Algorithms and Methods

Chapter 2

Bayesian learning of latent variable models

Juha Karhunen, Antti Honkela, Tapani Raiko, Markus Harva, Alexander Ilin, Matti Tornio, Harri Valpola

2.1 Bayesian modeling and variational learning: introduction

Unsupervised learning methods are often based on a generative approach where the goal is to find a latent variable model which explains how the observations were generated. It is assumed that there exist certain latent variables (also called in different contexts source signals, factors, or hidden variables) which have generated the observed data through an unknown mapping. The goal of generative learning is to identify both the latent variables and the unknown generative mapping.

The success of a specific model depends on how well it captures the structure of the phenomena underlying the observations. Various linear models have been popular, because their mathematical treatment is fairly easy. However, in many realistic cases the observations have been generated by a nonlinear process. Unsupervised learning of a nonlinear model is a challenging task, because it is typically computationally much more demanding than for linear models, and flexible models require strong regularization for avoiding overfitting.

In Bayesian data analysis and estimation methods, all the uncertain quantities are modeled in terms of their joint probability distribution. The key principle is to construct the joint posterior distribution for all the unknown quantities in a model, given the data sample. This posterior distribution contains all the relevant information on the parameters to be estimated in parametric models, or the predictions in non-parametric prediction or classification tasks [1, 2].

Denote by $\mathcal{H}$ the particular model under consideration, and by $\boldsymbol{\theta}$ the set of model parameters that we wish to infer from a given data set X. The posterior probability density $p(\boldsymbol{\theta}|X, \mathcal{H})$ of the parameters given the data X and the model $\mathcal{H}$ can be computed from the Bayes' rule

$$p(\boldsymbol{\theta}|X, \mathcal{H}) = \frac{p(X|\boldsymbol{\theta}, \mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(X|\mathcal{H})} \tag{2.1}$$

Here $p(X|\boldsymbol{\theta}, \mathcal{H})$ is the likelihood of the parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathcal{H})$ is the prior pdf of the parameters, and $p(X|\mathcal{H})$ is a normalizing constant. The term $\mathcal{H}$ denotes all the assumptions made in defining the model, such as the choice of a particular model class and structure, specific noise model, etc.

The parameters $\boldsymbol{\theta}$ of a particular model $\mathcal{H}_i$ are often estimated by seeking the peak value of a probability distribution. The non-Bayesian maximum likelihood (ML) method uses to this end the distribution $p(X|\boldsymbol{\theta}, \mathcal{H})$ of the data, and the Bayesian maximum a posteriori (MAP) method finds the parameter values that maximize the posterior probability density $p(\boldsymbol{\theta}|X, \mathcal{H})$. However, using point estimates provided by the ML or MAP methods is often problematic, because the model order estimation and overfitting (choosing too complicated a model for the given data) are severe problems [1, 2].

Instead of searching for some point estimates, the correct Bayesian procedure is to use all possible models to evaluate predictions and weight them by the respective posterior probabilities of the models. This means that the predictions will be sensitive to regions where the probability mass is large instead of being sensitive to high values of the probability density [3, 2]. This procedure optimally solves the issues related to the model complexity and choice of a specific model $\mathcal{H}_i$ among several candidates. In practice, however, the differences between the probabilities of candidate model structures are often very large, and hence it is sufficient to select the most probable model and use the estimates or predictions given by it.

A problem with fully Bayesian estimation is that the posterior distribution (2.1) has a highly complicated form except for in the simplest problems. Therefore it is too difficult

to handle exactly, and some approximative method must be used. Variational methods form a class of approximations where the exact posterior is approximated with a simpler distribution [4, 2]. In a method commonly known as *Variational Bayes (VB)* [1, 3, 2] the misfit of the approximation is measured by the Kullback-Leibler (KL) divergence between two probability distributions $q(v)$ and $p(v)$. The KL divergence is defined by

$$D(q \parallel p) = \int q(v) \ln \frac{q(v)}{p(v)} dv \tag{2.2}$$

which measures the difference in the probability mass between the densities $q(v)$ and $p(v)$.

A key idea in the VB method is to minimize the misfit between the actual posterior pdf and its parametric approximation using the KL divergence. The approximating density is often taken a diagonal multivariate Gaussian density, because the computations become then tractable. Even this crude approximation is adequate for finding the region where the mass of the actual posterior density is concentrated. The mean values of the Gaussian approximation provide reasonably good point estimates of the unknown parameters, and the respective variances measure the reliability of these estimates.

A main motivation of using VB is that it avoids overfitting which would be a difficult problem if ML or MAP estimates were used. VB method allows one to select a model having appropriate complexity, making often possible to infer the correct number of latent variables or sources. It has provided good estimation results in the very difficult unsupervised (blind) learning problems that we have considered.

Variational Bayes is closely related to information theoretic approaches which minimize the description length of the data, because the description length is defined to be the negative logarithm of the probability. Minimal description length thus means maximal probability. In the probabilistic framework, we try to find the latent variables or sources and the nonlinear mapping which most probably correspond to the observed data. In the information theoretic framework, this corresponds to finding the latent variables or sources and the mapping that can generate the observed data and have the minimum total complexity. This information theoretic view also provides insights to many aspects of learning and helps to explain several common problems [5].

