $\boldsymbol{\theta}$ and distributions for the other set $\mathbf{S}$. In such a viewpoint, a point estimate for a scalar can be seen as a uniform distribution defined on an interval of infinitely small, but fixed length. Then, the E-step and the M-step can be seen as the two steps on the alternate minimization of the cost function in Eq. (3.28) w.r.t. $q(\mathbf{S})$ and $q(\boldsymbol{\theta})$, respectively.

Using conjugate priors in VB models allows for the optimal update of the marginal approximations (e.g., $q(\mathbf{S})$ or $q(\boldsymbol{\theta})$) on each iteration (Attias, 2000b; Beal and Ghahramani, 2003). The cost function is then *minimized* on each step, which corresponds to *full* steps in the EM terminology. For example, Beal and Ghahramani (2003) present a variational Bayesian EM algorithm based on a family of conjugate-exponential models. The alternate update of $q(\mathbf{S})$ and $q(\boldsymbol{\theta})$ is simple there and the algorithm reduces to the full-step EM algorithm if the density $q(\boldsymbol{\theta})$ is restricted to point estimates.

Variational approximations have also been used for some LVMs learned by the EM-algorithm in which the full E-step is not tractable. There, the optimal posterior in Eq. (3.25) is approximated by minimizing the same type of cost function (see, e.g., Frey and Hinton, 1999; Attias, 1999; Ghahramani and Hinton, 2000).

Approximation of the posterior distribution is also done in *online Bayesian learning* (Opper, 1998) or *assumed-density filtering* (ADF) (Maybeck, 1982) as it is called in the control literature. This method considers the problem of updating the posterior distribution $p(\boldsymbol{\theta} \,|\, \mathbf{x}(1), \ldots, \mathbf{x}(t))$ after obtaining new measurements. For each new measurement $\mathbf{x}(t)$, the posterior is approximated by a convenient parametric distribution $q^{(t)}(\boldsymbol{\theta})$ by minimizing the KL divergence $D(\tilde{p} \,||\, q)$, where the new posterior $\tilde{p}$ is calculated using the previously found approximation:

$$\tilde{p}(\boldsymbol{\theta}) \propto p(\mathbf{x}(t) \,|\, \boldsymbol{\theta}) q^{(t-1)}(\boldsymbol{\theta}) \,. \tag{3.31}$$

The *expectation-propagation* (EP) method (Minka, 2001) modifies the basic ADF procedure such that the results become less dependent on the order in which the measurements are processed. Note that the ADF/EP approximation typically overestimates the posterior uncertainty as the form $D(p||q)$ of the minimized KL-divergence is different from the form $D(q \,||\, p)$ used in VB methods (see Fig. 3.2). The EP approach provides a better global approximation and therefore more accurate moments.

The EP and VB approximations are suited well to different problems. The VB approximation is more appropriate for parameter estimation (e.g., estimating parameters of an MLP network) where the posterior pdfs are often complex and severely multimodal. The EP approach would underfit hopelessly in this problem. However, the EP approximation can be better for state estimation (tracking the

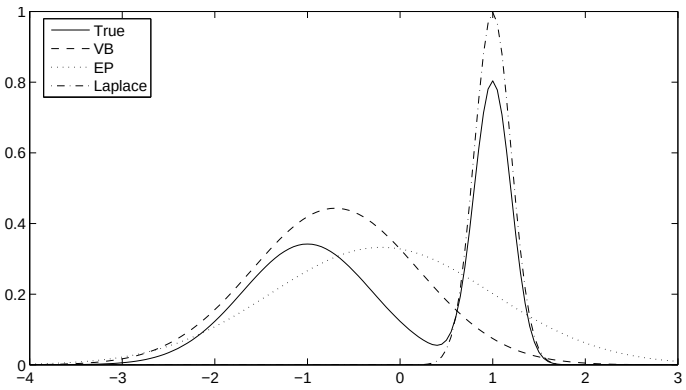


Figure 3.2: Approximating a hypothetical bimodal posterior by a Gaussian distribution using variational Bayesian methods (VB), expectation propagation (EP) and the Laplace approximation. The Laplace approximation is scaled by 0.5 for better presentation.

state of a dynamical system) as it can track several posterior modes, while the VB approach would track only one of the modes.

There are also variational approaches for approximating complex distributions which are not based on minimizing the KL-divergence. Jaakkola and Jordan (2000) use a family of adjustable bounds for the likelihood, which yields a tractable expression for the approximate posterior. The bounds are adjusted on each iteration in order to obtain the most accurate approximation around the points of interest. A similar approach was used by Girolami (2001) to derive an approximation based on a lower bound for the Laplacian prior in the problem of learning an overcomplete basis from a linear mixing model.

3.2.5 Basic LVMs with variational approximations

VB learning has been applied to various latent variable models reviewed in Section 2.1. The Gaussian model with linear mixing was considered by Bishop (1999b) in the technique called variational PCA. The model containing a mixture of linear factor analyzers was introduced by Ghahramani and Beal (2000). Valpola and Honkela extended the factor analysis model to the case of nonlinear mixing (Lappalainen and Honkela, 2000; Valpola et al., 2003b,a). The case with missing data was considered by Raiko et al. (2003).

A state-space model with switching between several linear regimes was considered by Ghahramani and Hinton (2000). Learning the standard state-space model using the VB principles was considered by Beal (2003) in the linear case,

and by Valpola and Karhunen (2002) for the more general nonlinear state-space models. A model with nonlinear state dynamics but with a linear mapping from the states to the observations was developed by Särelä et al. (2001).

Several researches have applied the VB principles to the ICA problem. Attias (1999) presented a model called *independent factor analysis* (IFA) in which the sources are described by mixtures of Gaussians. Later, Attias (2000a) extended the IFA model by taking into account the temporal statistical characteristics of the factors. The linear ICA model was considered by Valpola (Lappalainen, 1999), Attias (2000b), Miskin and MacKay (2001), and by Choudrey and Roberts (2001). The case of positive components was considered by Miskin and MacKay (2000), and later by Harva and Kabán (2005). Extensions with cluster ICA models were introduced by Chan et al. (2002) and by Choudrey and Roberts (2003). ICA problems with missing data were considered by Chan et al. (2003). A nonlinear source separation model based on the independence assumption was addressed by Valpola (2000) in the NIFA model.

Application of VB learning to other types of models have been considered, for example, by Hinton and van Camp (1993), Barber and Bishop (1998), Ghahramani and Hinton (2000).

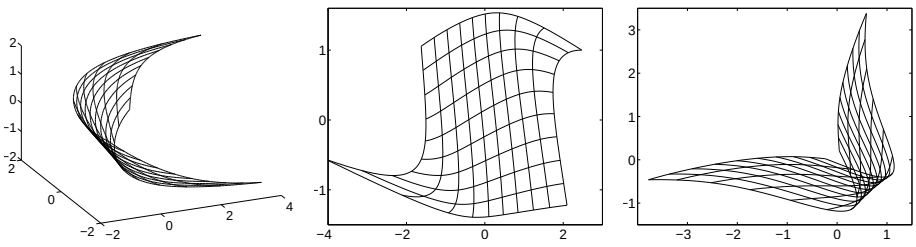
3.3 Post-nonlinear factor analysis

This section presents a latent variable model called *post-nonlinear factor analysis* (PNFA) which is learned by using the variational Bayesian approach. The motivation for the PNFA model is given based on the experiments reported in **Publication 3**. After that, the model structure is specified and the optimization algorithm is briefly described. Finally, the experimental results are presented. This section is largely based on **Publication 4** of this thesis.

3.3.1 Motivation

Publication 3 presents experimental comparison of two approaches to the nonlinear BSS problem: the NFA+FastICA approach based on the model developed by Valpola and Honkela (Lappalainen and Honkela, 2000) and Taleb and Jutten's (TJ) algorithm for post-nonlinear mixtures (Taleb and Jutten, 1999b, see also the general introduction of the two algorithms in Section 2.2.6). The comparison is performed on artificial test problems containing PNL mixtures, for which both algorithms are applicable. Both the classical case when the number M of the sources is equal to the number N of the observations and the case of overdetermined mixtures (when $M < N$) are considered.

A new interesting result of the experiments is that globally invertible PNL mixtures, but with non-invertible component-wise nonlinearities, can be iden-



(a) (b) (c)

Figure 3.3: (a): A two-dimensional manifold defined by a post-nonlinear mapping with one non-invertible post-nonlinearity. (b): The representation of the manifold in the source space estimated by NFA+FastICA. (c): The representation of the manifold in the source space estimated by Taleb and Jutten's algorithm.

tified and sources can be separated, extending the earlier results of Taleb and Jutten (1999b). In Publication 3, we explain this result using the following simple example of a three-dimensional PNL mixture of two sources. The sources are transformed using a PNL mapping of Eq. (2.26) with one non-invertible post-nonlinear distortion:

$$f_1(y) = y^2\,, \qquad f_2(y) = \tanh(y)\,, \qquad f_3(y) = \tanh(y)\,. \tag{3.32}$$

After the PNL transformation, the data lie on a two-dimensional manifold embedded in the three-dimensional space. If the sources are described in the original source space using an even grid, this manifold can be visualized by the transformed source grid, as shown in Fig. 3.3a. The PNL transformation is invertible as there exists a bijection from the two-dimensional source space to the data manifold in the three-dimensional observation space.

A nonlinear data representation modeled by an invertible generative mapping can be learned using the NFA algorithm. Fig. 3.3b shows the representation of the original source grid in the source space estimated by the NFA+FastICA approach. If $\mathbf{s}$ and $\hat{\mathbf{s}}$ are the original and the estimated sources, respectively, the algorithm implicitly estimates the mapping $\boldsymbol{\xi}$ such that $\hat{\mathbf{s}}(t) = \boldsymbol{\xi}(\mathbf{s}(t))\,,\ t = 1, \ldots, T$. The plot in Fig. 3.3b is the reconstruction of the even source grid using the mapping $\boldsymbol{\xi}$ explicitly estimated for this demonstration. As follows from the figure, the reconstruction of the original sources obtained using the Bayesian algorithm is pretty good.

Fig. 3.3c presents the same plot for the TJ algorithm. It shows that the TJ algorithm cannot achieve reconstruction of the sources. This happens due to its constrained structure as it estimates the inverse of the PNL transformation under the assumption that all the post-nonlinear distortions f_i are invertible. As

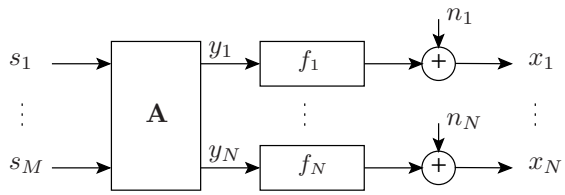


Figure 3.4: The structure of the PNFA model.

a result, it cannot unfold the curved data manifold.

The demonstrated result shows the relevance of exploiting more observations than sources and the relevance of learning a generative mapping instead of inverting the mixing transformation. This can be done by applying the Bayesian approach to the model in Eq. (2.2) with the restriction that the generative mapping $\mathbf{f}$ has the post-nonlinear structure. Combined with the Gaussian model for the sources $\mathbf{s}$, this yields the model that we call PNFA. Its structure is presented in Fig. 3.4.

The post-nonlinear ICA problem can be solved using PNFA in two steps similarly to the NFA+FastICA approach. First, the PNFA model is learned to find underlying *Gaussian* factors. After that, the factors found by PNFA are rotated using a linear ICA algorithm which is chosen to be FastICA in the experiments. These two steps are termed the *PNFA+FastICA* approach.

3.3.2 Density model

This section describes the density model $p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta})$ used for PNFA. The latent variables are introduced first. The sources s_j are assumed to be zero-mean Gaussian variables and the corresponding prior is

$$p(\mathbf{S} \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{M} \prod_{t=1}^{T} N\left(s_j(t) \,|\, 0,\, v_{s,j} \right) . \tag{3.33}$$

Variable $v_{s,j}$ is the variance parameter defining the prior distribution for the j-th source. Parameters defining priors for other variables are often called hyperparameters. The hyperparameters $v_{s,j}$ are assigned log-normal priors, making the source prior model *hierarchical*.

The variances of the source distributions are assumed different for individual sources to enable the *automatic relevance determination*, when irrelevant sources have posterior variances close to zero. This allows for automatic determination of the appropriate dimensionality of the latent space and avoids discrete model

selection (Bishop, 1999b). The idea of relevant input variable selection was first used by MacKay and Neal in the context of neural networks (Neal, 1998).

The observation model expresses the PNL structure of the generative mapping:

$$p(\mathbf{X} \mid \mathbf{S}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{t=1}^{T} N\left(x_i(t) \mid f_{i,t},\, v_{x,i} \right) , \tag{3.34}$$

where

$$f_{i,t} = f_i(y_i(t), \boldsymbol{\theta}_{f,i}) , \tag{3.35}$$

$$y_i(t) = \sum_{j=1}^{M} a_{ij} s_j(t) , \tag{3.36}$$

and $\boldsymbol{\theta}_{f,i}$ denotes the parameters of the post-nonlinearities f_i. The post-nonlinear distortions are modeled by multi-layer perceptron (MLP) networks with one hidden layer:

$$f_i(y, \boldsymbol{\theta}_{f,i}) = \mathbf{d}_{1,i}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{c}_{1,i}\, y + \mathbf{c}_{2,i}) + d_{2,i} . \tag{3.37}$$

and thus the parameters $\boldsymbol{\theta}_{f,i}$ include vectors $\mathbf{c}_{1,i}$, $\mathbf{c}_{2,i}$, $\mathbf{d}_{1,i}$ and a scalar $d_{2,i}$. A sigmoidal activation function $\boldsymbol{\phi}$ operates component-wise on its inputs.

The prior distributions for the parameters modeling the generative mapping are chosen as follows. The linear mixing part $\mathbf{A}$ containing the linear coefficients a_{ij} in Eq. (3.36) has a fixed Gaussian prior

$$p(\mathbf{A}) = \prod_{i,j} N\left(a_{ij} \mid 0,\, 1 \right) . \tag{3.38}$$

The variance of the weights are fixed to a constant because the scale of the weights can be defined by the changing variances $v_{s,j}$ of the sources (Lappalainen and Honkela, 2000). The nonlinearities in Eq. (3.37) are regularized by using zero mean Gaussian priors for the weights $\mathbf{c}_{1,i}$ and $\mathbf{d}_{1,i}$. Hierarchical Gaussian priors are also assigned to parameters $\mathbf{c}_{2,i}$, $d_{2,i}$ and the noise variance parameters $v_{x,i}$.

Thus, the overall pdf $p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta})$ has a simple factorial form

$$p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta}) = p(\mathbf{X} \mid \mathbf{S}, \boldsymbol{\theta}) p(\mathbf{S} \mid \boldsymbol{\theta}) \prod_{k} N\left(\theta_k \mid \theta_{k,\mathrm{m}},\, \theta_{k,\mathrm{v}} \right) \tag{3.39}$$

where the first two factors are defined in Eqs. (3.34) and (3.33), respectively, and $\theta_{k,\mathrm{m}}$, $\theta_{k,\mathrm{v}}$ denote the mean and variance parameters of the prior for a parameter or a hyperparameter θ_k. The parameters $\boldsymbol{\theta}$ include variables $\mathbf{A}$, $\mathbf{c}_{1,i}$, $\mathbf{c}_{2,i}$, $\mathbf{d}_{1,i}$, $d_{2,i}$ and various hyperparameters such as $\log v_{s,j}$ and $\log v_{x,i}$.

3.3.3 Optimization of the cost function

Learning the PNFA model is done using the variational Bayesian principles explained in Section 3.2.3. The posterior of the unknown parameters $\mathbf{S}$, $\boldsymbol{\theta}$ is approximated using a fully factorial distribution

$$q(\mathbf{S}, \boldsymbol{\theta}) = \prod_{j,t} q(s_j(t)) \prod_k q(\theta_k) \,, \tag{3.40}$$

where each individual factor $q(\theta)$ is a Gaussian distribution parameterized with the mean $\overline{\theta}$ and the variance $\widetilde{\theta}$. Such parameters $\overline{\theta}$ and $\widetilde{\theta}$ are called *variational* parameters. The approximation in Eq. (3.40) is fitted to the true posterior by minimizing the cost function in Eq. (3.28):

$$\mathcal{C}(q) = \left\langle \log \frac{q(\mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta})} \right\rangle = \langle \log q(\mathbf{S}, \boldsymbol{\theta}) \rangle - \langle \log p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta}) \rangle \,. \tag{3.41}$$

Due to the factorial structures of $q(\mathbf{S}, \boldsymbol{\theta})$ and $p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta})$, the cost function splits into a sum of simple terms:

$$\mathcal{C}(q) = \sum_{j,t} \langle \log q(s_j(t)) \rangle + \sum_k \langle \log q(\theta_k) \rangle \tag{3.42}$$

$$- \sum_{i,t} \langle \log N\left(x_i(t) \mid f_{i,t},\, v_{x,i} \right) \rangle \tag{3.43}$$

$$- \sum_{j,t} \langle \log N\left(s_j(t) \mid 0,\, v_{s,j} \right) \rangle - \sum_k \langle \log N\left(\theta_k \mid \theta_{k,\mathrm{m}},\, \theta_{k,\mathrm{v}} \right) \rangle \,, \tag{3.44}$$

where $\langle \cdot \rangle$ denotes the expectation over distribution $q(\mathbf{S}, \boldsymbol{\theta})$.

During learning, individual factors $q(s_j(t))$, $q(\theta_k)$ of the approximation in Eq. (3.40) are updated one at a time while keeping the others fixed. For each update of one factor, only the terms containing the corresponding variable are relevant. For example, for updating $q(\theta_k)$, the part of the cost function to be minimized is

$$\mathcal{C}_k = \langle \log q(\theta_k) \rangle - \sum_l \langle \log p(\theta_l \,|\, \theta_k) \rangle - \langle \log N\left(\theta_k \mid \theta_{k,\mathrm{m}},\, \theta_{k,\mathrm{v}} \right) \rangle \,, \tag{3.45}$$

where θ_l are all the variables whose distribution is conditioned on θ_k. Since each factor $q(\theta_k)$ is a univariate Gaussian distribution, one has to minimize the quantity in Eq. (3.45) w.r.t. the variational parameters $\overline{\theta}_k$ and $\tilde{\theta}_k$.

For the variables θ_k that do not contribute to the evaluation of the outputs $f_{i,t}$ (and therefore do not affect the likelihood terms in Eq. (3.43)), the cost terms in Eq. (3.45) and the gradients $\partial \mathcal{C}_k / \partial \bar{\theta}_k$ $\partial \mathcal{C}_k / \partial \tilde{\theta}_k$ can be evaluated exactly. Then,

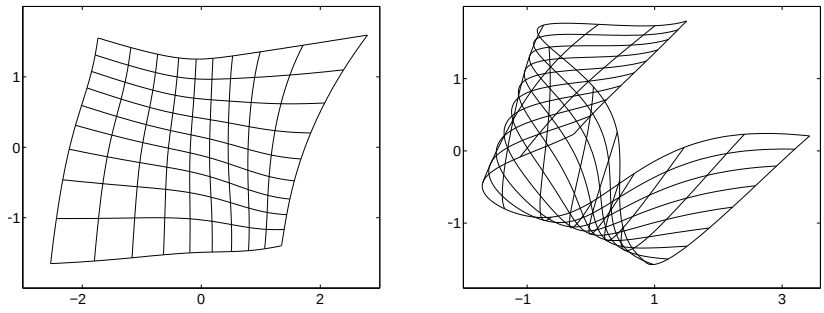


Figure 3.5: Experimental results for a test problem with a three-dimensional PNL mixture of two sources; two out of three post-nonlinearities are non-invertible. The plots show the representation of the even source grid found by PNFA+FastICA (left) and by Taleb and Jutten's algorithm (right).

a numerical optimization algorithm derived for the NFA model (Lappalainen and Honkela, 2000) can be used.

The difficulties arise when updating the posterior for the variables that contribute to the evaluation of $f_{i,t}$, because the likelihood terms in Eq. (3.43) and the corresponding gradients cannot be evaluated exactly. In **Publication 4**, it is shown how the likelihood terms depend on the means and variances of the outputs $f_{i,t}$ and it is explained how those means and variances (and therefore the cost function) can be calculated using first-order Taylor approximation and Gauss-Hermite quadrature. Using this approximation, the gradients of the likelihood terms can be propagated from the outputs $f_{i,t}$ to the rest of the parameters using a scheme resembling backpropagation (see, e.g., Haykin, 1999).

Since the cost function and its gradients can be computed, it is possible to do the minimization numerically. For many parameters, the resulting minimization procedure is similar to the gradient-based algorithm used in NFA (see Lappalainen and Honkela, 2000).

3.3.4 Experimental example

In **Publication 4**, we test the proposed PNFA algorithm on an artificial example of a three-dimensional PNL mixture of two independent sources. The PNL mappings is chosen such that it is globally invertible but contains two non-invertible post-nonlinear distortions. The mixtures are noisy, that is white Gaussian noise is added to the data after mixing. In the experiment, we use the PNFA+FastICA approach to find independent components underlying the test data.

The left plot in Fig. 3.5 is the representation of the original source grid in

the source space estimated by PNFA+FastICA (the interpretation of the plots is same as in Fig. 3.3). The results indicate that the source space and the PNL mapping are estimated pretty well. The achieved quality of the original source reconstruction is moderate but note that the classical PNL algorithms cannot achieve comparable quality (see the results of the TJ algorithm in the right plot of Fig. 3.5).

One of the probable reasons for the moderate quality of the source estimation is the coarse model for the sources, which is chosen to be Gaussian. A better approach might be to use a mixture model, as in the independent factor analysis model developed by Attias (1999). In order to obtain good separation quality, such an approach would most probably require a more complex posterior approximation because the fully factorial posterior approximation used in the presented PNFA algorithm is too simple to capture any posterior correlations in the vicinity of the correct BSS solution. The effect of the form of the posterior approximation is explained in more details in the following section.

3.4 Effect of posterior approximation

The computational complexity of the algorithms implementing the variational Bayesian principles significantly depends on the chosen form of the posterior approximation. In addition to the most commonly used factorization $q(\mathbf{S}, \boldsymbol{\theta}) = q(\mathbf{S})q(\boldsymbol{\theta})$, the source and parameter posterior approximations are typically factorized further. For example, the parameters can be divided into subsets

$$q(\boldsymbol{\theta}) = \prod_i q(\boldsymbol{\theta}_i)\,, \tag{3.46}$$

and each term $q(\boldsymbol{\theta}_i)$ captures the correlations between the variables in the set $\boldsymbol{\theta}_i$ while all posterior correlations with the variables in other sets $\boldsymbol{\theta}_j$ are neglected. The extreme case is the fully factorial approximation such as the one in Eq. (3.40) used in the PNFA algorithm.

Although assuming suitably factorial q usually results in computationally efficient learning algorithms, we show in **Publication 1** that the form of the posterior approximation can affect the solution found by VB methods. Two common cases are investigated in detail:

1. sources are approximated to be independent a posteriori

$$q(\mathbf{S}) = \prod_{j,t} q(s_j(t))\,; \tag{3.47}$$

2. the posterior correlations of the sources are modeled

$$q(\mathbf{S}) = \prod_{t=1}^{T} q(\mathbf{s}(t)) \,. \tag{3.48}$$

This effect is studied in Publication 1 both theoretically and experimentally by considering the source separation problem in linear mixtures

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \,, \tag{3.49}$$

when the sources are assumed to have either decoupled dynamics or non-Gaussian distributions. The analysis, however, extends to the case of nonlinear mixtures as well.

It is shown that neglecting the posterior correlations of the sources in Eq. (3.47) introduces a bias in favor of the PCA solution. By the PCA solution we mean the solution in which the mixing vectors, columns of mixing matrix $\mathbf{A}$, are orthogonal w.r.t. the inverse of the estimated noise covariance $\mathbf{\Sigma_n} = \mathrm{E}\{\mathbf{n}\mathbf{n}^\mathrm{T}\}$, that is $\mathbf{A}^\mathrm{T}\mathbf{\Sigma_n}^{-1}\mathbf{A}$ is a diagonal matrix. This effect can be unimportant in many latent variable models introduced in Section 2.1 where individual sources may not have meaningful interpretations. However, this matter is crucial for the source separation models discussed in Section 2.2.

3.4.1 Trade-off between posterior mass and posterior misfit

In variational methods, there is a general trade-off between the amount of posterior mass in the neighborhood of the solution and the misfit between the approximation and the true local probability distributions. This effect can be shown to exist for both Bayesian methods which are discussed in this thesis and also for ML methods which use variational approximations (e.g., Attias, 1999; Ghahramani and Hinton, 2000).

In general, Bayesian methods aim to find a solution which corresponds to a model whose neighborhood contains a large portion of the posterior probability mass. This implies that *the posterior density of the unknown parameters is high.* For linear models described by Eq. (3.49), this is achieved if

1. the sources and the mixing matrix together explain the observations well;

2. the source estimates fit their prior model.

Large posterior mass also implies that the solution corresponds to a *wide peak* in the posterior density, which means that

3. the solution is robust.

As was discussed in Section 3.2.3, VB learning is able to find a solution which meets these three requirements.

However, the restricted form of the posterior approximation results in an additional requirement:

4. the form of the posterior approximation $q(\mathbf{S}, \boldsymbol{\theta}) = q(\mathbf{S})q(\boldsymbol{\theta})$ should match the posterior $p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})$ around the solution.

In practice, the choice of the functional form of $q(\mathbf{S})$ may affect the optimal solution significantly, while the effect of the form of $q(\mathbf{A})$ is smaller. For the rest of the parameters, this effect is usually negligible as their number is typically much smaller than the number of unknown quantities in $\mathbf{A}$ and especially in $\mathbf{S}$.

The solution found by variational methods is usually a compromise between the amount of posterior mass (requirements 1–3) and the misfit between the approximation and the true local posterior (requirement 4). Usually it is desirable that the requirement 4 affects the solution as little as possible although sometimes it is possible to use it to select an appropriate solution among otherwise degenerate solutions (see Section 3.5 for an example of such regularization).

In the following, the trade-off between the misfit of the posterior approximation and the accuracy of the model is explained using a hypothetical example. Let us assume that the data are described well by a probabilistic LVM with the joint pdf $p(x, s, \theta)$ in the solution $s = s_{\text{true}}$ and $\theta = \theta_{\text{true}}$. Then, the joint posterior $p(s, \theta \,|\, x)$ has a peak in the vicinity of the correct solution $(\theta_{\text{true}}, s_{\text{true}})$ and its fragment could look like the one presented in Fig. 3.6. Note that there are typically correlations between the hidden variables and the other parameters. For example, in the linear model in Eq. (3.49), these correlations reflect the fact that rotating $\mathbf{A}$ could be compensated by rotating $\mathbf{s}$ correspondingly. These correlations are typically neglected in the posterior approximation.

Let us assume that the variational principles are used to approximate the posterior using a point estimate for θ and a *Gaussian* distribution $q(s)$ for the variable s. Then, VB learning reduces to the EM algorithm which uses a variational approximation for the posterior $p(s \,|\, X, \theta)$. Examples of this posterior are shown in Fig. 3.6 with the bold curves for two values of θ.

The peak in the posterior means that the cost of inaccurate modeling is minimized in the correct solution $(\theta_{\text{true}}, s_{\text{true}})$ where the model is most accurate. However, the posterior $p(s \,|\, X, \theta)$ is closest to Gaussian in the vicinity of another solution which we denote by (θ_q, s_q). There, the true posterior $p(s \,|\, X, \theta)$ can be approximated best by $q(s)$ and therefore the misfit between the optimal posterior and its approximation is minimized. The actual solution found by variational methods will generally be a compromise between these two solutions.

The presented example is rather illustrative as the mismatch between the true local posterior and its *Gaussian* approximation is more important in nonlinear

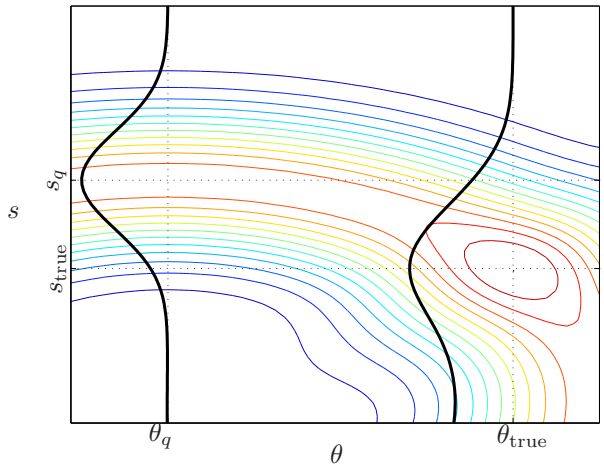


Figure 3.6: A hypothetical posterior $p(s, \theta \,|\, x)$. The data are explained best in the solution $(\theta_{\text{true}}, s_{\text{true}})$ where the posterior has a peak. The bold black curves represent the posterior pdfs $p(s \,|\, x, \theta_q)$ and $p(s \,|\, x, \theta_{\text{true}})$. The form of the posterior for s is closer to Gaussian in the solution (θ_q, s_q).

models (e.g., Valpola and Karhunen, 2002). The Gaussian form of the posterior approximation typically introduces a bias in favor of smooth mappings. For linear ICA models in Eq. (3.49), the more important factor is that the posterior approximation $q(\mathbf{S})$ often neglects the posterior correlations between the sources. As we show in **Publication 1**, this introduces a bias in favor of the PCA solution. Therefore, the found solution is a result of a trade-off between the ICA solution where the explanation of the sources is best and the PCA solution where the posterior approximation of the sources is most accurate. If the mixing vectors are close to orthogonal and the source model is strongly in favor of the ICA solution, the optimal solution can be expected to be close to the ICA solution. If the mixing matrix cannot be made more orthogonal (e.g., by pre-whitening), it is possible to end up close to the PCA solution even though the model should be able to judge the ICA solution to be better.

3.4.2 Factorial $q(\mathbf{S})$ favors orthogonality

The fully factorial approximation in Eq. (3.47) is often used in Bayesian ICA models. However, **Publication 1** shows that it favors solutions with an orthogonal mixing matrix, which is a characteristic of PCA.

Publication 1 considers three cases of linear models with different source mod-

els: temporally correlated sources, super-Gaussian sources and sources described with a mixture model. In the following, the form of the optimal unrestricted Gaussian approximation $q(\mathbf{s}(t))$ is presented for the three models.

Temporally correlated sources

In the simplest case, the temporal correlations in the sources can be modeled using a linear first-order autoregressive process with Gaussian innovations:

$$p(\mathbf{s}(t) \,|\, \mathbf{s}(t-1), \boldsymbol{\theta}) = N\left(\, \mathbf{s}(t) \,|\, \mathbf{B}\mathbf{s}(t-1),\, \boldsymbol{\Sigma}_{\mathbf{m}} \,\right) . \tag{3.50}$$

The matrix of dynamics $\mathbf{B}$ and the covariance matrix of innovations $\boldsymbol{\Sigma}_{\mathbf{m}}$ are assumed diagonal due to the independence of the sources.

It can be shown that the optimal unrestricted posterior $q(\mathbf{s}(t))$ for this model is a Gaussian distribution whose covariance for $t = 2, \ldots, T-1$ is given by

$$\boldsymbol{\Sigma}_{\mathbf{s},\mathrm{opt}} = \left\langle \mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{n}}^{-1} \mathbf{A} + \boldsymbol{\Sigma}_{\mathbf{m}}^{-1} + \mathbf{B}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{m}}^{-1} \mathbf{B} \right\rangle^{-1} , \tag{3.51}$$

where $\boldsymbol{\Sigma}_{\mathbf{n}}$ is the diagonal covariance matrix of the observation noise.

Super-Gaussian sources

If the sources are known to be super-Gaussian (i.e. their kurtosis is positive), each source can be modeled as a Gaussian variable whose variance changes with time. Then, the source prior model is

$$p(\mathbf{s}(t) \,|\, \boldsymbol{\theta}) = N\left(\, \mathbf{s}(t) \,|\, \mathbf{0},\, \boldsymbol{\Sigma}_{\mathbf{s}}(t) \,\right) \tag{3.52}$$

where $\boldsymbol{\Sigma}_{\mathbf{s}}(t)$ is the time-dependent diagonal covariance matrix. The diagonal elements of $\boldsymbol{\Sigma}_{\mathbf{s}}(t)$ are the variances of individual sources at different time instances, they are modeled in Publication 1 using log-normal parameterization.

The optimal unrestricted posterior $q(\mathbf{s}(t))$ is Gaussian for this model and its covariance matrix is

$$\boldsymbol{\Sigma}_{\mathbf{s}(t),\mathrm{opt}} = \left\langle \mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{n}}^{-1} \mathbf{A} + \boldsymbol{\Sigma}_{\mathbf{s}}^{-1}(t) \right\rangle^{-1} . \tag{3.53}$$

Mixture-of-Gaussians model

The source prior that is most commonly used in Bayesian ICA models is the mixture-of-Gaussians (MoG). The distribution of each source s_j is modeled by a mixture of K_j Gaussian components

$$p(s_j(t) \,|\, \boldsymbol{\theta}) = \sum_{k=1}^{K_j} \pi_{j,k} N\left(\, s_j(t) \,|\, m_{j,k},\, v_{j,k} \,\right) \tag{3.54}$$

and therefore the prior for the source vector $\mathbf{s}(t)$ is a mixture of $\prod_j K_j$ Gaussian components, each having a diagonal covariance matrix:

$$p(\mathbf{s}(t) \,|\, \boldsymbol{\theta}) = \prod_j p(s_j(t) \,|\, \boldsymbol{\theta}) = \sum_{\boldsymbol{\nu}} \pi_{\boldsymbol{\nu}} N\left(\, \mathbf{s}(t) \,|\, \boldsymbol{\mu}_{\boldsymbol{\nu}},\, \boldsymbol{\Sigma}_{\mathbf{s},\boldsymbol{\nu}} \,\right) . \tag{3.55}$$

Here, $\boldsymbol{\nu}$ is a vector whose j-th component $\nu_j \in \{1, \ldots, K_j\}$ defines the mixture component chosen for source s_j. The sum $\sum_{\boldsymbol{\nu}}$ means $\sum_{\nu_1=1}^{K_1} \cdots \sum_{\nu_M=1}^{K_M}$, $\pi_{\boldsymbol{\nu}} = \prod_j \pi_{j,\nu_j}$ denotes the prior probability that $\mathbf{s}(t)$ is drawn from the mixture component defined by $\boldsymbol{\nu}$, and $\boldsymbol{\Sigma}_{\mathbf{s},\boldsymbol{\nu}}$ are the diagonal covariance matrices of the mixture components.

The optimal unrestricted posterior $q(\mathbf{s}(t))$ for this model would be a mixture of Gaussians with $\prod_j K_j$ mixture components. The estimation of such posterior becomes computationally intractable in high dimensions and therefore a simpler approximation by only one Gaussian is sometimes used (Miskin and MacKay, 2001). The covariance matrix of this Gaussian approximation is given by

$$\Sigma_{\mathbf{s}(t),\mathrm{opt}} = \left\langle \mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{n}}^{-1} \mathbf{A} + \mathbf{D}(t) \right\rangle^{-1} \tag{3.56}$$

where $\mathbf{D}(t)$ is a diagonal matrix with the elements $d_j(t) = \sum_{k=1}^{K_j} \lambda_{tjk} v_{j,k}^{-1}$ on the main diagonal. The coefficients λ_{tjk} estimate the posterior probability that a sample $s_j(t)$ is drawn from the k-th mixture component $N\left(\, s_j(t) \,|\, m_{j,k},\, v_{j,k} \,\right)$.

The misfit between the factorial approximation in Eq. (3.47) and the optimal unrestricted $q(\mathbf{s}(t))$ is minimized when the form of the optimal $q(\mathbf{s}(t))$ agrees with Eq. (3.47). This is the case when the optimal covariance matrices given in Eqs. (3.51), (3.53), (3.56) are diagonal. This, in turn, happens if and only if the columns of $\mathbf{A}$ are orthogonal w.r.t. the inverse noise covariance $\boldsymbol{\Sigma}_{\mathbf{n}}^{-1}$. Since VB learning is trying to minimize the misfit, it favors orthogonal solutions for $\mathbf{A}$. A similar effect can be shown to exist for the posterior approximation of the mixing matrix when the fully factorial approximation favors uncorrelated sources.

The experiments reported in **Publication 1** confirm these theoretical results. Some results for a model with non-Gaussian sources are reproduced in Fig. 3.7. When the source distributions are close to Gaussian (experiment (a)), the PCA solution is found even after initialization in the correct solution. In experiment (c), the ICA solution is found because the sources are strongly non-Gaussian. Some other solution is obtained in the mediate case (experiment (b)).

Similar results were reported by other researchers. Højen-Sørensen et al. (2002) argue that posterior correlations should be taken into account in the application of variational methods to the ICA problem. Wang and Titterington (2004) consider a similar problem in which the parameters of a linear state-space model in Eqs. (2.8)–(2.9) are estimated using the variational Bayesian approach

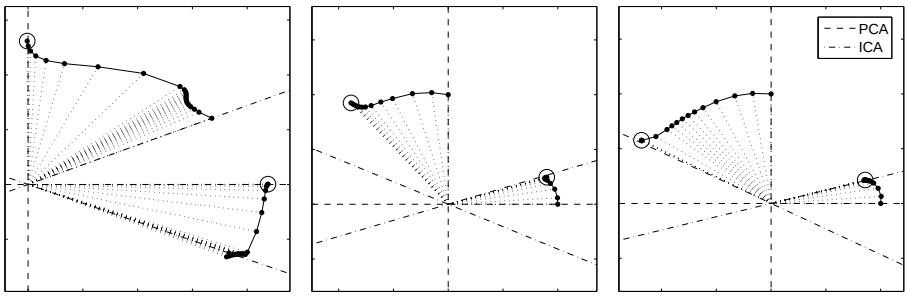


Figure 3.7: The two columns of the mixing matrix during learning an ICA model with two non-Gaussian sources. The sources are modeled by mixtures of Gaussians, and the factorial $q(\mathbf{s}(t))$ is used. The final solutions are circled. The degree of non-Gaussianity of the mixed signals grows from (a) to (c).

with a fully factorial approximation. In particular, they show that the estimate of the matrix $\mathbf{B}$ in Eq. (2.9) tends to the true value of $\mathbf{B}$ only for $\mathbf{B} = \mathbf{0}$. Thus, the fully factorial approximation introduces a bias in favor of a static factor analysis model.

3.5 Nonlinear state-space models

The effect of the posterior approximation is to introduce a bias in favor of certain types of solutions. This can be a negative result as was shown in the previous section. However, sometimes it is possible to use this effect to select an appropriate solution among otherwise degenerate solutions.

This section considers a *nonlinear dynamic factor analysis* (NDFA) method introduced by Valpola and Karhunen (2002) for estimation of nonlinear state-space models (see Section 2.1.3) using variational Bayesian learning. First, the modeling assumptions are briefly introduced. It is also shown how the method can achieve a meaningful representation of the sources by using a suitable posterior approximation. Then, it is demonstrated how the NDFA algorithm can be used for the problem of detecting changes in the dynamics of an observed dynamical system.

3.5.1 Nonlinear dynamic factor analysis

NDFA considers the classical nonlinear state-space model

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{n}(t) \tag{3.57}$$

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1)) + \mathbf{m}(t)\,, \tag{3.58}$$

in which the states $\mathbf{s}(t)$ and the noise terms $\mathbf{n}(t)$, $\mathbf{m}(t)$ are described by Gaussian distributions. All the structural assumptions of NDFA are expressed in the form of the density model $p(\mathbf{X}, \mathbf{S}, \boldsymbol{\theta})$. The observation equation (3.57) is expressed in the likelihood factor and the state equation (3.58) defines the source prior. The unknown nonlinear mappings $\mathbf{f}$ and $\mathbf{g}$ are modeled by MLP networks with one hidden layer of sigmoidal tanh nonlinearities. Gaussian distributions are used to describe the weights of the MLPs for computational tractability.

The posterior distribution of the unknown parts of the model is learned using the variational Bayesian principles. The posterior approximation $q(\boldsymbol{\theta}, \mathbf{S}) = q(\boldsymbol{\theta})q(\mathbf{S})$ is chosen to be fully factorial Gaussian for $q(\boldsymbol{\theta})$, but the posterior $q(\mathbf{S})$ is somewhat more complex. It takes into account the posterior dependences between the state values at successive time instants in order to avoid the problem described by Wang and Titterington (2004):

$$q(\mathbf{S}) = \prod_j \left[q(s_j(1)) \prod_{t=2}^{T} q(s_j(t) \,|\, s_j(t-1)) \right] . \tag{3.59}$$

The conditional distribution in Eq. (3.59) is assumed Gaussian

$$q(s_j(t) \,|\, s_j(t-1)) = N\left(s_j(t) \,|\, \mu_j(t),\, \widetilde{s}_j(t) \right) \tag{3.60}$$

with the mean $\mu_j(t)$ that depends linearly on the previous value $s_j(t-1)$:

$$\mu_j(t) = \overline{s}_j(t) + \rho_{j,(t-1),t}(s_j(t-1) - \overline{s}_j(t-1))\,. \tag{3.61}$$

A positive side-effect of the restrictions on the approximating distribution $q(\mathbf{S}, \boldsymbol{\theta})$ is that the nonlinear dynamical reconstruction problem is regularized and becomes well-posed. With linear $\mathbf{f}$ and $\mathbf{g}$, the true posterior distribution of the states $\mathbf{S}$ would be Gaussian, while nonlinear $\mathbf{f}$ and $\mathbf{g}$ result in a non-Gaussian posterior distribution. Restricting the approximation $q(\mathbf{S})$ to be Gaussian even in the nonlinear model therefore favors smooth mappings and regularizes the problem. The simpler Gaussian approximation $q(\mathbf{S}) = \prod_{t=1}^{T} q(\mathbf{s}(t))$ would still leave a rotational ambiguity within the source space, which would in practice yield a PCA-like solution. This is resolved by discouraging the posterior dependences between $s_j(t)$ and $s_l(t-1)$ with $j \neq l$.

NDFA favors decoupled dynamics of sources

It can be shown (see the Appendix at the end of this chapter) that the used parameterization of the posterior in Eq. (3.59) corresponds to modeling the posterior of all the source values

$$\begin{bmatrix}\mathbf{s}(1)^{\mathrm{T}} & \mathbf{s}(2)^{\mathrm{T}} & \mathbf{s}(3)^{\mathrm{T}} & \dots\end{bmatrix}^{\mathrm{T}} \tag{3.62}$$

with a Gaussian distribution whose covariance is parameterized as

$$\begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_{1,2} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{D}_{1,2} & \mathbf{D}_2 & \mathbf{D}_{2,3} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{D}_{2,3} & \mathbf{D}_3 & \mathbf{D}_{3,4} & \dots \\ \vdots & \vdots & \vdots & \ddots & \dots \end{bmatrix}^{-1}, \tag{3.63}$$

where $\mathbf{D}_t$ and $\mathbf{D}_{(t-1),t}$ are diagonal matrices made up from the elements $\widetilde{s}_j^{-1}(t) + \rho^2_{j,t,(t+1)}\widetilde{s}_j^{-1}(t+1)$ and $-\rho_{j,(t-1),t}\widetilde{s}_j^{-1}(t)$, respectively. There are also some exceptions for the last source values $\mathbf{s}(T)$.

Let us now assume for simplicity that the mappings $\mathbf{f}$ and $\mathbf{g}$ were restricted to be linear, that is the linear state-space model described by Eqs. (2.8)-(2.9) is learned. In this case, the optimal unrestricted posterior for the sources in Eq. (3.62) would be Gaussian with the covariance matrix

$$\begin{bmatrix} \widetilde{\mathbf{\Sigma}}_1 & -\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{m}}^{-1} & \mathbf{0} & \mathbf{0} & \dots \\ -\mathbf{\Sigma}_{\mathbf{m}}^{-1}\mathbf{B} & \widetilde{\mathbf{\Sigma}} & -\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{m}}^{-1} & \mathbf{0} & \dots \\ \mathbf{0} & -\mathbf{\Sigma}_{\mathbf{m}}^{-1}\mathbf{B} & \widetilde{\mathbf{\Sigma}} & -\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{m}}^{-1} & \dots \\ \vdots & \vdots & \vdots & \ddots & \dots \end{bmatrix}^{-1}, \tag{3.64}$$

where

$$\widetilde{\mathbf{\Sigma}}_1 = \mathbf{A}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{n}}^{-1}\mathbf{A} + \mathbf{\Sigma}_{\mathbf{s}_1}^{-1} + \mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{m}}^{-1}\mathbf{B} \tag{3.65}$$

$$\widetilde{\mathbf{\Sigma}} = \mathbf{A}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{n}}^{-1}\mathbf{A} + \mathbf{\Sigma}_{\mathbf{m}}^{-1} + \mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{m}}^{-1}\mathbf{B}\,, \tag{3.66}$$

with $\mathbf{\Sigma}_{\mathbf{s}_1}$ the prior covariance for the source values $\mathbf{s}(1)$. There are some exceptions in Eq. (3.64) for the last source values $\mathbf{s}(T)$. The misfit between the posterior approximation in Eq. (3.59) and the optimal unrestricted posterior is minimized when the covariance matrix in Eq. (3.63) agrees with the optimal structure in Eqs. (3.64)–(3.66). This is the case if and only if the columns of $\mathbf{A}$ are orthogonal w.r.t. $\mathbf{\Sigma}_{\mathbf{n}}^{-1}$ and the matrix of dynamics $\mathbf{B}$ is diagonal, that is the sources have independent dynamic models. This result can also be extended to the case of nonlinear mappings. Thus, the NDFA algorithm tries to find such a representation in which the dynamics of different sources are as decoupled as possible.

Subspace separation example

The experimental results reported by Valpola and Karhunen (2002) are reproduced here with the emphasis that the NDFA algorithm is able not only to learn a good dynamic model but also to find a meaningful source representation, that is the NDFA method can achieve nonlinear source separation.

The artificial dataset is produced by mixing in a nonlinear manner three independent dynamic processes, two of which are Lorenz processes and one is a harmonic oscillator. Only five linear projections of the eight states are used to produce the observations. Finally, the data are corrupted by observation noise. Five out of 10 observations are presented in Fig. 3.8.

As reported by Valpola and Karhunen (2002), the NDFA algorithm is able to learn a very good dynamic model for this artificial dataset. In addition to this, the dynamics of the sources is decoupled in such a way that three groups of sources reconstruct the three original dynamic processes (see Fig. 3.8). Within the three subspaces, the sources are estimated up to a nonlinear transformation but the subspaces are separated correctly. Note that the number of sources was set to 9 in the experiments but one of the sources was considered irrelevant by the algorithm and its values were estimated as zeros.

3.5.2 State change detection with NDFA

The presented experiment demonstrates that the NDFA algorithm is a powerful tool for estimating a good model for quite complex dynamical systems. The model can be used for many purposes, one of which could be monitoring the state of an industrial or natural process. In **Publication 2**, we demonstrate how the NDFA approach can be used for the problem of *detecting changes* in an observed dynamic process.

Change detection problem

The task of change detection is important in many fields of engineering and it is often related to fault diagnosis (Chen and Patton, 1999; Chiang et al., 2001). An abrupt change in the process usually indicates a fault, and the goal of change detection is to pinpoint the occurrence of the fault and to give an alarm. It would also be very desirable to be able to analyze exactly where in the process the original reason for the fault is. This may be quite difficult because a fault in some underlying subsystems or parameters may manifest itself in complicated ways in the observables, or sometimes be hardly observable at all.

Detection of changes in stochastic processes has been studied extensively (see, e.g., books by Basseville and Nikiforov, 1993; Gustafsson, 2000). Many classical methods monitor some direct indicators of the process observables and respond to

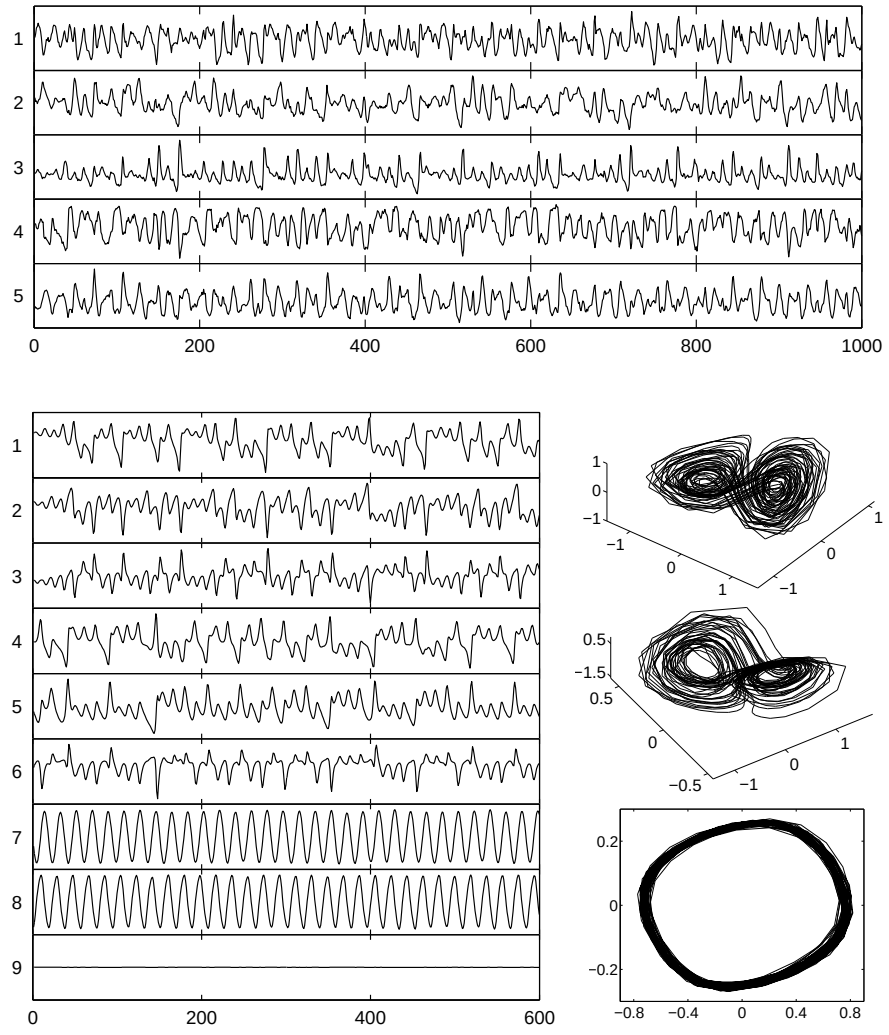


Figure 3.8: Above: Observations artificially generated as a nonlinear mixture of three dynamic processes. Only five out of ten observations used in the experiments are presented here. Below: The 9 sources estimated by NDFA. The left plot shows the time series for the beginning of the observation period (first 600 samples). The plots on the r.h.s. are the phase curves of the three separated subspaces: subspaces of components 1–3, 4–6 and 7–8 (from top to bottom).

changes in the indicators, such as the mean or variance of a process measurement. Such indicator-based approaches do not take all the relevant information about the process into account which usually means a delayed response to a change or neglect of the change in the worst case.

A better solution is to estimate a more sensitive model of the process and then use the goodness-of-fit of the new observations to the previously established model as the change indicator. For dynamical systems, state-space models are typical modeling tools. For *linear* SSMs, the change detection problem has been studied well and the most common technique is to test the statistical properties of the innovations generated by a Kalman filter (Basseville and Nikiforov, 1993; Gustafsson, 2000). For nonlinear SSMs, the results on detecting changes have been quite limited. The main approach to this problem is linearization like in the extended Kalman filter, and applying change detection methods to linearized systems.

The classical change detection methods often assume that the model of a process is known. In many cases, however, the model must be learned from available training data. For example, in real industrial processes, the state variables, dynamics and observation mapping are rarely known accurately enough to allow model-based approaches without estimating the process from the data. NDFA is a powerful tool for learning a nonlinear state-space model, which can efficiently be used in the problem of change detection.

NDFA for state change detection

The approach to change detection proposed in **Publication 2** makes use of the cost function provided by the NDFA algorithm in order to monitor the (differential) *entropy rate* of the observed process. For stationary stochastic processes, the entropy rate is defined as

$$h(\mathbf{x}) = \lim_{t\to\infty} \mathrm{E}\{-\log p(\mathbf{x}(t) \,|\, \mathbf{x}(t-1), \ldots, \mathbf{x}(1)\}\,, \tag{3.67}$$

where the expectation is taken over $p(\mathbf{x}(1), \ldots, \mathbf{x}(t))$ (Cover and Thomas, 1991). Using a process realization $\{\mathbf{x}(t-L+1), \ldots, \mathbf{x}(t)\}$ of length L, the entropy rate can be estimated as

$$\widehat{h}_L(t) = \frac{1}{L}\sum_{\tau=0}^{L-1} -\log p(\mathbf{x}(t-\tau) \,|\, \mathbf{x}(t-\tau-1), \ldots, \mathbf{x}(1)) \tag{3.68}$$

$$= -\frac{1}{L}\log p(\mathbf{x}(t-L+1), \ldots, \mathbf{x}(t) \,|\, \mathbf{x}(t-L), \ldots, \mathbf{x}(1))\,. \tag{3.69}$$

Now, based on the stationarity assumption, one can assume that short-time estimates $\widehat{h}_L(t)$ of the entropy rate fluctuate around a constant mean which is the

true value of the entropy rate. If the process changes, it is likely that its entropy changes as well and so does the mean value for $\widehat{h}_L(t)$. The entropy rate can therefore be taken as the indicator of change and alarm can be raised if the estimated value of the entropy rate either decreases or increases.

The estimation of the entropy rate can be done using the VB cost function provided by the NDFA algorithm. To see that, let us assume that the posterior approximation $q(\mathbf{S}, \boldsymbol{\theta})$ is close to the true posterior $p(\mathbf{S}, \boldsymbol{\theta} \,|\, \mathbf{X})$ and therefore the KL-divergence between the two densities is close to zero

$$D_t = D(q(\mathbf{S}_t, \boldsymbol{\theta}) \,||\, p(\mathbf{S}_t, \boldsymbol{\theta} \,|\, \mathbf{X}_t)) \approx 0\,. \tag{3.70}$$

In this equation, the subscript t emphasizes the fact that the data matrix $\mathbf{X}$ grows when new data arrive and therefore D depends on time. It follows from Eq. (3.70) that the cost function in Eq. (3.28) gives the estimate of the log-evidence:

$$\mathcal{C}(t) \approx -\log p(\mathbf{x}(1), \ldots, \mathbf{x}(t))\,. \tag{3.71}$$

Then, a short-time estimate of the entropy rate can be computed as

$$\widehat{h}_L(t) = \frac{1}{L}\Big(\mathcal{C}(t) - \mathcal{C}(t-L)\Big)\,. \tag{3.72}$$

The deviations of the quantity $\widehat{h}_L(t)$ from the entropy rate value calculated from the training sequence can then be monitored using the standard CUSUM test (Basseville and Nikiforov, 1993). Note that this approach is valid if D_t in Eq. (3.70) is not zero but represents a process with a stationary mean.

In **Publication 2**, we show how the cost function in Eq. (3.71) can be calculated efficiently when new measurements arrive. It is demonstrated that monitoring the terms of the cost function helps detect the states that undergo the most significant changes. Thus, an important feature of the proposed NDFA approach to change detection is that it is able not only to pinpoint the time of the change, but also to show which of the underlying states might be the reason for the change.

Example of state change detection

A change in a real process can take place in a variety of ways. In the NSSM model, it is reflected in a change either in the mapping $\mathbf{f}$ from the states to the observations, in the underlying state dynamics determined by the mapping $\mathbf{g}$, or in the noise levels. These changes can be detected by monitoring the NDFA cost function. **Publication 2** concentrates on the most demanding case where the nonlinearity $\mathbf{g}$ undergoes some change. The nonlinear mapping $\mathbf{f}$ can make this change hardly discernible in the observations, making the change detection problem very challenging.

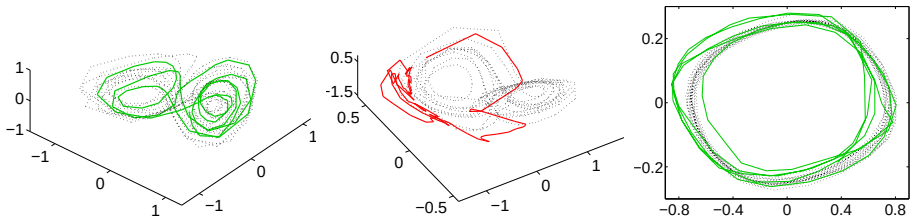


Figure 3.9: The phase curves of the three separated subspaces (refer to Fig. 3.8) for test data with a simulated pronounced change. The presented components are estimated using the NDFA model learned for the training data. The dotted and solid lines represent the estimated components before and after the moment of change, respectively. The cost function contribution changes most significantly for the components of the second subspace.

In the experiments, we consider the same artificial dynamic process for which Valpola and Karhunen (2002) estimated the NDFA model (see Fig. 3.8). The changes in the process are simulated by changing the parameters of one of the Lorenz processes or the harmonic oscillator. Both the case of pronounced changes in the dynamics (which become clearly visible in the measurements) and the case of slight changes (which are hardly visible in the observations) are investigated experimentally. The proposed approach is shown to detect the simulated changes and in the considered change detection tasks, it outperforms other approaches based on alternative models.

Fig. 3.9 presents the states estimated using the NDFA algorithm for test data which contains a simulated pronounced change. Note that the curves of the second decoupled subspace undergo the most significant changes. Also, the cost function terms corresponding to the states of this subspace changes most noticeably (see the cost function values in Publication 2).

3.6 Conclusions

In this chapter, several results on applying variational Bayesian methods to different LVMs have been presented. We started with a brief introduction to the basics of probability theory and the principles of Bayesian inference. Then, we outlined several popular methods for approximate evaluation of the posterior distribution of the unknown model parameters. Variational Bayesian learning, which is the main focus of this chapter, was emphasized.

The important characteristic of variational Bayesian learning have been discussed. The main advantages of VB methods include their elegant way to do

model selection, resistance against overfitting, and the possibility to regularize solutions by choosing a suitable form of the posterior approximation. We also discussed potential problems with applying VB learning in practice. They include high computational complexity, multiple local minima, the possibility to underfit, and a possible bias in favor of some types of solutions. Methods related to VB estimation have been outlined as well.

This chapter presented a model called post-nonlinear factor analysis which is learned using the VB approach. PNFA is a generative LVM where the hidden variables are described by the Gaussian distribution and the generative mapping is restricted to the post-nonlinear type. The proposed PNFA method can be applied to the ICA problem in post-nonlinear mixtures and it can overcome some limitations of the existing alternative methods. In particular, it can separate sources from mixtures with non-invertible PNL distortions provided that the number of the observed variables is greater than the number of the sources and the full generative mapping is invertible.

The computational complexity of VB methods depends significantly on the chosen form of the posterior approximation. A simpler, factorial form usually yields a faster learning algorithm. However, the form of the posterior approximation can introduce a bias in favor of some type of solutions and the result of VB learning is usually a compromise between the solutions where the explanation of the data is best and the solutions where the posterior approximation is most accurate. In this chapter, this effect was discussed first using a hypothetical example. Then, it was shown both theoretically and experimentally that a fully factorial approximation in linear ICA models introduces a bias in favor of the PCA solution. This result also generalizes to the case of nonlinear mixtures.

The effect of posterior approximation can be a negative result but sometimes it is possible to use it to select an appropriate solution among otherwise degenerate solutions. In this chapter, this regularization was shown to exist in the nonlinear dynamic factor analysis model introduced by Valpola and Karhunen (2002) for estimation of nonlinear state-space models. The NDFA algorithm based on VB learning can achieve a meaningful representation of the sources by using a suitable posterior approximation. This was shown here by emphasizing the subspace separation results in the experiments reported by Valpola and Karhunen (2002).

The last part of this chapter presented a potential application for the models learned using the VB approach. It was demonstrated how the NDFA algorithm can be applied to the problem of detecting changes in the dynamics of a complex process. The proposed approach uses the VB cost function in order to calculate a short-time estimate of the entropy rate of the process. This quantity is assumed stationary if the process does not undergo any changes and therefore it can be used as the indicator of change.

Appendix to Chapter 3: proofs

Posterior approximation $q(\mathbf{S})$ in NDFA

The following derivations show that the parameterization of the posterior in Eqs. (3.59)–(3.61) used in the NDFA algorithm corresponds to modeling the posterior of all the source values with a Gaussian distribution whose covariance is parameterized as presented in Eq. (3.63).

It follows from Eq. (3.59) that the sources are modeled to be independent a posteriori, that is

$$q(\mathbf{S}) = \prod_j q(s_j(1), \ldots, s_j(T)) \,. \tag{3.73}$$

Let us first consider the posterior $q(s_j(1), \ldots, s_j(T))$ describing the values of one source s_j. We use the following notation $z_t = s_j(t)$, $\overline{z}_t = \overline{s}_j(t)$, $\widetilde{z}_t = \widetilde{s}_j(t)$ where $\overline{s}(t)$, $\widetilde{s}(t)$ parameterize the posterior as presented in Eqs. (3.60) and (3.61).

The approximate pdf $q(z_1, z_2)$ for two successive values is equal to

$$\begin{aligned}
q(z_1, z_2) &= q(z_1) q(z_2 \,|\, z_1) \\
&\propto \exp\left(-\frac{1}{2}\widetilde{z}_1^{-1}(z_1 - \overline{z}_1)^2\right) \exp\left(-\frac{1}{2}\widetilde{z}_2^{-1}(z_2 - \overline{z}_2 - \rho_{1,2}(z_1 - \overline{z}_1))^2\right) \\
&= \exp\Bigg(-\frac{1}{2}\Big[\widetilde{z}_1^{-1}(z_1 - \overline{z}_1)^2 + \widetilde{z}_2^{-1}(z_2 - \overline{z}_2)^2 \\
&\qquad\qquad -2\widetilde{z}_2^{-1}\rho_{1,2}(z_1 - \overline{z}_1)(z_2 - \overline{z}_2) + \widetilde{z}_2^{-1}\rho_{1,2}^2(z_1 - \overline{z}_1)^2\Big]\Bigg) \\
&= \exp\Bigg(-\frac{1}{2}\Big[(z_1 - \overline{z}_1)^2(\widetilde{z}_1^{-1} + \widetilde{z}_2^{-1}\rho_{1,2}^2) \\
&\qquad\qquad +\widetilde{z}_2^{-1}(z_2 - \overline{z}_2)^2 - 2\widetilde{z}_2^{-1}\rho_{1,2}(z_1 - \overline{z}_1)(z_2 - \overline{z}_2)\Big]\Bigg) \\
&\propto \exp\left(-\frac{1}{2}\mathbf{z}_{1..2}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_{1..2}^{-1}\mathbf{z}_{1..2}\right) ,
\end{aligned}$$

where

$$\mathbf{z}_{1..2} = \begin{bmatrix} z_1 - \overline{z}_1 & z_2 - \overline{z}_2 \end{bmatrix}^{\mathrm{T}}$$

and

$$\widetilde{\mathbf{\Sigma}}_{1..2}^{-1} = \begin{bmatrix} \widetilde{z}_1^{-1} + \widetilde{z}_2^{-1}\rho_{1,2}^2 & -\widetilde{z}_2^{-1}\rho_{1,2} \\ -\widetilde{z}_2^{-1}\rho_{1,2} & \widetilde{z}_2^{-1} \end{bmatrix} .$$

It can be shown likewise that the approximate pdf $q(z_1, z_2, z_3)$ is equal to

$$q(z_1, z_2, z_3) = q(z_1, z_2) q(z_3 \,|\, z_2) \propto \exp\left(-\frac{1}{2}\mathbf{z}_{1..3}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_{1..3}^{-1}\mathbf{z}_{1..3}\right) ,$$

where

$$\mathbf{z}_{1..3} = \begin{bmatrix} z_1 - \overline{z}_1 & z_2 - \overline{z}_2 & z_3 - \overline{z}_3 \end{bmatrix}^{\mathrm{T}}$$

and $\widetilde{\mathbf{\Sigma}}_{1..3}^{-1}$ is a tridiagonal matrix

$$\widetilde{\mathbf{\Sigma}}_{1..3}^{-1} = \begin{bmatrix} \widetilde{z}_1^{-1} + \widetilde{z}_2^{-1}\rho_{1,2}^2 & -\widetilde{z}_2^{-1}\rho_{1,2} & 0 \\ -\widetilde{z}_2^{-1}\rho_{1,2} & \widetilde{z}_2^{-1} + \widetilde{z}_3^{-1}\rho_{2,3}^2 & -\widetilde{z}_3^{-1}\rho_{2,3} \\ 0 & -\widetilde{z}_3^{-1}\rho_{2,3} & \widetilde{z}_3^{-1} \end{bmatrix}. \tag{3.74}$$

These results can easily be generalized to T source values. Thus, the approximate pdf $q(s_j(1), \ldots, s_j(T))$ is Gaussian and the inverse of the corresponding covariance matrix has a tridiagonal structure, similar to Eq. (3.74). Note that non-zero elements in Eq. (3.74) appear only on the main diagonal and in the elements corresponding to two successive source values.

Now taking into account other sources and formatting all source values according to Eq. (3.62) yields a Gaussian pdf whose covariance is given by Eq. (3.63).

Chapter 4

Faster separation algorithms

4.1 Introduction

The source separation methods considered in this chapter assume the linear mixing model in Eq. (2.13) in which the noise term $\mathbf{n}(t)$ is typically omitted:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{j=1}^{M} \mathbf{a}_j s_j(t)\,. \tag{4.1}$$

Using the matrix notation of Eq. (2.1), this can be written as

$$\mathbf{X} = \mathbf{A}\mathbf{S}\,. \tag{4.2}$$

As was discussed in Section 2.2, the reconstruction of the sources $s_j(t)$ can be achieved based on some prior assumptions or by using knowledge about the unknown parts of the model. Independence of sources is often used when very little is known about the underlying processes and therefore ICA has become a popular tool for exploratory data analysis. As was reviewed in Section 2.2, independence can be utilized in different ways by using such assumptions as non-Gaussianity of source distributions, distinct autocorrelation or frequency structures of the sources, or non-stationarity of source variances. Different approaches may be suited better for particular problems or applications. Sometimes it is also possible to combine several approaches in order to improve the quality of separation.

Very often, one may have some idea about the nature of the sources which might be underlying the data. Relevant signals are often expected to have specific temporal, spectral or spatial characteristics and it would be very useful to incorporate such prior knowledge into the separation algorithm directly. For example, in biomedical applications, some idea about the waveform of the heart

beat can help extract cardiac artifacts from MEG recordings. Such prior information can also be used in exploratory data analysis when one investigates what kind of components it is possible to find in the data by using different types of assumptions. This kind of problem setting, with some prior knowledge available, is often called *semiblind*.

Bayesian methods considered in Chapter 3 are popular for their principled way to express modeling assumptions and prior knowledge in terms of probability distributions. For example, the known characteristics of the sources and the mixing matrix could be modeled by properly chosen priors for $\mathbf{S}$ and $\mathbf{A}$, respectively. Thus, Bayesian methods are good candidates to be used in semiblind source separation problems. However, the main drawback of Bayesian methods is their *high computational burden*. For example, learning a model like NDFA with a decent number of unknown parameters may take several days on a modern computer. This makes these methods hardly applicable to large-scale problems and complicates exploratory data analysis when different types of models are likely to be tried.

This chapter considers semiblind methods which are not Bayesian as they do not have an explicit density model for all the unknown parameters. It is shown however, that the resulting algorithms can sometimes have an interpretation as approximate Bayesian inference. All the algorithms presented in this chapter follow the unifying algorithmic framework of *denoising source separation* introduced by Särelä and Valpola (2005). This framework allows for easy development of source separation methods which can be either completely blind, or combine such blind criteria as independence with some prior knowledge (which is done in constrained ICA methods, James and Hesse, 2005), or use the prior information alone to achieve separation.

The methods proposed in this chapter were originally designed for exploratory analysis of climate data. The dataset considered in this thesis is a huge collection of global climate measurements obtained for the last 56 years and thus the high dimensionality of the dataset (more than 20,000 time instances in about 10,000 spatial locations) was one of the main reasons for applying *fast* and relatively simple separation algorithms. Most of the presented algorithms were motivated by the patterns and regularities found in the considered climate dataset. Yet, the proposed methods are quite general and could be applied to other types of data as well.

4.2 The general algorithmic framework

The algorithmic framework of denoising source separation (DSS), as presented by Särelä and Valpola (2005), is a general sequence of steps used by different source separation algorithms. The sources estimated in that framework are generally

assumed

1. to be mutually uncorrelated,
2. to have some structure known from the available prior information.

Typically, maximizing the structure of components makes them more independent and thus DSS can be seen as generalization of ICA with relaxed independence assumption.

4.2.1 Preprocessing and demixing

The requirement that the sources are uncorrelated is assured by using a preprocessing step called whitening or sphering (Hyvärinen et al., 2001). Whitening makes the covariance structure of the data uniform in such a way that any linear projection of the data has unit variance. The positive effect of such a transformation is that any orthogonal basis in the whitened space defines uncorrelated sources. Therefore, whitening is used as a preprocessing step in many ICA algorithms, and the mixing matrix can be restricted to be orthogonal afterwards.

Whitening is usually performed by PCA with normalizing the principal components to unit variances. If measurements $\mathbf{X}$ are centered, the matrix of sphered data $\mathbf{Y}$, defined similarly to Eq. (2.1), is calculated as

$$\mathbf{Y} = \mathbf{D}^{-1/2}\mathbf{V}^{\mathrm{T}}\mathbf{X}\,, \tag{4.3}$$

where $\mathbf{D}$ is the diagonal matrix of the eigenvalues of the data covariance matrix defined in Eq. (2.6). The columns of matrix $\mathbf{V}$ are the corresponding eigenvectors. The dimensionality of the data can also be reduced at this stage by retaining only the principal components corresponding to the largest eigenvalues in $\mathbf{D}$.

It is easy to show that the covariance matrix calculated for the whitened data $\mathbf{Y}$ is the identity matrix. Matrix $\mathbf{Y}$ is not unique, though; any *orthogonal* rotation of its columns produces a matrix

$$\mathbf{S} = \mathbf{W}\mathbf{Y} \tag{4.4}$$

that also has unit covariance. Therefore, a set of uncorrelated sources can be found by using Eq. (4.4) with the restriction that $\mathbf{W}$ is an orthogonal matrix. Matrix $\mathbf{W}$ (or the overall transformation matrix $\mathbf{W}\mathbf{D}^{-1/2}\mathbf{V}^{\mathrm{T}}$) is often called a *demixing matrix* in the ICA literature.

The matrix $\mathbf{S}$ of the source values is defined similarly to Eq. (2.1). Each row of $\mathbf{S}$ contains all the values of one source for the whole observation period. In this chapter, one row of $\mathbf{S}$ is denoted by

$$\mathbf{s}_{1..T}^{\mathrm{T}} = \begin{bmatrix} s(1) & \dots & s(T) \end{bmatrix}\,. \tag{4.5}$$

In some applications, it can be desirable to extract only one source at a time. Then, the rows $\mathbf{s}_{1..T,j}^{\mathrm{T}}$ are estimated one after another as

$$\mathbf{s}_{1..T,j}^{\mathrm{T}} = \mathbf{w}_j^{\mathrm{T}}\mathbf{Y}\,, \tag{4.6}$$

where the demixing vectors $\mathbf{w}_j^{\mathrm{T}}$ are the rows of the matrix $\mathbf{W}$.

The optimal matrix $\mathbf{W}$ (or its rows $\mathbf{w}_j^{\mathrm{T}}$) is found so as to maximize the desired properties of components $\mathbf{S}$, that is by using the second DSS requirement.

4.2.2 Special case with linear filtering

In some cases, the interesting properties of a source signal can be obtained by applying a linear temporal filter. For example, the sources are sometimes known to be cyclic over a certain period of time or to have prominent variability in a certain timescale and filtering would emphasize this characteristic structure of the sources.

Using the notation of Eq. (4.5), linear filtering is written as

$$\widehat{\mathbf{s}}_{1..T}^{\mathrm{T}} = \mathbf{s}_{1..T}^{\mathrm{T}}\mathbf{F}\,, \tag{4.7}$$

where $\mathbf{F}$ is the filtering matrix of dimensionality $T\times T$. The amount of structure in the signal can then be measured by a quantity that gives the ratio of the variance of the filtered component $\widehat{s}$ and the variance of the non-filtered component s:

$$\mathcal{F}(\mathbf{s}_{1..T}(\mathbf{w})) = \frac{\mathrm{var}\{\widehat{s}\}}{\mathrm{var}\{s\}} = \frac{\sum_{t=1}^{T}\widehat{s}^{\,2}(t)}{\sum_{t=1}^{T}s^2(t)} = \frac{\|\mathbf{s}_{1..T}^{\mathrm{T}}\mathbf{F}\|^2}{\|\mathbf{s}_{1..T}\|^2} = \frac{\|\mathbf{w}^{\mathrm{T}}\mathbf{Y}\mathbf{F}\|^2}{\|\mathbf{w}^{\mathrm{T}}\mathbf{Y}\|^2}\,, \tag{4.8}$$

where $\widehat{s}(t)$ denotes one element of $\widehat{\mathbf{s}}_{1..T}$. The measure in Eq. (4.8) can be understood as the relative amount of energy contained in the interesting part of the signal and it attains its maximum value of unity if filtering does not change the signal. In Publication 7, we use the term *clarity* for this quantity.

The sources can be estimated one by one using Eq. (4.6) so as to maximize the objective function in Eq. (4.8). It can be shown, however, that for many practical cases such estimation can be performed in just three steps, when whitening is followed by filtering and PCA (see Fig. 4.1).

The intuition behind this approach is that filtering on the second step renders the variances of the sphered components different and the covariance matrix of $\widehat{\mathbf{Y}}$ is no more equal to the identity matrix. Note that in many practical situations, this filtering can be done using the same filter $\mathbf{F}$ as in Eq. (4.8) (Särelä and Valpola, 2005). Then, PCA can identify the directions which maximize the properties of interest. The eigenvalues obtained from PCA on the third step give the values of the objective function $\mathcal{F}$ for the found sources. Thus, the components

Figure 4.1: Separation algorithm in case of linear denoising.

are ranked according to the prominence of the desired properties (their clarity values) the same way as the principal components in PCA are ranked according to the amount of variance they explain.

The procedure presented in Fig. 4.1 is basically equivalent to joint diagonalization of the data covariance matrix $\mathbf{C}$ and the covariance of the filtered data $\mathbf{C}_f$ given in Eqs. (2.6) and (2.20), respectively. Thus, this algorithm can solve the source separation problem, that is it can reconstruct the original sources, under the following conditions: 1) the original sources and their filtered versions are mutually uncorrelated and 2) the clarity values of the components are different. Note also that the considered three-step algorithm optimizes the same type of cost function as the maximum noise fraction transform proposed by Green et al. (1988).

4.2.3 General case of nonlinear denoising

In the general case, the interesting properties of the sources could be quite sophisticated and the quantity $\mathcal{F}(\mathbf{s}_{1..T}(\mathbf{w}))$ measuring the amount of desired structure in a signal could be quite complex. This measure depends on the source values which are estimated using the demixing vector $\mathbf{w}$. Therefore, $\mathcal{F}$ should be optimized w.r.t. $\mathbf{w}$.

The optimization of such an objective function could be done by the following gradient-based algorithm. It follows from Eq. (4.6) that for whitened data $\mathbf{Y}$ it holds that

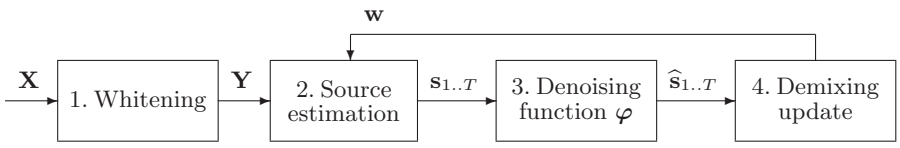
$$\mathbf{w} = \frac{1}{T}\mathbf{Y}\mathbf{s}_{1..T}\,. \tag{4.9}$$

Using the chain rule for computing derivatives, it follows from Eq. (4.6) that the gradient of $\mathcal{F}(\mathbf{s}_{1..T}(\mathbf{w}))$ w.r.t. $\mathbf{w}$ can be computed from the gradient w.r.t. $\mathbf{s}_{1..T}$ as

$$\frac{\partial \mathcal{F}}{\partial \mathbf{w}} = \mathbf{Y}\frac{\partial \mathcal{F}}{\partial \mathbf{s}_{1..T}}\,. \tag{4.10}$$

Now using Eqs. (4.9)–(4.10), the gradient ascent step for $\mathbf{w}$ can be written as

$$\mathbf{w}_{\text{new}} = \mathbf{w} + \mu\frac{\partial \mathcal{F}}{\partial \mathbf{w}} = \frac{1}{T}\mathbf{Y}\left(\mathbf{s}_{1..T} + T\mu\frac{\partial \mathcal{F}}{\partial \mathbf{s}_{1..T}}\right)\,, \tag{4.11}$$



Deflation:	$\mathbf{s}_{1..T,j}^{\mathrm{T}} = \mathbf{w}_j^{\mathrm{T}}\mathbf{Y}$	$\widehat{\mathbf{s}}_{1..T,j} = \boldsymbol{\varphi}(\mathbf{s}_{1..T,j})$	$\mathbf{w}_j = \mathrm{orth}(\mathbf{Y}\,\widehat{\mathbf{s}}_{1..T,j})$
Symmetric:	$\mathbf{S} = \mathbf{W}\mathbf{Y}$	$\widehat{\mathbf{S}} = \boldsymbol{\varphi}(\mathbf{S})$	$\mathbf{W}^{\mathrm{T}} = \mathrm{orth}(\mathbf{Y}\widehat{\mathbf{S}}^{\mathrm{T}})$

Figure 4.2: The general sequence of steps in the algorithmic framework of denoising source separation. The equations explain the operations on different steps for the deflation and symmetric approaches.

where μ is the step size. Eq. (4.11) shows that one step for optimizing $\mathbf{w}$ can be performed by first updating the sources with the step size μ_s:

$$\widehat{\mathbf{s}}_{1..T} = \mathbf{s}_{1..T} + \mu_s \frac{\partial \mathcal{F}}{\partial \mathbf{s}_{1..T}} = \boldsymbol{\varphi}(\mathbf{s}_{1..T}) \tag{4.12}$$

and then calculating the new value for $\mathbf{w}$ using Eq. (4.9). This yields the sequence of steps presented in Fig. 4.2, which is iterated until convergence.

In the *deflation* approach, the components $\mathbf{s}_j$ are estimated one after another. Then, the function orth(.) in Step 4 implements the Gram-Schmidt orthogonalization, when the demixing vector $\mathbf{w}_j$ is made orthogonal to the previously found vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}$ (see, e.g., Hyvärinen et al., 2001).

In the *symmetric* approach, all the components are estimated simultaneously, as in Eq. (4.4). Then, the denoising function in Step 3 is applied to all the sources, that is $\widehat{\mathbf{S}} = \boldsymbol{\varphi}(\mathbf{S})$, which means that the values of one source can affect the new values for another source. The operator orth(.) in Step 4 gives the orthogonal projection of the matrix $\mathbf{Y}\widehat{\mathbf{S}}^{\mathrm{T}}$ onto the set of orthogonal matrices.

The basic idea of the algorithmic framework called *denoising source separation* (Särelä and Valpola, 2005) is to design separation algorithms following the general sequence of steps presented in Fig. 4.2. The separation criterion is introduced in the procedure in the form of a suitably chosen *denoising function* $\boldsymbol{\varphi}$. In case the algorithm is derived from an optimized measure $\mathcal{F}$, the corresponding denoising function is given by Eq. (4.12). For many practical cases, however, it can be easier to construct an update rule

$$\widehat{\mathbf{s}}_{1..T} = \boldsymbol{\varphi}(\mathbf{s}_{1..T}) \tag{4.13}$$

with a sensible function $\boldsymbol{\varphi}$ than to derive a gradient-based rule in Eq. (4.12) from an objective function. First, the interesting signal structure could be difficult to measure using a simple index $\mathcal{F}$. Second, the derivation of the gradient $\partial\mathcal{F}/\partial\mathbf{s}_{1..T}$

could be cumbersome, especially for complex $\mathcal{F}$. It is also possible that the gradient-based update rule in Eq. (4.12) is not robust as, for example, it can be sensitive to some particular values of s (see an example in Section 4.3.4).

In general, the denoising function $\boldsymbol{\varphi}(\mathbf{s}_{1..T})$ should be designed such that it emphasizes the desired (interesting) properties of the signal and removes irrelevant information from $\mathbf{s}_{1..T}$. It can represent a gradient-based update rule or its modification. Sometimes, it is possible to derive an appropriate denoising function from rather heuristic principles. Also note that for any $\boldsymbol{\varphi}(\mathbf{s}_{1..T})$, it is possible to modify Eq. (4.13) by adding a term $\alpha + \beta\mathbf{s}_{1..T}$, with α, β some constants, as in

$$\widehat{\mathbf{s}}_{1..T} \propto \alpha + \beta\mathbf{s}_{1..T} + \boldsymbol{\varphi}(\mathbf{s}_{1..T})\,, \tag{4.14}$$

without changing the fixed points of the algorithm (Särelä and Valpola, 2005).

In DSS terminology, the iterative procedure in Fig. 4.2 is usually interpreted as extension of the power method for computing the principal components of $\mathbf{Y}$. Without denoising, this procedure is indeed equivalent to the power method, because then Steps 2 and 4 give $\mathbf{w} = \mathrm{orth}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{w})$. Since $\mathbf{Y}$ is white, all the eigenvalues are equal and the solution without denoising becomes degenerate. Therefore, even slightest changes made by denoising $\boldsymbol{\varphi}$ can determine the rotation. Since the denoising procedure emphasizes the desired properties of the sources, the algorithm can find the rotation where the properties of interest are maximized.

It should be noted that the presented procedure is very general. The essential part of any specific algorithm implemented in this framework is the denoising procedure. In fact, many existing ICA algorithms fall into the pattern of DSS although they have been derived from other perspectives, typically from a properly chosen cost function. Examples of such algorithms include the FastICA algorithm where the maximized structure is non-Gaussianity of the sources (Hyvärinen et al., 2001), the semiblind algorithm which uses the knowledge of the source autocorrelation function (Barros and Cichocki, 2001) and the blind algorithm for extraction of sources which are expected to have prominent frequencies in their spectra (Cichocki et al., 2002; Cichocki and Amari, 2002).

4.2.4 Calculation of spatial patterns

In the applications, we are interested not only in the sources $\mathbf{S}$, but also in the matrix $\mathbf{A}$ in Eq. (4.1). From Eqs. (4.2)–(4.4), it follows that

$$\mathbf{X} = \mathbf{A}\mathbf{S} = \mathbf{A}\mathbf{W}\mathbf{Y} = \mathbf{A}\mathbf{W}\mathbf{D}^{-1/2}\mathbf{V}^{\mathrm{T}}\mathbf{X}\,. \tag{4.15}$$

Thus $\mathbf{A}$ should be chosen as the (pseudo)inverse of $\mathbf{W}\mathbf{D}^{-1/2}\mathbf{V}^{\mathrm{T}}$ which is

$$\mathbf{A} = \mathbf{V}\mathbf{D}^{1/2}\mathbf{W}^{\mathrm{T}}\,. \tag{4.16}$$

Since the extracted components are normalized to unit variances, the columns of $\mathbf{A}$ have a meaningful scale. If the sensor array has a spatial arrangement, which is the case for spatio-temporal datasets, each column $\mathbf{a}_j$ of the mixing matrix can be visualized as a spatial map showing how the effect of the j-th source is distributed over the sensor array.

Note that the signs of the extracted components cannot generally be determined, which is a well-known property of the classical ICA problem. Such ambiguity arises when $\boldsymbol{\varphi}(\mathbf{s}_{1..T}) = -\boldsymbol{\varphi}(-\mathbf{s}_{1..T})$. The sign indeterminacy can be resolved if there exists some information about the asymmetry of the source distributions.

The ambiguity of the solution is even higher for subspace models such as independent subspace analysis (Hyvärinen and Hoyer, 2000) or independent dynamics subspace analysis presented in Publication 8. There, the sources are decomposed into groups and the sources within a group are generally assumed dependent while components from different groups are mutually independent. Such models can be estimated only up to orthogonal rotations of sources within the groups.

A subspace of sources can be visualized by the observation variance explained by its components. For the model in Eq. (4.1), the variance of one observation x_i equals

$$\operatorname{var}\{x_i\} = \sum_{j=1}^{M} a_{ij}^2 \operatorname{var}\{s_j\} = \sum_{j=1}^{M} a_{ij}^2 \,, \tag{4.17}$$

which follows from the condition that the sources s_j are mutually uncorrelated and have unit variances. Thus, the variances explained by the sources from one subspace $\{s_j | j \in \mathcal{J}_k\}$ equal

$$\operatorname{var}_{\mathcal{J}_k}\{\mathbf{x}\} = \sum_{j \in \mathcal{J}_k} \mathbf{a}_j^2 \,, \tag{4.18}$$

where $\mathbf{a}_j^2$ denotes the vector of the squared elements of the mixing vector $\mathbf{a}_j$. The quantity in Eq. (4.18) is a vector whose dimensionality equals the number of sensors and therefore, for datasets with a spatial arrangement, it can be represented as a spatial pattern showing the effect on the observation variance in different spatial locations.

4.2.5 Connection to Bayesian methods

This section shows that learning Baysian ICA models can often be done in the presented algorithmic framework. This can be shown, for example, under the assumption that the mixing matrix $\mathbf{A}$ is point estimated and the source posterior is modeled using probability distributions, that is learning the posterior is done using the EM-algorithm: The source distributions are updated on the E-step and the mixing matrix is reestimated on the M-step.

In the following, let us consider the noisy model in Eq. (2.13) and assume that the data $\mathbf{X}$ have been prewhitened. Therefore, the mixing matrix is restricted to be orthogonal, that is $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$, and the transpose of $\mathbf{A}$ gives the demixing matrix $\mathbf{W}$. Another assumption made here is that the observation noise is isotropic in the whitened space, that is the observation noise covariance is $\mathbf{\Sigma_n} = v_x\mathbf{I}$.

The update rules are derived here by simplifying the learning rules used in variational Bayesian methods discussed in Chapter 3. Some of the notation used in this section is taken from Chapter 3, where $\langle\cdot\rangle$ denotes the expectation over the (approximating) posterior, and $\overline{\theta}$ is the variational parameter giving the (approximate) posterior mean of the parameter θ.

Reestimation of the mixing matrix

Let us first show that the new values of the mixing matrix obtained on the M-step are defined mostly by the means of the source posterior distributions.

When the Gaussian distribution is used both as the prior for the elements of $\mathbf{A}$ and to model the observation noise, the optimal posterior $q(\mathbf{A})$ is also Gaussian. Then, a natural choice for the point estimates for $\mathbf{A}$ is the mean of the Gaussian distribution $q(\mathbf{A})$. The i-th row of the mixing matrix $\mathbf{A}$ is here denoted by $\boldsymbol{\alpha}_i$. If $\boldsymbol{\alpha}_i$ has a zero-mean prior, the update rule for its posterior mean can be shown to be

$$\overline{\boldsymbol{\alpha}}_i = \left\langle v_{x,i}^{-1} \sum_{t=1}^{T} \mathbf{s}(t)\mathbf{s}(t)^{\mathrm{T}} + \mathbf{\Sigma}_{\boldsymbol{\alpha},i}^{-1} \right\rangle^{-1} \left\langle v_{x,i}^{-1} \right\rangle \sum_{t=1}^{T} x_i(t) \left\langle \mathbf{s}(t) \right\rangle \,, \tag{4.19}$$

where $\Sigma_{\boldsymbol{\alpha},i}$ is the covariance of the Gaussian prior for $\boldsymbol{\alpha}_i$ and $v_{x,i}$ denotes the variance of the noise in the i-th measurement channel. If the prior for the rows is very flat, $\mathbf{\Sigma}_{\boldsymbol{\alpha},i}^{-1}$ is close to zero and Eq. (4.19) simplifies to

$$\overline{\boldsymbol{\alpha}}_i = \left\langle \sum_{t=1}^{T} \mathbf{s}(t)\mathbf{s}(t)^{\mathrm{T}} \right\rangle^{-1} \sum_{t=1}^{T} x_i(t) \left\langle \mathbf{s}(t) \right\rangle \tag{4.20}$$

or in the matrix notation

$$\overline{\mathbf{A}} = \mathbf{X} \left\langle \mathbf{S} \right\rangle^{\mathrm{T}} \left\langle \mathbf{S}\mathbf{S}^{\mathrm{T}} \right\rangle^{-1} \,. \tag{4.21}$$

For whitened data, the factor $\left\langle \mathbf{S}\mathbf{S}^{\mathrm{T}} \right\rangle^{-1}$ accounts mostly for scaling the solution for $\mathbf{A}$. This follows from the fact that the sources should practically be uncorrelated when the estimates are close to the optimal solution, which yields

$$\left\langle \mathbf{S}\mathbf{S}^{\mathrm{T}} \right\rangle = \left\langle \mathbf{S} \right\rangle \left\langle \mathbf{S} \right\rangle^{\mathrm{T}} + \sum_{t=1}^{T} \mathbf{\Sigma}_{\mathbf{s}(t)} \approx T\,\mathbf{I} + \sum_{t=1}^{T} \mathbf{\Sigma}_{\mathbf{s}(t)} \,, \tag{4.22}$$

where the posterior covariances of the sources $\mathbf{\Sigma}_{\mathbf{s}(t)}$ are diagonal due to the orthogonality restriction on $\mathbf{A}$ (see Publication 1).

Now it follows from Eqs. (4.21)–(4.22) that the update of $\mathbf{A}$ can be done as

$$\mathbf{A} \leftarrow \text{orth}(\mathbf{X} \langle \mathbf{S} \rangle^{\mathrm{T}}) , \tag{4.23}$$

which is equivalent to Step 3 of the general algorithmic framework as $\mathbf{W} = \mathbf{A}^{\mathrm{T}}$.

ICA model with super-Gaussian sources

Let us now consider an example of the E-step, that is the update rules for the source distribution $q(\mathbf{s}(t))$. We consider here the ICA model with super-Gaussian sources presented in Publication 1. There, each source is modeled a priori as a Gaussian variable with zero mean and a time-dependent variance $v_{s,j}(t)$:

$$p(s_j(t) \,|\, \boldsymbol{\theta}) = N\left(s_j(t) \,|\, 0,\, v_{s,j}(t) \right) . \tag{4.24}$$

The mean of the fully factorial posterior approximation $q(\mathbf{s}(t))$ is updated using the following rule:

$$\overline{\mathbf{s}}(t) = \left\langle \mathbf{A}^{\mathrm{T}} \mathbf{\Sigma}_{\mathbf{n}}^{-1} \mathbf{A} + \mathbf{\Sigma}_{\mathbf{s}}^{-1}(t) \right\rangle^{-1} \left\langle \mathbf{A}^{\mathrm{T}} \mathbf{\Sigma}_{\mathbf{n}}^{-1} \mathbf{x}(t) \right\rangle , \tag{4.25}$$

where $\mathbf{\Sigma}_{\mathbf{s}}(t)$ is a diagonal matrix made up from the variances $v_{s,j}(t)$. Note that the mean values $\overline{\mathbf{s}}(t)$ are the most important for the M-step as was showed previously. It can be shown after straightforward calculations that each element of $\overline{\mathbf{s}}(t)$ in Eq. (4.25) can be computed as

$$\overline{s}_j(t) = \frac{1}{1 + \left\langle v_{s,j}^{-1}(t) \right\rangle / \left\langle v_x^{-1} \right\rangle} s_{x,j}(t) , \tag{4.26}$$

where $s_{x,j}(t)$ denotes the j-th element of the source vector computed from the data $\mathbf{x}(t)$ as

$$\mathbf{s}_x(t) = \mathbf{A}^{\mathrm{T}} \mathbf{x}(t) . \tag{4.27}$$

Since $\mathbf{W} = \mathbf{A}^{\mathrm{T}}$, Eq. (4.27) is equivalent to Step 2 of the general alorithmic framework and therefore Eq. (4.26) defines the denoising function.

Fig. 4.3a presents the function in Eq. (4.26) if $\left\langle v_x^{-1} \right\rangle = 1$ and the source variance is estimated such that $\left\langle v_{s,j}^{-1}(t) \right\rangle = s_{x,j}^{-2}(t)$. This function is a typical *shrinkage* function that can be used for extracting super-Gaussian sources (Hyvärinen, 1999b). Thus, the EM algorithm for this model can be simplified to a DSS procedure which uses a shrinkage function as denoising.

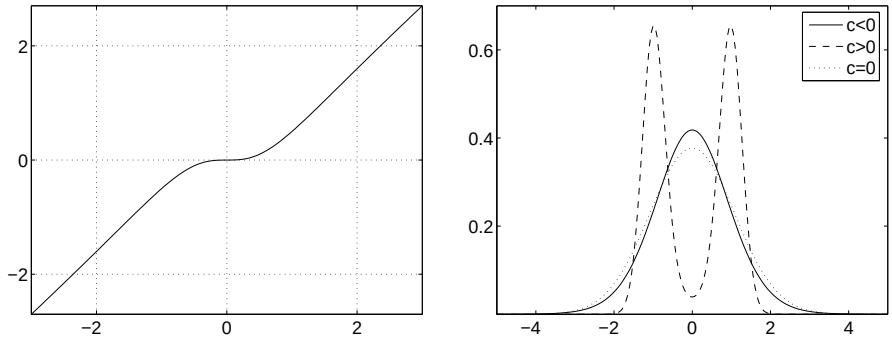


Figure 4.3: Left: The denoising function corresponding to the Bayesian model with super-Gaussian sources. Right: The prior model for the sources used in the Bayesian interpretation of FastICA with the tanh nonlinearity. The curves represent probability density functions defined by Eq. (4.39) with negative, zero and positive values for c.

Bayesian interpretation of FastICA

It is also possible to show that some algorithms which are derived without using the Bayesian principles and which follow the general DSS framework can have a Bayesian interpretation. In the following, the FastICA algorithm (Hyvärinen et al., 2001) is shown to have an interpretation as the EM-algorithm for a linear LVM with specific prior distributions for the sources. The presented derivations follow the view of the fast separation algorithms presented by Valpola and Pajunen (2000).

Let us first assume that each source is modeled as a Gaussian random variable with time-dependent mean and variance:

$$p(s_j(t) \,|\, \boldsymbol{\theta}) = N\left(s_j(t) \,|\, \mu_j(t),\, v_{s,j}(t) \right) . \tag{4.28}$$

The rule for updating the posterior mean for $q(\mathbf{s}(t))$ is then given by

$$\bar{\mathbf{s}}(t) = \left\langle \mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{n}}^{-1} \mathbf{A} + \boldsymbol{\Sigma}_{\mathbf{s}}^{-1}(t) \right\rangle^{-1} \left\langle \mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{n}}^{-1} \mathbf{x}(t) + \boldsymbol{\Sigma}_{\mathbf{s}}^{-1}(t) \boldsymbol{\mu}(t) \right\rangle , \tag{4.29}$$

where $\boldsymbol{\mu}(t)$ is made up from the elements $\mu_j(t)$, and $\boldsymbol{\Sigma}_{\mathbf{s}}(t)$ is a diagonal matrix made up from the variances $v_{s,j}(t)$. This can be transformed to

$$\bar{s}_j(t) = \frac{s_{x,j}(t) v_{s,j}(t) + \mu_j(t) v_x}{v_x + v_{s,j}(t)} , \tag{4.30}$$

where $\mathbf{s}_x(t)$ is defined in Eq. (4.27) and parameters $v_{s,j}$, v_x are assumed to be point estimated.

It is convenient to reformulate Eq. (4.30) using the score function $\psi(s) = \frac{\partial}{\partial s} \ln p(s)$ and its derivative, defined for a Gaussian distribution with mean μ and variance v as

$$\psi_{\mathrm{g}}(s) = \frac{\mu - s}{v}\,, \qquad \psi'_{\mathrm{g}} = \frac{\partial \psi_{\mathrm{g}}(s)}{\partial s} = -\frac{1}{v}\,. \tag{4.31}$$

This transforms Eq. (4.30) to the following update rule:

$$\overline{s}_j(t) = s_{x,j}(t) + \frac{\psi_{\mathrm{g},j}(s_{x,j}(t))v_x}{1 - v_x \psi'_{\mathrm{g},j}(s_{x,j}(t))} = s_{x,j}(t) + \frac{\psi_{\mathrm{g},j}(s_{x,j}(t))v_x}{1 + v_x / v_{s,j}(t)}\,. \tag{4.32}$$

Let us assume now that the prior model for each source is not restricted to Gaussian and the distribution in Eq. (4.28) is just a *local Gaussian approximation* of the true prior distribution. The noise variance v_x is typically much smaller than the variance $v_{s,j}(t)$ of the local approximation and therefore Eq. (4.32) can be approximated as

$$\overline{s}_j(t) \approx s_{x,j}(t) + \psi_{\mathrm{g},j}(s_{x,j}(t))v_x = s_{x,j}(t) + \Big(\mu_j(t) - s_{x,j}(t)\Big)\frac{v_x}{v_{s,j}(t)}\,, \tag{4.33}$$

which means that the solution for $\overline{s}_j(t)$ would be close to $s_{x,j}(t)$. Therefore, the Gaussian approximation in Eq. (4.28) can be computed in the vicinity of $s_{x,j}(t)$ by choosing the parameters $\mu_j(t)$ and $v_{s,j}(t)$ such that

$$\psi_{\mathrm{true},j}(s_{x,j}(t)) = \psi_{\mathrm{g},j}(s_{x,j}(t))\,, \qquad \psi'_{\mathrm{true},j}(s_{x,j}(t)) = \psi'_{\mathrm{g},j}\,. \tag{4.34}$$

This transforms Eq. (4.33) to the update rule

$$\overline{s}_j(t) \approx s_{x,j}(t) + \psi_{\mathrm{true},j}(s_{x,j}(t))v_x\,, \tag{4.35}$$

which is equivalent to the following denoising function

$$\widehat{\mathbf{s}}_{1..T} = \beta'\,\mathbf{s}_{1..T} + \psi_{\mathrm{true}}(\mathbf{s}_{1..T})\,, \tag{4.36}$$

with β' some constant. Eq. (4.36) should be compared with the update rule used in FastICA:

$$\widehat{\mathbf{s}}_{1..T} = \beta''\mathbf{s}_{1..T} + g(\mathbf{s}_{1..T}) \tag{4.37}$$

where g is some chosen nonlinearity applied component-wise and β'' is an updated constant. The criteria optimized with Eqs. (4.36) and (4.37) are equivalent if

$$\psi_{\mathrm{true}}(s) \propto \alpha + \beta s + g(s)\,. \tag{4.38}$$

A popular choice for $g(s)$ is the hyperbolic tangent and then it follows from Eq. (4.38) that the corresponding prior density model for the sources is defined by

$$p(s) = Z \exp(as + bs^2 + c \log\cosh s)\,, \tag{4.39}$$

where Z is the normalization constant. In the noiseless case, coefficients a, b, c should be chosen such that

$$\int_{-\infty}^{\infty} p(s)\,ds = 1\,, \quad \int_{-\infty}^{\infty} sp(s)\,ds = 0\,, \quad \int_{-\infty}^{\infty} s^2 p(s)\,ds = 1\,. \tag{4.40}$$

The requirement that s has zero mean yields $a = 0$, and the pair (b, c) has only one degree of freedom since the variance of s is constrained to unity. Depending on the sign of the parameter c, the distribution in Eq. (4.39) can model either super-Gaussian or sub-Gaussian sources as demonstrated in Fig. 4.3b. Negative c correspond to super-Gaussian distributions while positive c define sub-Gaussian distributions. Thus, one denoising $\boldsymbol{\varphi}$ used in FastICA suits a family of source distributions.

4.3 Fast algorithms proposed in this thesis

This section presents several source separation algorithms proposed in this thesis. All the presented algorithms follow the unifying algorithmic framework described in Section 4.2. Some of the proposed methods are derived so as to maximize an objective function $\mathcal{F}$ measuring the amount of the desired structure, while others are based on properly designed denoising procedures.

The following sections describe the optimized signal structure for each algorithm and outline the corresponding denoising procedure. Artificial source separation examples are presented for some of the algorithms. This section is based on **Publications 5-9** of this thesis.

4.3.1 Clarity-based analysis

Publication 5 presents a simple frequency-based analysis based on a linear filtering procedure, as explained in Section 4.2.2. In this analysis, filtering means passing spectral components within a certain frequency band and removing all other frequencies. Therefore, the algorithm can be used if relevant sources are expected to have prominent variability in a certain timescale.

The filtering matrix $\mathbf{F}$ used in the objective function in Eq. (4.8) is implemented in practice using the discrete cosine transform (DCT):

$$\mathbf{F} = \mathbf{V}_{\mathrm{dct}}^{\mathrm{T}} \boldsymbol{\Lambda} \mathbf{V}_{\mathrm{dct}}\,, \tag{4.41}$$

where $\mathbf{V}_{\mathrm{dct}}$ is the orthogonal matrix of the DCT basis in which one row $\mathbf{v}_f^{\mathrm{T}}$ corresponds to one DCT component with frequency f. $\boldsymbol{\Lambda}$ is a diagonal matrix with elements $\lambda_f \in [\,0,\,1\,]$ on the main diagonal. Then, the filtered signal can be

written as

$$\widehat{\mathbf{s}}_{1..T}^{\mathrm{T}} = \mathbf{s}_{1..T}^{\mathrm{T}}\mathbf{F} = \sum_f \lambda_f (\mathbf{v}_f^{\mathrm{T}}\mathbf{s}_{1..T})\mathbf{v}_f^{\mathrm{T}} \,. \tag{4.42}$$

Thus, a spectral component $\mathbf{v}_f$ is fully passed if the corresponding element λ_f equals unity, and it is removed from the signal if $\lambda_f = 0$.

The analysis is tuned to a specific frequency band by assigning large values to the elements λ_f corresponding to the frequencies of interest and setting $\lambda_f = 0$ for other frequencies. Then, the three-step procedure described in Section 4.2.2 can find the components that contain the largest relative amount of interesting frequencies in their power spectra. The extracted components are ordered according to their clarity values defined in Eq. (4.8).

The algorithm can be considered semiblind as it uses the knowledge of the frequency band of the prominent source variations. With low-pass filtering, the analysis is similar to the maximum autocorrelation factor transform proposed by Switzer (1985) and the linear case of slow feature analysis (Wiskott and Sejnowski, 2002). Therefore, we refer to this step as slow feature analysis in Publication 7. The application of this algorithm to global climate data is discussed in Section 4.4.3.

4.3.2 Frequency-based blind source separation

The algorithm described in the previous section is useful for extracting components with prominent structures in a certain frequency range. This requires some knowledge about the expected power spectra of the original components. In blinder settings, this information does not exist and the prominent spectral characteristics of the sources should be found automatically.

In **Publications 6** and **7**, we present an algorithm which can be seen as an extension of the previous approach. It achieves signal separation based on the assumption that the sources have *distinct power spectra*. Similarly to the previous approach, the interesting signal properties are emphasized by linear temporal filtering. However, since the sources are expected to have distinct frequency contents, an *individual* filter is applied to each source. The characteristic spectral properties of the sources are not known in advance, and therefore the filters are adjusted to the prominent spectral characteristics of the sources which emerge during the learning procedure. This approach is implemented using the general sequence of steps presented in Fig. 4.2, where the denoising function performs temporal filtering using a set of *adaptive* filters.

The corresponding denoising procedure is briefly outlined in the following, see also Table 4.1 for details. Note that each filter is in practice implemented using the filtering matrix $\mathbf{F}_j$ defined similarly to Eq. (4.41). Note also that the filtering matrix in Eq. (4.41) is defined by the diagonal elements λ_f of the matrix $\mathbf{\Lambda}$. A

Table 4.1: Denoising procedure for frequency-based separation

1. Compute DCT: $\mathbf{S}_{\mathrm{dct}} = \mathbf{S}\mathbf{V}_{\mathrm{dct}}^{\mathrm{T}}\mathbf{\Lambda}_{\mathrm{f}}$, where $\mathbf{\Lambda}_{\mathrm{f}}$ retains only interesting frequencies similarly to $\mathbf{\Lambda}$ in Eq. (4.41).

1. Estimate matrix $\mathbf{P}$ of power spectra values $P_j(f)$ for different sources j (in rows) and different frequencies f (in columns). This is done by, e.g., low-pass filtering the squares of the elements $\mathbf{S}_{\mathrm{dct}}$ in each row.

3. To increase the competition in weak frequencies, normalize $\mathbf{P}$ such the $\frac{1}{M}\sum_{j=1}^{M} P_j(f) = 1$, for all interesting frequencies.

4. Compute the eigen decomposition $\frac{1}{T}\mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{V}_p\mathbf{D}_p\mathbf{V}_p^{\mathrm{T}}$ and do partial whitening to a degree α:

$$\mathbf{\Lambda}_{\mathrm{m}} = \max\left(\mathbf{V}_p\mathbf{D}_p^{-\alpha/2}\mathbf{V}_p^{\mathrm{T}}\mathbf{P}\ ,\ 0\right)\ .$$

Each row of matrix $\mathbf{\Lambda}_{\mathrm{m}}$ defines one frequency mask $\boldsymbol{\lambda}_j$.

5. Implement topographic idea by, e.g., low-pass filtering the columns of $\mathbf{\Lambda}_{\mathrm{m}}$.

6. Calculate new source values: $\widehat{\mathbf{S}} = (\mathbf{S}_{\mathrm{dct}} \circ \mathbf{\Lambda}_{\mathrm{m}})\mathbf{V}_{\mathrm{dct}}$, where $\circ$ denotes element-wise multiplication.

vector of these elements defining the filter used for the j-th source is denoted here by $\boldsymbol{\lambda}_j$. We call each vector $\boldsymbol{\lambda}_j$ a *frequency mask*.

The first step of the denoising procedure is to compute the power spectra of the current source estimates. This gives an idea about the characteristic spectral properties of each source and suggests which frequencies should be emphasized by filtering. The next step is to calculate the individual frequency masks $\boldsymbol{\lambda}_j$ such that they are distinctive compared to each other. The intuition here is to make a coefficient $\lambda_{f,j}$ large if the frequency f is more prominent in s_j compared to the other sources. Correspondingly, $\lambda_{f,j}$ is made small if the frequency f is less prominent in s_j. Such a competition procedure naturally requires that all the sources are estimated simultaneously and the deflation approach is not applicable here. The competition mechanism is in practice implemented using rather heuristic principles and it is based on partial whitening the power spectra. This is somewhat similar to the whitening-based estimation of the source variances proposed by Valpola and Särelä (2004). The algorithm also uses some ideas similar to topographic ICA (Hyvärinen et al., 2001) in order to relax the competition in the power spectra of the neighboring sources. The final step of the denoising procedure is filtering the source estimates, as in Eq. (4.42), using the filters

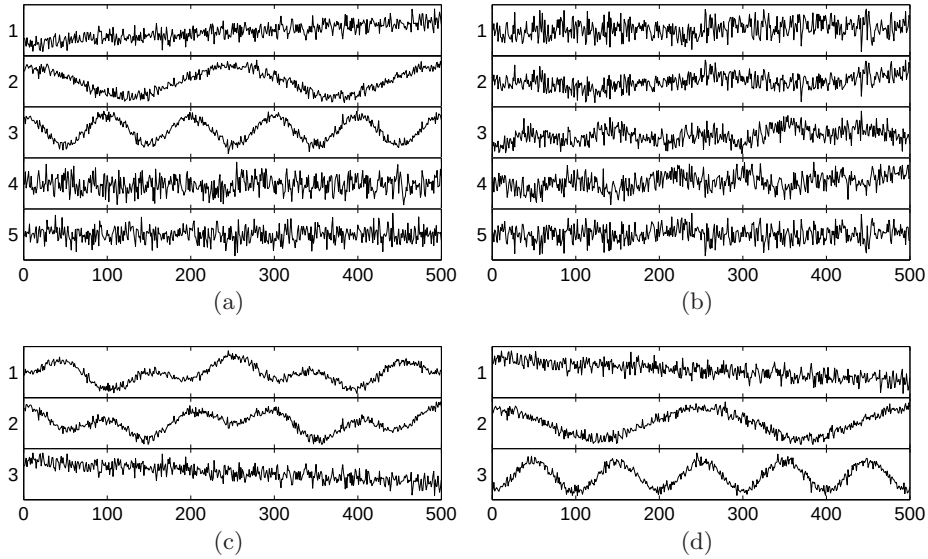


Figure 4.4: (a): Artificially generated sources three of which (sources 1, 2 and 3) have prominent variability in the slow timescale. (b): Observations generated as a linear mixture of the five sources. (c): Three components extracted by the clarity-based analysis with the emphasis on the prominent slow variability, where the period of slow spectral components is assumed to be longer than 80. (d): The result of the frequency-based rotation of components in (c).

defined by the estimated frequency masks.

Note that the proposed algorithm essentially performs separation in the frequency domain using an approach closely related to structured variances, which is discussed in Section 2.2.5.

Let us demonstrate an example of a frequency-based analysis using the two presented approaches. The test signals are generated by mixing linearly five sources, as shown in Fig. 4.4a,b. Sources 1–3 have prominent variability in the slow timescale while the other two signals are white Gaussian noise.

First, the clarity-based analysis is applied to extract three sources with the most prominent *slow* variability. The period of slow spectral components is chosen to be longer than 80. The extracted sources are shown in Fig. 4.4c to reconstruct the subspace of the original components 1–3. The first original component is reconstructed by source 3 due to its distinct clarity value. However, the original components 2 and 3 are still mixed in sources 1 and 2. These components cannot be separated using the clarity-based analysis as their clarity values are identical.

On the second stage, the frequency-based rotation is applied to the three extracted sources. It is easy to see from Fig. 4.4d that now the resulting components reconstruct the original components 1–3.

4.3.3 Independent dynamics subspace analysis

Frequency-based approach can find a meaningful representation of complex multidimensional data as it can separate different phenomena by the timescales of their prominent variations. This approach is not applicable, however, if the mixed phenomena have similar frequency contents. In this case, a combined time-frequency analysis (see, e.g., Särelä and Valpola, 2005) could be useful provided that interesting spectral components of different sources have distinct activation structures. However, the time-frequency analysis is difficult when the observation period is short compared to the timescale of the interesting data variations.

It is also possible that several components are related to the same phenomenon and their separation is not really possible. This might be the case, for example, in climate data, which is explained in **Publication 7**. Climate phenomena constantly interact with each other and cannot be independent. Most probably, they can be described by multidimensional dynamic processes and a meaningful separation criterion would be making the dynamics of different groups of sources as decoupled as possible.

Publication 8 presents a model called *independent dynamics subspace analysis* (IDSA) which implements the aforementioned assumptions. Now the sources are decomposed into groups as in Eq. (2.18). Each group $\mathbf{s}_k$ is assumed to be of known dimensionality and to follow an independent first-order nonlinear dynamic model:

$$\mathbf{s}_k(t) = \mathbf{g}_k(\mathbf{s}_k(t-1)) + \mathbf{m}_k(t)\,, \quad k = 1, \ldots, K\,, \tag{4.43}$$

where $\mathbf{g}_k$ is an unknown nonlinear function and $\mathbf{m}_k(t)$ accounts for modeling errors and noise. Assuming separate $\mathbf{g}_k$ in Eq. (4.43) means that the subspaces have decoupled dynamics, that is sources from one subspace do not affect the development of sources from other subspaces (see Fig. 4.5). In the linear case

$$\mathbf{s}(t) = \mathbf{B}\mathbf{s}(t-1) + \mathbf{m}(t)\,, \tag{4.44}$$

decoupled dynamics is equivalent to having a block-diagonal matrix $\mathbf{B}$ with nonzero blocks $\mathbf{B}_k$.

The IDSA model resembles linear dynamic factor analysis (DFA) considered by Särelä et al. (2001). The main difference is that the IDSA model requires all the sources be directly visible in the observations, which implies that they can be estimated using Eq. (4.4). The DFA model is more general as it permits sources which are important only for explaining the source dynamics and which cannot be identified as certain linear projections of the data.

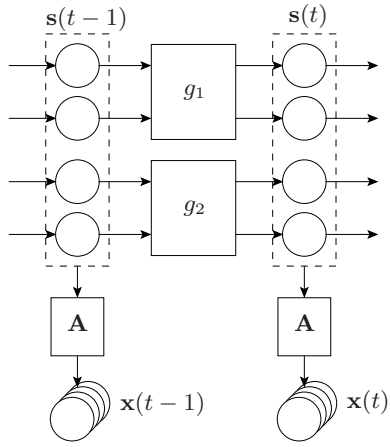


Figure 4.5: The model used in independent dynamics subspace analysis.

Without loss of generality, we can retain the assumption that all the sources are mutually uncorrelated and have unit variances. The sources from different subspaces are uncorrelated due to independence and the correlations within the subspaces can always be removed by a linear transformation (whitening). Note that IDSA identifies the sources only up to linear rotations within the subspaces, which is a known indeterminacy of multidimensional ICA (Cardoso, 1998).

Each subspace is estimated so as to minimize the prediction error of the corresponding subspace dynamic model in Eq. (4.43). Hence, the minimized objective function is

$$\mathcal{C} = \frac{1}{2} \sum_t \|\mathbf{s}_k(t) - \mathbf{g}_k(\mathbf{s}_k(t-1))\|^2 \,. \tag{4.45}$$

The source values are calculated using the separating structure in Eq. (4.4), and therefore

$$\mathbf{s}_k(t) = \mathbf{W}_k \mathbf{x}(t) \,, \tag{4.46}$$

where each row of the matrix $\mathbf{W}_k$ defines one source of the k-th subspace. The objective function in Eq. (4.45) should be optimized w.r.t. the nonlinear function $\mathbf{g}_k$ and the sources $\mathbf{s}_k(t)$ with the constraint that the demixing matrix is orthogonal. This can be done using the general algorithmic framework outlined in Fig. 4.2, as was explained in Section 4.2.3. Therefore, the corresponding denoising procedure alternately updates $\mathbf{g}_k$ and $\mathbf{s}_k(t)$ (see Table 4.2).

The nonlinearity $\mathbf{g}_k$ is updated so as to minimize the cost function in Eq. (4.45) keeping the current source estimates $\mathbf{s}_k(t)$ fixed. The exact implementation of

Table 4.2: Denoising procedure for independent dynamics subspace analysis

1. Update dynamics $\mathbf{g}_k$ so as to minimize $\mathcal{C}$ for current $\mathbf{s}_k(t)$.
2. Calculate the new source estimates $\widehat{\mathbf{s}}_k(t) = \mathbf{s}_k(t) - \mu\partial\mathcal{C}/\partial\mathbf{s}_k(t)$, where

$$\frac{\partial\mathcal{C}}{\partial\mathbf{s}_k(t)} = \mathbf{s}_k(t) - \mathbf{g}_k(\mathbf{s}_k(t-1)) - \left[\frac{\partial\mathbf{g}_k(\mathbf{s}_k(t))}{\partial\mathbf{s}_k}\right]^{\mathrm{T}}\Big[\mathbf{s}_k(t+1) - \mathbf{g}_k(\mathbf{s}_k(t))\Big]$$

with the following exceptions: when $t = 1$, the term $\mathbf{s}_k(t) - \mathbf{g}_k(\mathbf{s}_k(t-1))$ is omitted; and when $t = T$, the term $[\,\partial\mathbf{g}_k(\mathbf{s}_k(t))/\partial\mathbf{s}_k\,]^{\mathrm{T}}[\ldots]$ is omitted. The Jacobian matrix of $\mathbf{g}_k$ calculated at $\mathbf{s}_k(t)$ is denoted by $\partial\mathbf{g}_k(\mathbf{s}_k(t))/\partial\mathbf{s}_k$.

this step depends on the chosen mathematical model for $\mathbf{g}_k$. For example, the case of linear dynamics is trivial as minimizing the cost function w.r.t. the blocks $\mathbf{B}_k$ of the matrix $\mathbf{B}$ yields

$$\mathbf{B}_k = \mathbf{S}_{k,t+1}\mathbf{S}_{k,t}^{\dagger}\,, \tag{4.47}$$

where $\mathbf{S}_{k,t}$ and $\mathbf{S}_{k,t+1}$ are matrices whose columns contain the source values $\mathbf{s}_k(t)$ at times $t = 1,\ldots,T-1$ and $t = 2,\ldots,T$, respectively, and † denotes a pseudoinverse matrix.

In **Publication 8**, an MLP network is used to model $\mathbf{g}_k$:

$$\mathbf{g}_k(\mathbf{s}) = \mathbf{D}_k\boldsymbol{\phi}(\mathbf{C}_k\mathbf{s} + \mathbf{c}_k) + \mathbf{d}_k\,, \tag{4.48}$$

where $\mathbf{D}_k$, $\mathbf{C}_k$, $\mathbf{c}_k$, $\mathbf{d}_k$ are the parameters of the MLP and $\boldsymbol{\phi}$ is a sigmoidal function that operates component-wise on its inputs. The parameters of the MLP can be updated using the standard backpropagation procedure (see, e.g., Haykin, 1999). It should be noted that the solution for $\mathbf{g}_k$ should be regularized. If $\mathbf{g}_k$ is overfitted to the current source estimates $\mathbf{s}_k(t)$, yielding $\mathcal{C} = 0$, the algorithm stops in a degenerate solution.

The update of the sources $\mathbf{s}(t)$ is done using the gradient descent step similarly to Eq. (4.12). If the MLP model in Eq. (4.48) is used, the Jacobian matrix required for computing the gradient is given by

$$\partial\mathbf{g}_k(\mathbf{s})/\partial\mathbf{s}_k = \mathbf{D}_k\,\mathrm{diag}(\boldsymbol{\phi}'(\mathbf{C}_k\mathbf{s}_k + \mathbf{c}_k))\mathbf{C}_k\,, \tag{4.49}$$

where $\boldsymbol{\phi}'$ denotes the derivative of $\boldsymbol{\phi}$. Note that in practice the update of the dynamics $\mathbf{g}_k$ can be done more rarely than the update of the sources.

The independent subspaces can be estimated either symmetrically or one after another using deflation. The possibility to extract subspaces one by one provides a useful tool for extracting dynamically coupled components with the

most predictable time course from multivariate data. This is an important advantage compared to other methods where the model is learned for all data (e.g., Särelä et al., 2001), which can be very difficult for highly multidimensional and noisy data. Another important advantage of the proposed method is its computationally efficient learning algorithm, which is fast compared to the models estimated using the variational Bayesian approach.

Fig. 4.6 reproduces the experimental results reported in **Publication 8**. The artificial dataset is generated by mixing linearly three independent dynamic processes, two of which are Lorenz processes and one is a harmonic oscillator, and two white Gaussian noise signals. Five out of 10 observations are presented in Fig. 4.6. The algorithm is set to estimate symmetrically three independent subspaces: a two-dimensional subspace with linear dynamics and two three-dimensional subspaces with nonlinear dynamics. The recovered sources are shown in Fig. 4.6 to reconstruct the three subspaces of the original dynamic processes.

The current implementation of IDSA is based on the first-order autoregressive model for subspace dynamics. Including more time delays in the dynamic model can be useful when some of the subspace dimensions are not present in the data (like in the DFA model). However, one should be careful as introducing a higher-order memory to the dynamic model may cause the problem when the dynamics of *any* linear projection of the data can be perfectly modeled, which makes the separation of the subspaces impossible.

In practice, a frequency-based representation of data considered in Section 4.3.2 might be useful before performing IDSA. Slower components are generally easier to predict and the algorithm can favor them. Then, a good initialization is important for obtaining meaningful results. Therefore, it is preferable that all subspaces in the data would have the same timescale of prominent variations.

4.3.4 Extraction of components with structured variance

The previous sections considered algorithms for extracting prominent components with slowly changing time course. However, interesting slow behavior can be found in fast changing components as well. **Publication 9** introduces an algorithm which seeks fast components with prominent temporal structure of variances. The motivation of the proposed analysis comes from the inspection of the global weather measurements and the observation that fast weather variations have distinct yearly structure. This raises the question whether there are similar variations on slower timescales. The aim of the algorithm is to capture such prominent slow variability of the variances with the possibility to put emphasis on different timescales.

An assumption made in our analysis is that the interesting sources have nonstationary variances, that is their level of activation changes with time. More-

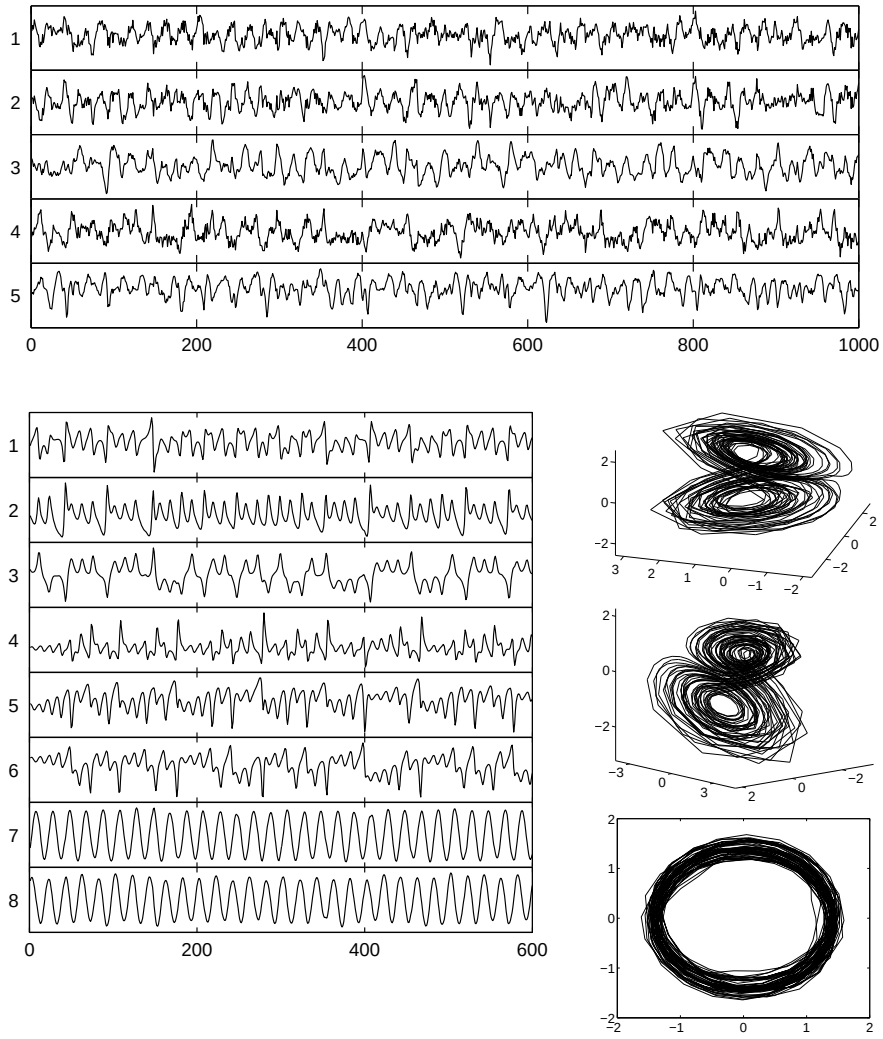


Figure 4.6: Above: Observations artificially generated as a linear mixture of three dynamic processes and two noise signals. Only five out of 10 observations used in the experiments are presented here. Below: The eight sources estimated by IDSA. The left plot shows the time series for the beginning of the observation period (first 600 samples). The plots on the r.h.s. are the phase curves of the three separated subspaces: subspaces of components 1–3, 4–6 and 7–8 (from top to bottom).

over, the variances of the sources have prominent temporal structure in a specific timescale. In the derivation of the algorithm, the source values $\{s(t)|t = 1, \ldots, T\}$ are regarded as a realization of a stochastic process $\{s_t\}$ consisting of random variables s_t. Note the difference in notations: $s(t)$ denotes a sample from the random variable s_t. The variables s_t are assumed Gaussian with zero mean and changing variances $v(t)$. We also define the *mean variance* of $\{s_t\}$ as

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} v(t)\,. \tag{4.50}$$

The following quantity is proposed to measure the amount of structure in each source:

$$\mathcal{F} = h(\nu) - h(s)\,, \tag{4.51}$$

where $h(s)$ denotes the (differential) entropy rate of $\{s_t\}$ and $h(\nu)$ is the entropy rate of a Gaussian process $\{\nu_t\}$ with i.i.d. zero-mean variables ν_t whose variances $\mathrm{E}\{\nu_t^2\}$ are stationary and equal to the mean variance of $\{s_t\}$ defined in Eq. (4.50). The Gaussian process with stationary variances has the highest entropy rate among all the processes with the same mean variance. Therefore, $\mathcal{F}$ is a good measure of non-stationarity, it is always nonnegative and it attains its minimum value of zero if and only if $\{s_t\}$ is a Gaussian process with stationary variances. The proposed measure resembles negentropy in Eq. (2.16) which is used as a measure of non-Gaussianity of a random variable.

The assumption that variances $v(t)$ have prominent variability in the known timescale helps estimate $v(t)$ from one realization of the stochastic process. Then, given a realization of length T, the quantity in Eq. (4.50) can be estimated as $\frac{1}{T}\sum_t v(t)$. The Gaussian variables s_t are assumed independent given $v(t)$ and therefore the entropy rate of $\{s_t\}$ can be estimated as

$$h(s) \approx \frac{1}{T} \sum_t H(s_t) = \frac{1}{T} \sum_t \frac{1}{2} \log 2\pi e v(t)\,, \tag{4.52}$$

where $H(s_t)$ denotes the entropy of s_t. This yields

$$\mathcal{F} = \frac{1}{2} \log \frac{1}{T} \sum_t v(t) - \frac{1}{T} \sum_t \frac{1}{2} \log v(t) \geq 0\,. \tag{4.53}$$

In practice, whitening makes $\frac{1}{T}\sum_t s^2(t) = 1$ for any source estimate, which allows for the assumption that

$$\frac{1}{T} \sum_t v(t) = 1\,. \tag{4.54}$$

This simplifies Eq. (4.53) to

$$\mathcal{F}_1 = -\frac{1}{T}\sum_t \frac{1}{2}\log v(t)\,. \tag{4.55}$$

The statistic $\mathcal{F}$ is a good measure of the structure which is related to non-stationarity of variances and has some connection to non-Gaussianity. The latter can be seen by noting from Eq. (4.54) that the variances $v(t)$ fluctuate around unity and therefore one can use the approximation $\log(1+\epsilon) \approx \epsilon - \frac{1}{2}\epsilon^2$. This yields from Eq. (4.55) the quantity

$$\mathcal{F}_2 \propto \frac{1}{T}\sum_t v^2(t) - 1 \tag{4.56}$$

which measures the magnitude of the variance fluctuations around the mean variance. For a process with stationary and unit variance, $\mathcal{F}_2$ equals zero. Now note that if the local variance $v(t)$ is approximated by $s^2(t)$, Eq. (4.56) gives the fourth moment of the random variable s. Such higher-order moments are often used for measuring non-Gaussianity (Hyvärinen et al., 2001).

In order to use the proposed measure, one needs to estimate the variances $v(t)$ of a signal at each time instant. This is usually done by estimating local sample variances because the variance is assumed to change slowly. We, however, want to concentrate on a *specific timescale* of variance variability and therefore we assume that the variance can be estimated in practice by filtering the squared signal values $s^2(t)$ such that only the interesting frequencies are preserved:

$$\mathbf{v}_{1..T} = \mathbf{F}\mathbf{s}^2_{1..T}\,. \tag{4.57}$$

Here, $\mathbf{v}_{1..T}$ is the vector of variances $v(t)$ and $\mathbf{s}^2_{1..T}$ is the vector made up from the squared source values $s^2(t)$, both defined similarly to Eq. (4.5), and $\mathbf{F}$ is the symmetric filtering matrix defined as in Eq. (4.41).

The measures $\mathcal{F}_1$ and $\mathcal{F}_2$ are functions of the variances $v(t)$ which are estimated from the sources $s(t)$ using Eq. (4.57). Thus, $\mathcal{F}_1$ and $\mathcal{F}_2$ are functions of $s(t)$ and can be maximized w.r.t. $s(t)$ by the gradient ascent method explained in Section 4.2.3. The required gradient can be approximated as

$$\frac{\partial \mathcal{F}}{\partial s(t)} \approx \frac{\partial \mathcal{F}}{\partial v(t)} s(t)\,, \tag{4.58}$$

which yields the denoising function

$$\widehat{s}(t) = g(v(t))s(t)\,, \tag{4.59}$$

where the nonlinearity g is given by

$$\text{for } \mathcal{F}_1: \quad g(v) \propto \beta - 1/v\,, \tag{4.60}$$

$$\text{for } \mathcal{F}_2: \quad g(v) \propto \beta + v\,, \tag{4.61}$$

and β is an arbitrary constant. The values $g(v(t))$ can be termed *masks* as they are applied to the current source estimates to get the new ones.

However, neither of the two nonlinearities is robust. The nonlinearity in Eq. (4.61) behaves well for small values of v but it gives too much weight to large v. This makes the algorithm very sensitive to outliers and very often results in overfitting (Hyvärinen et al., 1999). Note that $\mathcal{F}_2$ is related to higher-order moments which often suffer from this problem. In contrast, the nonlinearity in Eq. (4.60) saturates for large v but it is sensitive to small v where the gradient goes to infinity.

More robust algorithms can be derived by adjusting the nonlinearity g. For example, Eq. (4.60) could be transformed into

$$g(v) \propto \beta - \frac{1}{v+\alpha}\,, \tag{4.62}$$

where α accounts for the uncertainty of the local variance estimate $v(t)$. The exact shape of the nonlinearity g is usually not important and one can approximate Eq. (4.62) by another function which saturates for large v, for example, by

$$g(v) = \beta + \tanh(\alpha v)\,, \tag{4.63}$$

where α is a constant.

In general, the update rule in Eq. (4.59) with an arbitrary smooth g can be shown to maximize the following criterion:

$$\mathcal{F}_3 = \left(\frac{1}{T}\sum_t G(v(t)) - G(1)\right)^2\,, \tag{4.64}$$

where $g(v) = \partial G(v)/\partial v$. Note that $\mathcal{F}_3$ with $G(v) = v^2$ and $\mathcal{F}_2$ defined in Eq. (4.56) are maximized at the same points. To decrease overfitting, G can be chosen to be a function growing slower than v^2. For example, $G(v) = \log\cosh v$ yields $g(v) = \tanh(v)$. Note that the measure in Eq. (4.64) bears some similarity to the approximation of negentropy used by Hyvärinen (1999a).

The outline of the denoising procedure is presented in Table 4.3. It starts with estimating the local variances using Eq. (4.57). Then, the nonlinearity g is applied to the variance estimates in order to calculate the masks. In practice, we have used the nonlinearities defined in Eqs. (4.63) and (4.61). In order to emphasize the dominant signal activations, the constant β was chosen such that

Table 4.3: Denoising procedure for ICA with structured variance

1. Calculate the variance estimates as $\mathbf{v}_{1..T,j} = \mathbf{F}\mathbf{s}^2_{1..T,j}$
2. Compute the masks $\mathbf{m}_j = g(\mathbf{v}_j)$
3. Shift the mask: $\mathbf{m}_j = \mathbf{m}_j - \min_t m_j(t)$
4. Calculate the new source estimates $\widehat{s}_j(t) = m_j(t)s_j(t)$

the minimum values of the masks are put to zero. This does not change the fixed points of the algorithm but speeds up convergence. Finally, the denoised source estimates are calculated by applying the mask to the current source values.

The proposed algorithm can be modified for subspace analysis where several sources are assumed to share the same variance structure. In this case, the subspace activation can be estimated on Step 1 by taking the average of the squared sources from the same subspace:

$$\mathbf{v}_{1..T} = \mathbf{F}\left(\frac{1}{K}\sum_{j=1}^{K}\mathbf{s}^2_{1..T,j}\right). \tag{4.65}$$

Then, the same mask calculated from $\mathbf{v}_{1..T}$ is applied to each component from the corresponding subspace.

In **Publication 9**, we present an example of applying the proposed algorithm to artificial data. The example shows that focusing on a specific timescale of the variance variability helps extract the most relevant components from data. In blinder settings, the method can be used as a tool for exploratory data analysis. Different interesting phenomena can be found in the same dataset by concentrating on different timescales. The focus of the analysis is changed by simply using another filter in the variance estimation. The results of such exploratory analysis for climate data are presented in **Publication 9**. The emphasis on a properly chosen timescale can also be important for solving the BSS problem as it can improve the separation results, especially for noisy data when other separation criteria cannot provide reliable components.

4.4 Application to climate data analysis

4.4.1 Extraction of patterns of climate variability

One of the main goals of statistical analysis of climate data is to extract physically meaningful patterns of climate variability from highly multivariate weather

measurements. The classical technique for defining such dominant patterns is PCA, or empirical orthogonal functions (EOF), as it is called in climatology (see, e.g., von Storch and Zwiers, 1999). However, the maximum remaining variance criterion used in PCA can lead to such problems as mixing different physical phenomena in one extracted component (Richman, 1986). This makes PCA a useful tool for information compression but limits its ability to isolate individual modes of climate variations.

To overcome this problem, rotation of the principal components has proven useful. The classical rotation criteria used in climatology are based on the general concept of "simple structure" which can provide spatially or temporally localized components (Richman, 1986). Independent component analysis is a technique which can also be used for the rotation of principal components (Aires et al., 2002). The criterion used by ICA is the assumption of the statistical independence of the components. Even though ICA can sometimes give a meaningful representation of weather data (see, e.g., Aires et al., 2000; Lotsch et al., 2003; Basak et al., 2004), the statistical independence is quite a restrictive assumption which can often lead to naive solutions.

In the algorithmic framework of DSS, it is easy to implement various rotation criteria. One can efficiently incorporate prior knowledge about the interesting properties of the sources of data variability. The motivation for seeking a particular type of components can come from general statistical principles (e.g., maximizing non-Gaussianity of components gives the ICA solution), expert knowledge (e.g., some information about the spectral structure of components), or based on some elementary inspection of data (e.g., by observing some regular patterns in them). For example, in the climate data analysis we might be interested in some phenomena that would have prominent variability in a certain timescale or exhibit slow changes. Thus, DSS presents a powerful tool for *exploratory analysis* of large spatio-temporal climate datasets. In **Publications 5**, **6**, **7** and **9**, we present several algorithms designed in this algorithmic framework and apply them to global long-term climate measurements.

4.4.2 Climate data and preprocessing method

In the publications of this thesis, measurements of three major atmospheric variables are analyzed. The considered set of variables includes surface temperature, sea level pressure and precipitation and it is often used for describing global climate phenomena such as El Niño–Southern Oscillation (ENSO) (Trenberth and Caron, 2000). The datasets are provided by the reanalysis project of the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP/NCAR) (Kalnay et al., 1996; NCEP data, 2004). The data represent globally gridded daily measurements over a long period of time. The

spatial grid is regularly spaced over the globe with $2.5° \times 2.5°$ resolution.

The reanalysis data is not fully real because the missing measurements have been reestimated based on the available data and approximation models. Yet, the data is as close to the real measurements as possible. Although the quality of the data varies over time and spatial location, we used the whole period of 1948–2004 and the whole global grid. Thus, the data contain more than 10,000 spatial locations and about 20,000 time instances.

To preprocess the data, the long-term mean was removed and the data points were weighted to diminish the effect of a denser sampling grid around the poles: each data point was multiplied by a weight proportional to the square root of the corresponding area of its location. The spatial dimensionality of the data was then reduced using the PCA/EOF analysis applied to the weighted data. We retained 100 principal components which explain more than 90% of the total variance, which is due to the high spatial correlation between nearby points on the global grid. In Publication 9, where fast changing phenomena are of interest, the principal components are additionally preprocessed by high-pass filtering.

4.4.3 Clarity-based extraction of slow components

Publications 5, **6** and **7** concentrate on slowly changing sources of climate variability. The clarity-based analysis presented in Section 4.3.1 is applied to extract components exhibiting the most prominent variability in a specific timescale. In **Publication 5**, the components with the most prominent *interannual variability* are found to be related to the well-known ENSO phenomenon. For all three datasets that were tested, the time course of the most prominent component provides a good ENSO index and the corresponding spatial patterns contain many features traditionally associated with ENSO. Several other components with prominent interannual structures are extracted as well. For example, the second component extracted from the dataset combining the three variables resembles the derivative of the first component. Thus, it is likely to be related to ENSO as well. The time courses and the spatial patterns of the two most prominent component extracted from the combined dataset are reproduced in Figs. 4.7 and 4.8, respectively.

4.4.4 Frequency-based separation of slow components

Publications 6-7 extend the analysis of slow climate variations to a wider frequency range. First, the slow subspace of the climate system is identified using the clarity-based approach applied to the combined measurements of the three variables. Then, the found slow components are separated based on their frequency contents using the algorithm from Section 4.3.2. Preliminary results of

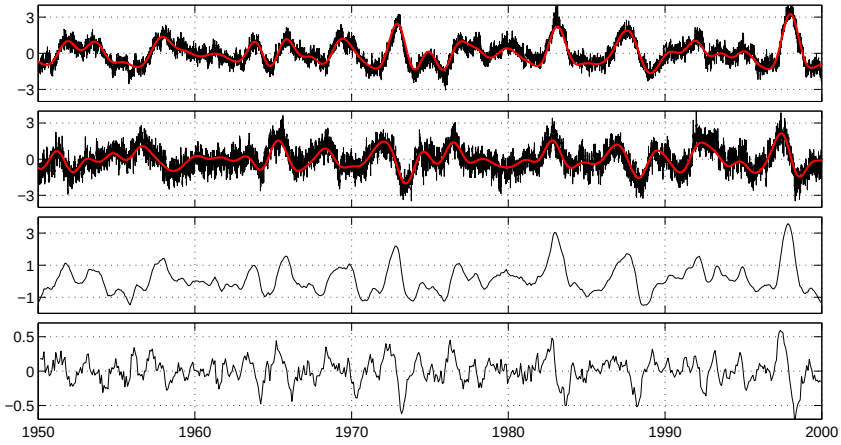


Figure 4.7: The dark curves on the two upper plots are the time courses of the two components with the most prominent interannual variability. They are extracted from the dataset combining surface temperature, sea level pressure and precipitation. The red curves are the same components after filtering in the interannual timescale. The two lower plots present the index which is used in climatology to measure the strength of El Niño (above) and its derivative (below).

this analysis are reported in **Publication 6**, and somewhat improved results are presented in **Publication 7**.

The extracted components turn out to represent the subspace of the slow climate phenomena as a linear combination of trends, decadal-interannual quasi-periodic signals, the annual cycle and other phenomena with distinct spectral contents. Using this approach, the known climate phenomena are identified as certain subspaces of the climate system and some other interesting phenomena hidden in the weather measurements are found.

Figs. 4.9–4.11 reproduce the surface temperature and sea level patterns of some of the 16 slow components reported in Publication 7. Only the components with prominent loadings around the poles are presented here.

4.4.5 Components with structured variance

In **Publication 9**, the algorithm presented in Section 4.3.4 is used in order to extract fast changing components whose variances have prominent temporal structure. When we concentrate on the dominant, annual variance variations, two

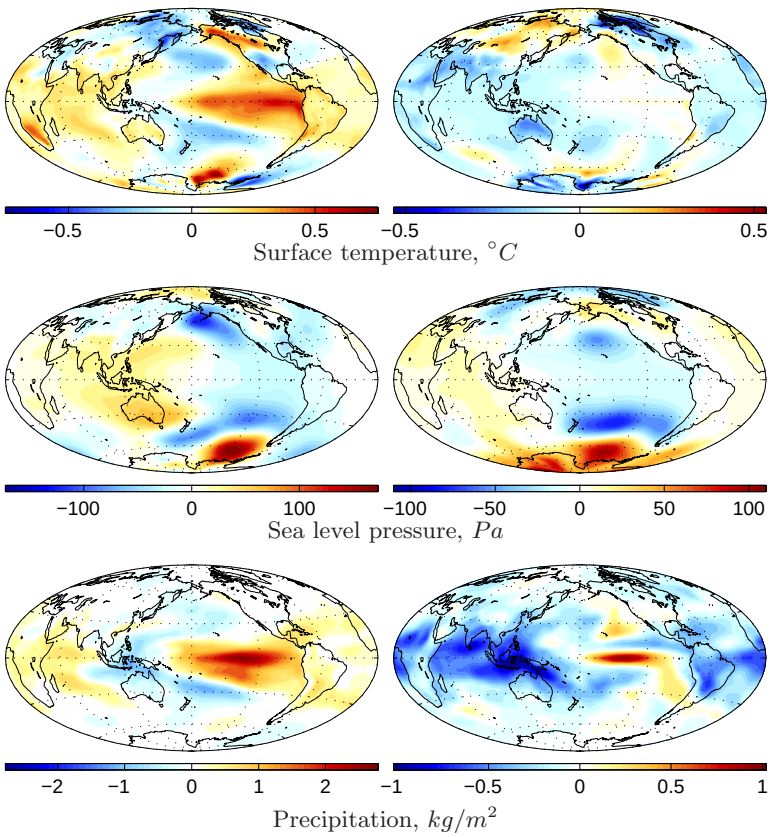


Figure 4.8: The spatial patterns corresponding to the first (left column) and second (right column) components with the most prominent interannual variability. The maps tell how strongly the component is expressed in the measurement data.

subspaces with different phases of the yearly activations are extracted. The first subspace explains the fast temperature variability in the Northern Hemisphere and has higher activations during Northern Hemisphere (NH) winters. The second subspace corresponds to the fast oscillations in the Southern Hemisphere with higher activations during NH summers.

In the second experiment, we concentrate on the slower, decadal timescale of the fast temperature variations. Several components with prominent temporal and spatial structures are extracted. Fig. 4.12 reproduces the temporal patterns of some of the components found in the data. The prominent slow structure of the variance emerge very clearly in the extracted components.

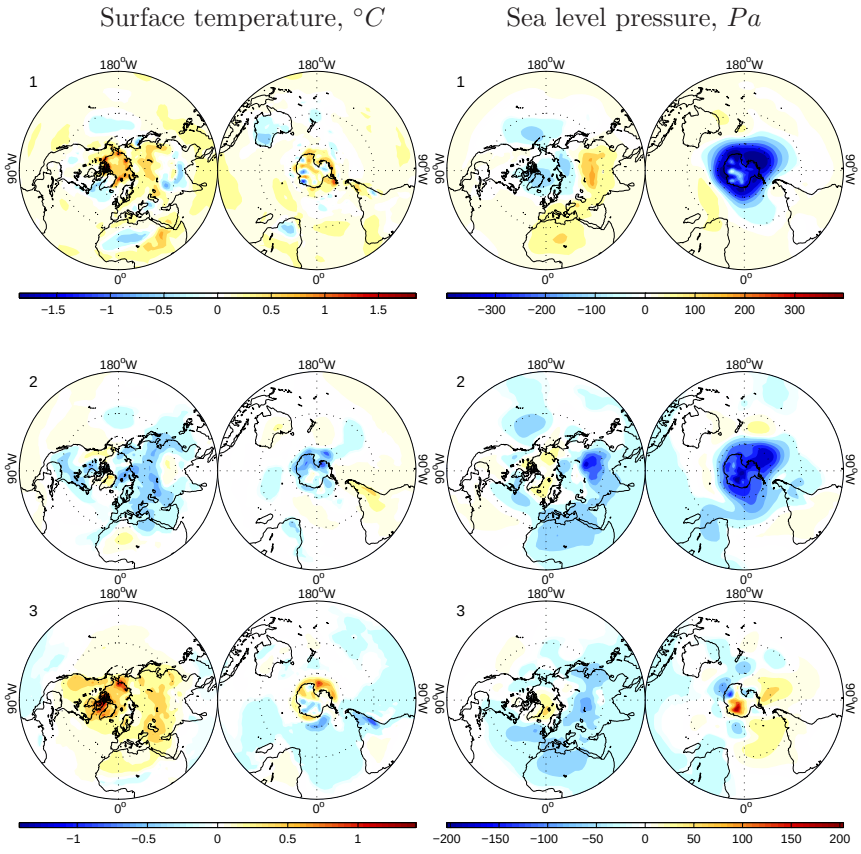


Figure 4.9: The spatial patterns of components 1–3 (trends) found by frequency-based rotation of the 16 most prominent slow components.

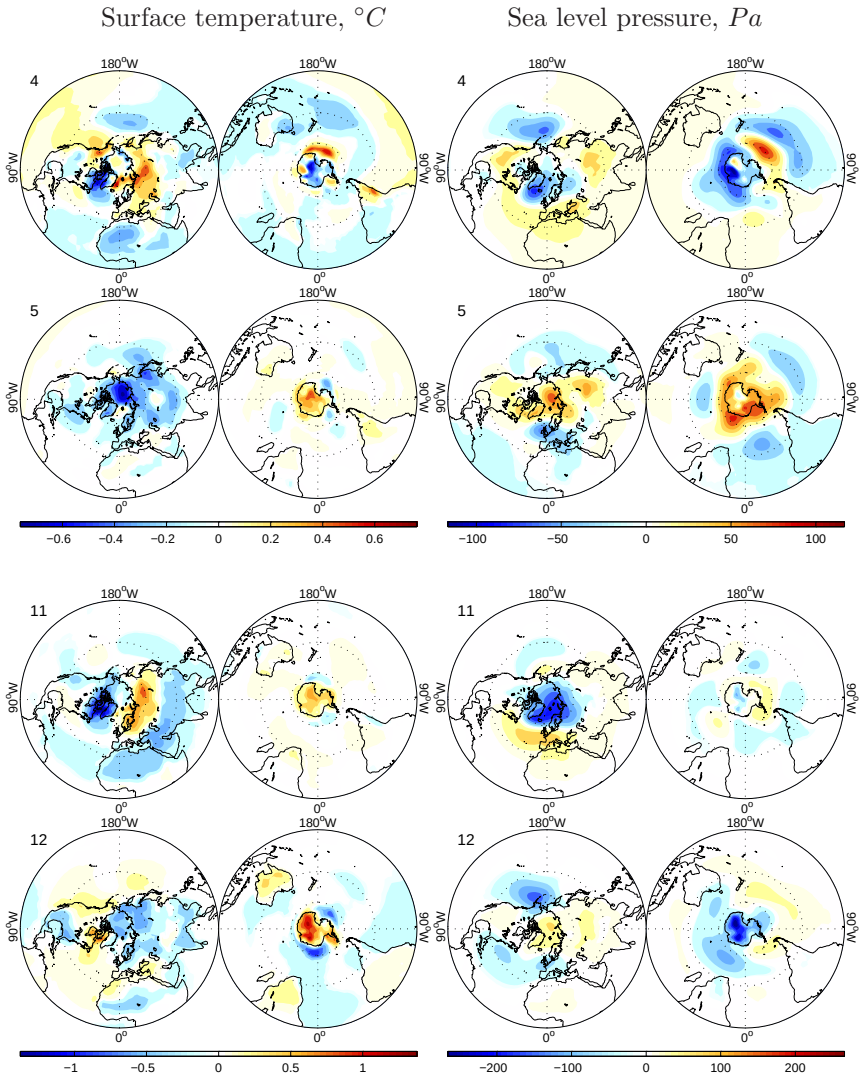


Figure 4.10: The spatial patterns of components 4–5 (trends) and components 11–12 (prominent slow and annual frequencies) found by frequency-based rotation of the 16 most prominent slow components.

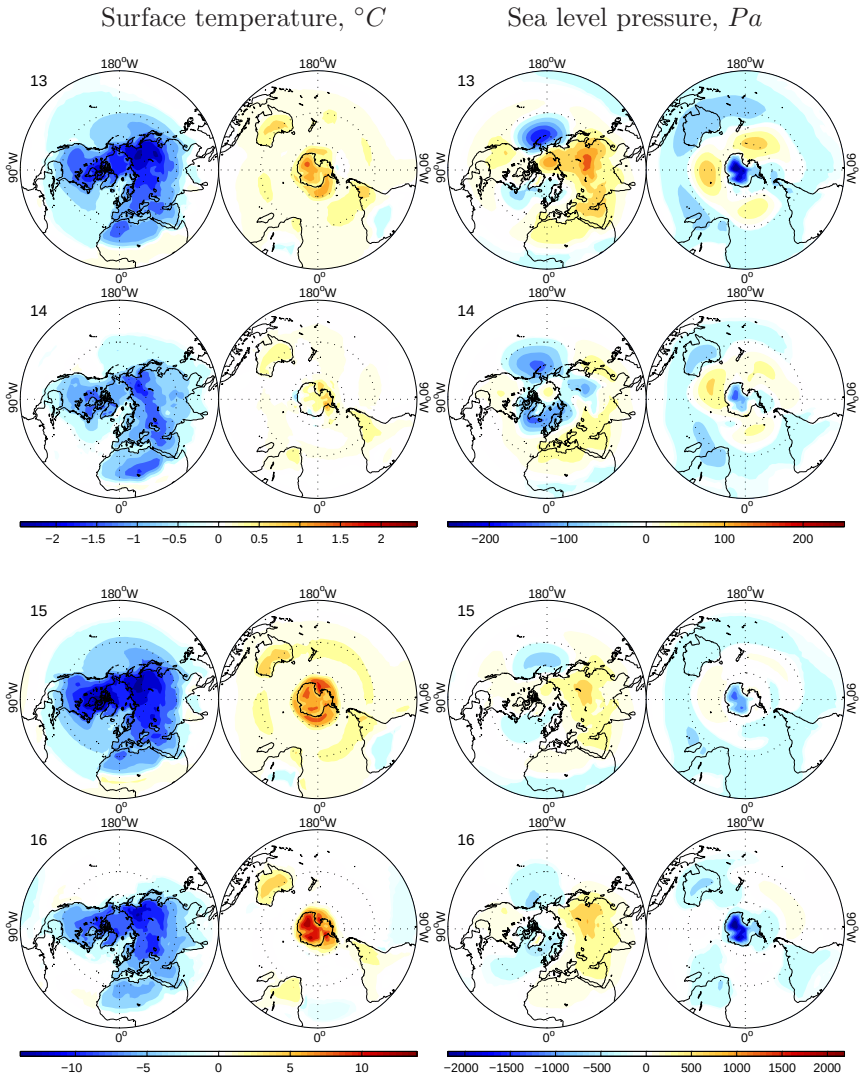


Figure 4.11: The spatial patterns of components 13–14 (prominent close-to-annual oscillations) and components 15–16 (the annual cycle) found by frequency-based rotation of the 16 most prominent slow components.

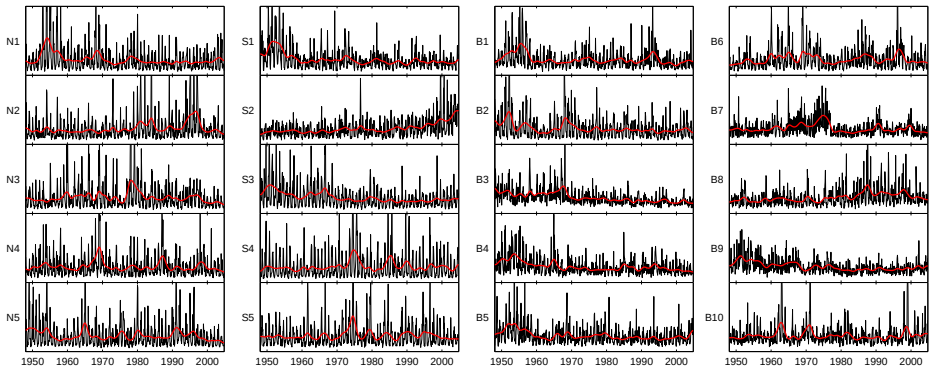


Figure 4.12: The temporal patterns of the fast components extracted from surface temperature measurements (black). The red curves emphasize the prominent slow structure.

4.4.6 Discussion and future directions

This section presented some results of exploratory analysis of global weather measurements using several algorithms which follow the algorithmic framework of denoising source separation. The obtained results are very promising but the meaning of the results needs to be further investigated, as some of the found components may correspond to significant climate phenomena while others may reflect some artifacts produced during the data acquisition. A third alternative would be that the components may have been overfitted to the data. In some of the experiments, for example, in the extraction of components with structured variance, some of the results looked like typical overfits. To be sure, the reliability of the results could be tested by cross-validation.

The results of the analysis open up many possible directions for future research. The results on prominent slow climate variability presented in **Publications 5–7** suggest that there might be phenomena that could be described by multidimensional processes with complex nonlinear dynamics. This makes the IDSA model presented in **Publication 8** very promising in this application. The fact that there are climate phenomena like ENSO which can be observed in different weather variables (such as temperature, air pressure, precipitation) raises the question whether there are other climate phenomena like that. It might be that such phenomena manifest themselves in more complicated ways in the observables and could be extracted using more complex (nonlinear, hierarchical) models.

The results on prominent variance structures reported in **Publication 9** indicate what kind of features could be found in the fast climate variations when the

emphasis is put on different timescales. The presented analysis of the variance structures can be extended in many different ways. For example, it would be interesting to relate the components with prominent variance structures to the known climate phenomena visible as specific projections of global weather data. It would also be possible to use more information for more robust variance estimation. The additional information could be in the form of other components extracted from climate data or a hierarchical variance model (Valpola et al., 2004).

The presented algorithms can easily be applied to other weather measurements with the possibility to concentrate on various properties of interest, different timescales and spatial localizations. It is also possible that some new interesting properties emerge during such exploratory analysis. This could motivate other types of models and algorithms, and the algorithmic framework used in this chapter can be a useful tool.

4.5 Conclusions

In this chapter, faster source separation algorithms based on the linear mixing model have been considered. The presented algorithms have been implemented following the unifying algorithmic framework of denoising source separation. This framework allows for fast development of semiblind algorithms which use available prior knowledge in the separation process. Thus, the framework provides a useful tool for exploratory data analysis.

The general algorithmic framework has been presented in the beginning of this chapter. It includes the preprocessing step called whitening followed by rotation using an orthogonal demixing matrix. This matrix is found so as to optimize the signal properties that are known from the prior information. This is generally done using an iterative procedure in which the desired (interesting) properties are emphasized by means of a denoising function. In the special case when the interesting part of a signal is obtained by linear temporal filtering, the whole procedure can be reduced to three simple steps: whitening, filtering and PCA. The presented exposition of the algorithmic framework shows the connection between the denoising function and the measure of structure that is optimized either implicitly or explicitly.

The approaches considered in this chapter have some connection to Bayesian methods studied in Chapter 3. It has been shown that approximate Bayesian methods applied to source separation problems can often be implemented in the considered algorithmic framework. For example, a simple Bayesian model with super-Gaussian sources was shown to correspond to using a shrinkage function in the denoising procedure. A Bayesian interpretation of the FastICA algorithm was also presented.

After the general introduction to the used framework, the algorithms proposed in this thesis have been presented. Two algorithms perform separation of signals based on their spectral contents. A simple algorithm, that was presented first, focuses on a specific timescale of prominent signal variations. In the second algorithm, this approach was extended to blinder case when sources are separated by making their frequency contents as distinctive as possible. The model called independent dynamics subspace analysis considers the case when a group of sources may share a common dynamic model. The proposed algorithm performs separation of the different groups by decoupling explicitly their dynamic models. The approach presented last allows for finding components with prominent variance structures. The proposed algorithm can easily be tuned to concentrate on different timescales of variance variations.

The last part of this chapter presents several results on applying some of the proposed algorithms to exploratory analysis of climate data. In fact, the proposed algorithms were largely motivated by this particular application. Some of the components extracted from global climate data with the proposed techniques have evident and meaningful interpretations, while other results may require some further investigations.

Chapter 5

Conclusions

Latent variable models are important tools for statistical analysis of spatio-temporal datasets. Using these models, it is possible to capture basic data regularities or to find interesting and meaningful patterns hidden in the data. LVMs with meaningful interpretations can be learned by source separation methods which assume that the hidden variables correspond to some significant sources generating the data. Independence of the processes reflected in the source signals is the typical assumption used by the methods of this kind.

This thesis considered several source separation models and different approaches to their estimation. The first half considered Bayesian estimation methods which describe all the unknown variables using probability distributions. The main focus has been variational Bayesian methods based on approximating complex posterior distributions using simpler and tractable distributions. Three basic results include a study of the effect of the posterior approximation, a new model for solving post-nonlinear ICA problems, and the application of the nonlinear dynamic factor analysis approach to the problem of state change detection.

The first result is a theoretical and experimental study of the properties of VB methods using linear ICA models. It shows that the solution provided by VB methods is always a compromise between the accuracy of the model (i.e., a good explanation of data) and the accuracy of the posterior approximation. This may be a negative effect as too simple posterior approximations may introduce a significant bias in favor of some types of solutions. This problem can be overcome by either modeling posterior correlations or by applying suitable preprocessing. Otherwise the found solution may not be meaningful. Sometimes, however, it is possible to use this effect to regularize otherwise degenerate solutions.

Another important result is the application of the VB approach to the model called post-nonlinear factor analysis. The model is a special case of the general

NFA with a restriction that the generative mapping has a special, post-nonlinear structure. The proposed technique can be used in post-nonlinear ICA problems and it can overcome some of the limitations of the existing alternative methods.

The thesis presents a study of the nonlinear dynamic factor analysis presented by Valpola and Karhunen (2002) in which the VB approach is applied to estimation of nonlinear state-space models. In the introductory part of this thesis, it has been shown that the NDFA algorithm can be considered a source separation method as it can find representations with dynamically decoupled subspaces. In Publication 2, the NDFA algorithm has been applied to the problem of detecting changes in complex dynamic processes. The VB cost function provided by the NDFA algorithm was used to calculate the estimate of the process entropy rate. This estimate was proposed to be taken as the indicator of change. The extensive experimental study has shown that the proposed approach can outperform greatly other alternative techniques applicable to the change detection problem.

The second half of this thesis considered faster source separation algorithms which use point estimates for the unknown parameters. Several algorithms assuming the linear mixing model have been proposed. The algorithms were largely motivated by the analysis of the highly-multidimensional spatio-temporal datasets containing daily weather measurements all over the globe for a period of 56 years. The algorithms follow the unifying algorithmic framework of denoising source separation. Three basic approaches to source separation have been used: frequency-based analysis, separation by decoupling dynamic models and extraction of components with structured variances.

The frequency-based analysis aims to find components with prominent spectral contents. The first algorithm concentrates on a specific timescale of data variations and extracts components in which such variations are most prominent. When applied to the global climate data with concentration on the interannual timescale, the first extracted components were clearly related to the El Niño–Southern Oscillation phenomenon. The first component extracted from surface temperature, sea level pressure and precipitation data provided a good ENSO index, and the second component somewhat resembled the derivative of the first one. The second frequency-based algorithm extends the previous approach to the more general case when the sources are estimated so as to make their frequency contents as distinctive as possible. The application of the technique to the global climate data turned out to give a meaningful representation of the slow climate variability as a combination of slowest trends, interannual quasi-periodical signals, the annual cycle and components which slowly modify the seasonal variations. Several components which might be related to ENSO emerged in the results. This fact suggests that there might exist complex climate phenomena which could be described by a group of components, and such groups of components could have a predictable time course.

Another technique proposed in this thesis is called independent dynamics subspace analysis. Its model takes into account the assumptions motivated by the results obtained by the application of the frequency-based analysis to the climate data. The sources are decomposed into groups and each group is assumed to share a common dynamic model. An efficient algorithm for learning this model has been proposed. It is much faster than alternative methods, for example, based on the VB principles. The proposed model is rather general and could be used in different applications for finding groups of the most predictable components.

The third approach considered in the second half is the analysis of components based on their variance structures. The algorithm that can extract components with prominent variance variations in a specific timescale has been proposed. It was derived as an approximate algorithm for optimizing a measure of non-stationarity which somewhat resembles negentropy. The results obtained for the global climate data contained some remarkable patterns both in spatial localization and in time courses. This result suggests that the algorithm can potentially extract components that would correspond to meaningful climate phenomena.

There are many open research questions related to the results presented in this thesis. For example, the proposed Bayesian post-nonlinear model could be improved by using a more complex model for the hidden variables. Using the mixture model similar to independent factor analysis (Attias, 1999) could potentially improve the quality of the source reconstruction. The effect of the posterior approximation in this type of nonlinear models could be investigated in more details. Modeling posterior correlations of the sources may be required in order to diminish the bias introduced by simple approximations. Improved approximation techniques (e.g., similar to the ideas presented by Barber and Bishop, 1998) could be useful in this problem.

An important line of future research is application of the proposed techniques to real-world problems. For example, the change detection approach based on variational Bayesian learning could be applied to real process monitoring tasks. The faster algorithms presented in the second half of this thesis could be useful for analysis of other types of spatio-temporal datasets (e.g., biomedical data). These algorithms could easily be modified in order to capture interesting data properties which might emerge in a specific application. Hierarchical and nonlinear extensions (e.g., similar to Wiskott and Sejnowski, 2002) of the faster algorithms might be useful as well.

The presented analysis of the global climate data can be continued in many ways. Some ideas were outlined in the discussion of Section 4.4.6. The important directions include investigation of the meaning of the found components, a potential application of the proposed subspace model with independent dynamics, nonlinear extensions of the proposed techniques and finding out the relations between components with different kinds of prominent structures.

Bibliography

Aires, F., Chédin, A., and Nadal, J.-P. (2000). Independent component analysis of multivariate time series: Application to the tropical SST variability. *Journal of Geophysical Research*, 105(D13):17437–17455.

Aires, F., Rossow, W. B., and Chédin, A. (2002). Rotation of EOFs by the independent component analysis: Toward a solution of the mixing problem in the decomposition of geophysical time series. *Journal of the Atmospheric Sciences*, 59(1):111–123.

Almeida, L. B. (2003). MISEP – linear and nonlinear ICA based on mutual information. *Journal of Machine Learning Research*, 4 (Dec):1297 – 1318.

Almeida, L. B. (2005). Separating a real-life nonlinear image mixture. *Journal of Machine Learning Research*, 6:1199–1232.

Almeida, L. B. (2006). *Nonlinear Source Separation*. Synthesis Lectures on Signal Processing. Morgan and Claypool Publishers.

Amari, S.-I. and Cardoso, J.-F. (1997). Blind source separation – semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700.

Amari, S.-I., Cichocki, A., and Yang, H. (1996). A new learning algorithm for blind signal separation. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, Cambridge, MA, USA.

Amari, S.-I. and Nagaoka, H. (2000). *Methods of Information Geometry*. American Mathematical Society, Providence.

Attias, H. (1999). Independent factor analysis. *Neural Computation*, 11(4):803–851.

Attias, H. (2000a). Independent factor analysis with temporally structured sources. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 386–392. MIT Press, Cambridge, MA, USA.

Attias, H. (2000b). A variational Bayesian framework for graphical models. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, Cambridge, MA, USA.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.

Barber, D. and Bishop, C. (1998). Ensemble learning for multi-layer networks. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems 10*, pages 395–401. The MIT Press, Cambridge, MA, USA.

Barros, A. K. and Cichocki, A. (2001). Extraction of specific signals with temporal structure. *Neural Computation*, 13(9):1995–2003.

Basak, J., Sudarshan, A., Trivedi, D., and Santhanam, M. S. (2004). Weather data mining using independent component analysis. *Journal of Machine Learning Research*, 5:239–253.

Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes: Theory and Application*. Information and system science series. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Beal, M. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University of London, UK.

Beal, M. J. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics 7*, pages 453–464.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159.

Belouchrani, A., Meraim, K. A., Cardoso, J.-F., and Moulines, E. (1997). A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press.

Bishop, C. (1999a). Latent variable models. In Jordan, M., editor, *Learning in Graphical Models*, pages 371–403. The MIT Press, Cambridge, MA, USA.

Bishop, C. M. (1999b). Variational principal components. In *Proceedings of the 9th International Conference on Artificial Neural Networks, (ICANN '99)*, pages 509–514.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1995). EM optimization of latent variable density models. In *Advances in Neural Information Processing Systems 8*, pages 465–471. MIT Press, Cambridge, MA.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10:215–234.

Blaschke, T. and Wiskott, L. (2004). Independent slow feature analysis and nonlinear blind source separation. In Puntonet, C. G. and Prieto, A., editors, *Proceedings of the Fifth International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, volume 3195 of *Lecture Notes in Computer Science*, pages 742–749. Springer-Verlag, Berlin.

Briegel, T. and Tresp, V. (1999). Fisher scoring and a mixture of modes approach for approximate inference and learning in nonlinear state space models. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems 11*, pages 403–409. The MIT Press, Cambridge, MA, USA.

Cardoso, J.-F. (1989). Source separation using higher order moments. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP '89)*, pages 2109–2112.

Cardoso, J.-F. (1997). Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114.

Cardoso, J. F. (1998). Multidimensional independent component analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, pages 1941–1944, Seattle, WA.

Cardoso, J.-F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192.

Cardoso, J.-F. and Laheld, B. H. (1996). Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030.

Chan, K., Lee, T.-W., and Sejnowski, T. J. (2002). Variational learning of clusters of undercomplete nonsymmetric independent components. *Journal of Machine Learning Research*, 3:99–114.

Chan, K., Lee, T.-W., and Sejnowski, T. J. (2003). Variational Bayesian learning of ICA with missing data. *Neural Computation*, 15(8):1991–2011.

Chen, J. and Patton, R. J. (1999). *Robust Model-Based Fault Diagnosis for Dynamic Systems*. Kluwer Academic Publishers, Boston/Dordrecht/London.

Cherkassky, V. and Mulier, F. (1998). *Learning from Data: Concepts, Theory, and Methods*. Information and system science series. John Wiles & Sons.

Chiang, L. H., Russell, E. L., and Braatz, R. D. (2001). *Fault Detection and Diagnosis in Industrial Systems*. Springer-Verlag, London.

Choi, S., Cichocki, A., and Belouchrani, A. (2002). Second order nonstationary source separation. *Journal of VLSI Signal Processing*, 32(1-2):93–104.

Choudrey, R. A. and Roberts, S. J. (2001). Flexible Bayesian independent component analysis for blind source separation. In *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA 2001)*, pages 90–95, San Diego, USA.

Choudrey, R. A. and Roberts, S. J. (2003). Variational mixture of Bayesian independent component analyzers. *Neural Computation*, 15(1):213–252.

Cichocki, A. and Amari, S.-I. (2002). *Adaptive Blind Signal and Image Processing*. John Wiley & Sons.

Cichocki, A. and Belouchrani, A. (2001). Source separation of temporally correlated sources using bank of band-pass filters. In *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA 2001)*, pages 173–178, San Diego, USA.

Cichocki, A., Rutkowski, T., and Siwek, K. (2002). Blind signal extraction of signals with specified frequency band. In *Neural Networks for Signal Processing XII: Proceedings of the 2002 IEEE Signal Processing Society Workshop*, pages 515–524, Martigny, Switzerland.

Cichocki, A. and Thawonmas, R. (2000). On-line algorithm for blind signal extraction of arbitrarily distributed, but temporally correlated sources using second order statistics. *Neural Processing Letters*, 12:91–98.

Cichocki, A. and Unbehauen, R. (1996). Robust neural networks with on-line learning for blind identification and blind separation of sources. *IEEE Transactions on Circuits and Systems*, 43(11):894–906.

Comon, P. (1994). Independent component analysis – a new concept? *Signal Processing*, 36:287–314.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons.

Darmois, G. (1951). Analyse des liaisons de probabilité. In *Proceedings of International Statastics Conferences 1947*, volume IIIA, page 231, Washington, D.C.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38.

Diamantaras, K. I. and Kung, S. Y. (1996). *Principal Component Neural Networks: Theory and Applications*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons.

Doucet, A., de Freitas, N., and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.

Feller, W. (1968). *Probability Theory and Its Applications*. Wiley.

Frey, B. J. and Hinton, G. E. (1999). Variational learning in nonlinear Gaussian belief networks. *Neural Computation*, 11(1):193–214.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, Florida.

Ghahramani, Z. and Beal, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press, Cambridge, MA, USA.

Ghahramani, Z. and Beal, M. J. (2001). Graphical models and variational methods. In Saad, D. and Opper, M., editors, *Advanced Mean Field Methods - Theory and Practice*. MIT Press, Cambridge, MA, USA.

Ghahramani, Z. and Hinton, G. E. (1996). Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, Department of Computer Science, University of Toronto.

Ghahramani, Z. and Hinton, G. E. (1998). Hierarchical non-linear factor analysis and topographic maps. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*, pages 486–492. The MIT Press, Cambridge, MA, USA.

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural Computation*, 12(4):831–864.

Gharieb, R. R. and Cichocki, A. (2003). Second-order statistics based blind source separation using a bank of subband filters. *Digital Signal Processing*, 13(2):252–274.

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532.

Green, A. A., Berman, M., Switzer, P., and Craig, M. D. (1988). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74.

Grewal, M. S. and Andrews, A. P. (1993). *Kalman filtering: Theory and Practice*. Information and system science series. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Gustafsson, F. (2000). *Adaptive Filtering and Change Detection*. John Wiley & Sons.

Harman, H. H. (1960). *Modern Factor Analysis*. The University of Chicago Press.

Harmeling, S., Ziehe, A., Kawanabe, M., and Müller, K.-R. (2003). Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124.

Harva, M. and Kabán, A. (2005). A variational Bayesian method for rectified factor analysis. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2005)*, pages 185–190, Montréal, Canada.

Haykin, S. (1999). *Neural Networks – A Comprehensive Foundation, 2nd ed.* Prentice-Hall.

Haykin, S. and Principe, J. (May 1998). Making sense of a complex world. *IEEE Signal Processing Magazine*, 15(3):66–81.

Hinton, G. and Sejnowski, T. J. (1999). *Unsupervised Learning – Fondations of Neural Computation*. MIT Press, Cambridge, MA.

Hinton, G. and van Camp, D. (1993). Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory*, pages 5–13, Santa Cruz, CA, USA.

Højen-Sørensen, P., Winther, O., and Hansen, L. K. (2002). Mean-field approaches to independent component analysis. *Neural Computation*, 14(4):889–918.

Honkela, A. and Valpola, H. (2005). Unsupervised variational Bayesian learning of nonlinear models. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 593–600. MIT Press, Cambridge, MA, USA.

Hyvärinen, A. (1999a). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A. (1999b). Sparse code shrinkage: Denoising by maximum likelihood estimation. *Neural Computation*, 12(3):429–439.

Hyvärinen, A. (2001). Blind source separation by nonstationarity of variance: a cumulant-based approach. *IEEE Transactions on Neural Networks*, 12(6):1471–1474.

Hyvärinen, A. (2005). A unifying model for blind separation of independent sources. *Signal Processing*, 85(7):1419–1427.

Hyvärinen, A., Hoyer, P., and Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, 13(7):1525–1558.

Hyvärinen, A. and Hoyer, P. O. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley.

Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439.

Hyvärinen, A., Särelä, J., and Vigário, R. (1999). Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size. In *Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '99)*, pages 425–429, Aussois, France.

Jaakkola, T. (2000). Tutorial on variational approximation methods. In *Advanced Mean Field Methods: Theory and Practice*. MIT Press, Cambridge, MA.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37.

James, C. J. and Hesse, C. W. (2005). Independent component analysis for biomedical signals. *Physiological Measurement*, 26:R15–R39.

Jones, M. and Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society, Series A*, 150:1–36.

Julier, S. and Uhlmann, J. K. (1996). A general method for approximating nonlinear transformations of probability distributions. Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10.

Jutten, C. and Karhunen, J. (2004). Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. *International Journal of Neural Systems*, 14(5):267–292.

Kalnay, E. and coauthors (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77:437–471.

Karhunen, J. and Joutsensalo, J. (1994). Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127.

Kawamoto, M., Matsuoka, K., and Oya, M. (1997). Blind separation of sources using temporal correlation of the observed signals. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E80-A(4):695–704.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer-Verlag, Berlin, Heidelberg, New York.

Kramer, M. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243.

Lappalainen, H. (1999). Ensemble learning for independent component analysis. In *Proceedings of International Workshop on Independent Component Analysis and Signal Separation (ICA '99)*, pages 7–12, Aussois, France.

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

Lappalainen, H. and Miskin, J. (2000). Ensemble learning. In Girolami, M., editor, *Advances in Independent Component Analysis*, pages 75–92. Springer-Verlag, Berlin.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816.

Lee, J. (2003). *From Principal Component Analysis to Non-Linear Dimensionality Reduction and Blind Source Separation*. PhD thesis, Université Catholique de Louvain-La-Neuve.

Lee, J., Jutten, C., and Verleysen, M. (2004). Non-linear ICA by using isometric dimensionality reduction. In *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA 2004)*, pages 710–717, Granada, Spain.

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, New Jersey.

Lotsch, A., Friedl, M. A., and Pinzón, J. (2003). Spatio-temporal deconvolution of NDVI image sequences using independent component analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 41(12):2938–2942.

Lütkepohl, H. (1993). *Introduction to multiple time series analysis*. Springer-Verlag, Berlin.

MacKay, D. J. C. (1995a). Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research, Section A*, 354(1):73–80.

MacKay, D. J. C. (1995b). Ensemble learning and evidence maximization. Technical report, Cavendish Laboratory, University of Cambridge.

MacKay, D. J. C. (1995c). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6:469–505.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

Matsuoka, K., Ohya, M., and Kawamoto, M. (1995). A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419.

Maybeck, P. S. (1982). *Stochastic Models, Estimation and Control*. Academic Press.

Minka, T. (2001). *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusets Institute of Technology.

Miskin, J. and MacKay, D. J. C. (2000). Ensemble learning for blind image separation and deconvolution. In Girolami, M., editor, *Advances in Independent Component Analysis*, pages 123–141. Springer-Verlag.

Miskin, J. and MacKay, D. J. C. (2001). Ensemble learning for blind source separation. In Roberts, S. and Everson, R., editors, *Independent Component Analysis: Principles and Practice*, pages 209–233. Cambridge University Press.

Molgedey, J. and Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637.

NCEP data (2004). NCEP Reanalysis data provided by the NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, USA. Available from `http://www.cdc.noaa.gov/`.

Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-92-1, Department of Computer Science, University of Toronto.

Neal, R. M. (1998). Assessing relevance determination methods using DELVE. In Bishop, C. M., editor, *Neural Networks and Machine Learning*, pages 97–129. Springer-Verlag.

Neal, R. M. and Hinton, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 355–368. The MIT Press, Cambridge, MA, USA.

Oja, E. (1983). *Subspace Methods of Pattern Recognition*. Research Studies, Press, Letchworth, England.

Oja, E. (1991). Data compression, feature extraction, and autoassociation in feedforward neural networks. In Kohonen, T., Mäkisara, K., Simula, O., and Kangas, J., editors, *Artificial Neural Networks (Proceedings of International Conference on Artificial Neural Networks, ICANN '91)*, pages 737–745. Elsevier, Amsterdam.

Oja, E. (2002). Unsupervised learning in neural computation. *Theoretical Computer Science*, 287:187–207.

Opper, M. (1998). A Bayesian approach to online learning. In Saad, D., editor, *On-line Learning in Neural Networks*, pages 363–378. Cambridge University Press.

Pajunen, P., Hyvärinen, A., and Karhunen, J. (1996). Nonlinear blind source separation by self-organizing maps. In *Proceedings of International Conference on Neural Information Processing*, pages 1207–1210, Hong Kong.

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd edition.

Pham, D.-T. (2000). Blind separation of instantaneous mixture of sources based on order statistics. *IEEE Transactions on Signal Processing*, 48(2):363–375.

Pham, D.-T. and Cardoso, J.-F. (2001). Blind separation of instantaneous mixtures of non stationary sources. *IEEE Transactions on Signal Processing*, 49:1837–1848.

Pham, D.-T., Garrat, P., and Jutten, C. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, pages 771–774.

Raiko, T., Valpola, H., Östman, T., and Karhunen, J. (2003). Missing values in hierarchical nonlinear factor analysis. In *Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP 2003)*, pages 185–189, Istanbul, Turkey.

Richman, M. B. (1986). Rotation of principal components. *Journal of Climatology*, 6:293–335.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural Computation*, 11(2):305–346.

Roweis, S. and Ghahramani, Z. (2001). An EM algorithm for identification of nonlinear dynamical systems. In Haykin, S., editor, *Kalman Filtering and Neural Networks*, pages 175–220. Wiley, New York.

Roweis, S. T. (1998). EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems 10*, pages 626–632. MIT Press, Cambridge, MA.

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

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409.

Särelä, J. and Valpola, H. (2005). Denoising source separation. *Journal of Machine Learning Research*, 6:233–272.

Särelä, J., Valpola, H., Vigário, R., and Oja, E. (2001). Dynamical factor analysis of rhythmic magnetoencephalographic activity. In *Proceedings of Internationl Conference on Independent Component Analysis and Signal Separation (ICA 2001)*, pages 451–456, San Diego, USA.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series*, 3(4):253–264.

Stone, J. V. (2001). Blind source separation using temporal predictability. *Neural Computation*, 13(7):1559–1574.

Switzer, P. (1985). Min/max autocorrelation factors for multivariate spatial imagery. In Billard, L., editor, *Computer Science and Statistics*, pages 13–16. Elsevier Science Publishers B.V.

Takens, F. (1981). Detecting strange attractors in turbulence. In Rand, D. and Young, L.-S., editors, *Dynamical Systems and Turbulence*, pages 366–381. Springer-Verlag, Berlin.

Taleb, A. and Jutten, C. (1999a). Batch algorithm for source separation in post-nonlinear mixtures. In *Proceedings of International Workshop on Independent Component Analysis and Signal Separation (ICA '99)*, pages 155–160, Aussois, France.

Taleb, A. and Jutten, C. (1999b). Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820.

Tan, Y., Wang, J., and Zurada, J. M. (2001). Nonlinear blind source separation using a radial basis function network. *IEEE Transactions on Neural Networks*, 12(1):124–134.

Tenenbaum, J. B., da Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.

Tipping, M. E. (1996). *Topographic Mappings and Feed-Forward Neural Networks*. PhD thesis, Astom University, Aston Street, Birmingham B4 7ET, UK.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, Series B 21(3):611–622.

Tong, L., Soo, V., Liu, R., and Huang, Y. (1991). Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38(5):499–509.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17:401–419.

Trenberth, K. E. and Caron, J. M. (2000). The Southern Oscillation revisited: Sea level pressures, surface temperatures, and precipitation. *Journal of Climate*, 13:4358–4365.

Valpola, H. (2000). Nonlinear independent component analysis using ensemble learning: theory. In *Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, pages 251–256, Helsinki, Finland.

Valpola, H., Harva, M., and Karhunen, J. (2004). Hierarchical models of variance sources. *Signal Processing*, 84(2):267–282.

Valpola, H. and Karhunen, J. (2002). An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692.

Valpola, H., Oja, E., Ilin, A., Honkela, A., and Karhunen, J. (2003a). Nonlinear blind source separation by variational Bayesian learning. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E86-A(3):532–541.

Valpola, H., Östman, T., and Karhunen, J. (2003b). Nonlinear independent factor analysis by hierarchical models. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, pages 257–262, Nara, Japan.

Valpola, H. and Pajunen, P. (2000). Fast algorithms for Bayesian independent component analysis. In *Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, pages 233–237, Helsinki, Finland.

Valpola, H. and Särelä, J. (2004). Accurate, fast and stable denoising source separation algorithms. In Puntonet, C. G. and Prieto, A., editors, *Proceedings of Fifth International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, volume 3195 of *Lecture Notes in Computer Science*, pages 65–72. Springer-Verlag, Berlin.

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468.

von Storch, H. and Zwiers, W. (1999). *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, U.K.

Wan, E. A. and van der Merwe, R. (2001). The unscented Kalman filter. In Haykin, S., editor, *Kalman Filtering and Neural Networks*, pages 221–280. Wiley, New York.

Wang, B. and Titterington, D. M. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, 20:151–170.

Williams, C. K. I. (1995). On a connection between kernel PCA and metric multidimensional scaling. In *Advances in Neural Information Processing Systems 13*, pages 675–681. MIT Press, Cambridge, MA.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14:715–770.

Yang, H. H., Amari, S.-I., and Cichocki, A. (1998). Information-theoretic approach to blind separation of sources in non-linear mixture. *Signal Processing*, 64(3):291–300.

Ziehe, A. and Müller, K.-R. (1998). TDSEP — an effective algorithm for blind separation using time structure. In *Proceedings of the 8th International Conference on Artificial Neural Networks (ICANN '98)*, pages 675–680, Skövde, Sweden.