In the following subsections, we first discuss a natural conjugate gradient algorithm which speeds up learning remarkably compared with alternative variational Bayesian learning algorithms. We then briefly present a practical building block framework that can be used to easily construct new models. This work has been for the most part carried out already before the years 2006-2007 covered in this biennial report. After this we consider the difficult nonlinear blind source separation (BSS) problem using our Bayesian methods. This section has been placed into the Bayes chapter instead of the ICA/BSS because the methods used are all Bayesian. This section is followed by variational Bayesian learning of nonlinear state-space models, which are applied to time series prediction, improving inference of states, and stochastic nonlinear model predictive control. After this we consider an approach for non-negative blind source separation, and then principal component analysis in the case of missing values using both Bayesian and non-Bayesian approaches. We then discuss predictive uncertainly and probabilistic relational models. Finally we present applications of the developed Bayesian methods to astronomical data analysis problems. In most of these topics, variational Bayesian learning is used, but for relational models and estimation of time delays in astronomical applications other Bayesian methods are applied.

2.2 Natural conjugate gradient in variational inference

Variational methods for approximate inference in machine learning often adapt a parametric probability distribution to optimize a given objective function. This view is especially useful when applying variational Bayes (VB) to models outside the conjugate-exponential family. For them, variational Bayesian expectation maximization (VB EM) algorithms are not easily available, and gradient-based methods are often used as alternatives.

In previous machine learning algorithms based on natural gradients [6], the aim has been to use maximum likelihood to directly update the model parameters $\boldsymbol{\theta}$ taking into account the geometry imposed by the predictive distribution for data $p(\boldsymbol{X}|\boldsymbol{\theta})$. The resulting geometry is often very complicated as the effects of different parameters cannot be separated and the Fisher information matrix is relatively dense.

Recently, in [7], we propose using natural gradients for free energy minimisation in variational Bayesian learning using the simpler geometry of the approximating distributions $q(\boldsymbol{\theta}|\boldsymbol{\xi})$. Because the approximations are often chosen to minimize dependencies between different parameters $\boldsymbol{\theta}$, the resulting Fisher information matrix with respect to the variational parameters $\boldsymbol{\xi}$ will be mostly diagonal and hence easy to invert.

While taking into account the structure of the approximation, plain natural gradient in this case ignores the structure of the model and the global geometry of the parameters $\boldsymbol{\theta}$. This can be addressed by using conjugate gradients. Combining the natural gradient search direction with a conjugate gradient method yields our proposed *natural conjugate gradient (NCG)* method, which can also be seen as an approximation to the fully Riemannian conjugate gradient method.

The NCG algorithm was compared against conjugate gradient (CG) and natural gradient (NG) algorithms in learning a nonlinear state-space model [8]. The results for a number of datasets ranging from 200 to 500 samples of 21 dimensional speech spectrograms can be seen in Figure 2.1. The plain CG and NG methods were clearly slower than others and the maximum runtime of 24 hours was reached by most CG and some NG runs. NCG was clearly the fastest algorithm with the older heuristic method of [8] between these extremes. The results with a larger data set are very similar with NCG outperforming all alternatives by a factor of more than 10.

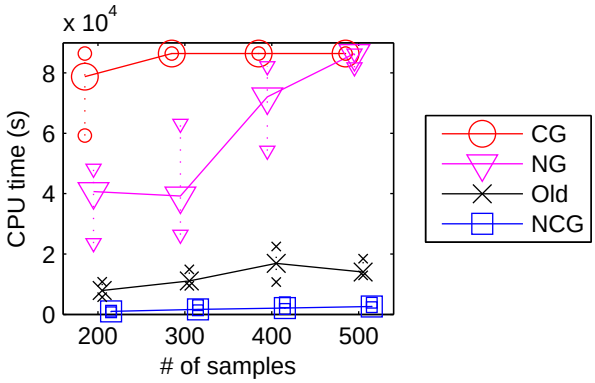


Figure 2.1: Convergence speed of the natural conjugate gradient (NCG), the natural gradient (NG) and the conjugate gradient (CG) methods as well as the heuristic algorithm (Old) with different data sizes. The lines show median times with 25 % and 75 % quantiles shown by the smaller marks. The times were limited to at most 24 hours, which was reached by a number of simulations.

The experiments in [7] show that the natural conjugate gradient method outperforms both conjugate gradient and natural gradient methods by a large margin. Considering univariate Gaussian distributions, the regular gradient is too strong for model variables with small posterior variance and too weak for variables with large posterior variance. The posterior variance of latent variables is often much larger than the posterior variance of model parameters and the natural gradient takes this into account in a very natural manner.

2.3 Building blocks for variational Bayesian learning

In graphical models, there are lots of possibilities to build the model structure that defines the dependencies between the parameters and the data. To be able to manage the variety, we have designed a modular software package using C++/Python called the Bayes Blocks [9]. The theoretical framework on which it is based on was published in [10] and a description of the software package was published in [11].

The design principles for Bayes Blocks have been the following. Firstly, we use standardized building blocks that can be connected rather freely and can be learned with local learning rules, i.e. each block only needs to communicate with its neighbors. Secondly, the system should work with very large scale models. We have made the computational complexity linear with respect to the number of data samples and connections in the model.

The building blocks include Gaussian variables, summation, multiplication, nonlinearity, mixture-of-Gaussians, and rectified Gaussians. Each of the blocks can be a scalar or a vector. Variational Bayesian learning provides a cost function which can be used for updating the variables as well as optimizing the model structure. The derivation of the cost function and learning rules is automatic which means that the user only needs to define the connections between the blocks. Examples of structures which can be build using the Bayes Blocks library can be found in Figure 2.2.

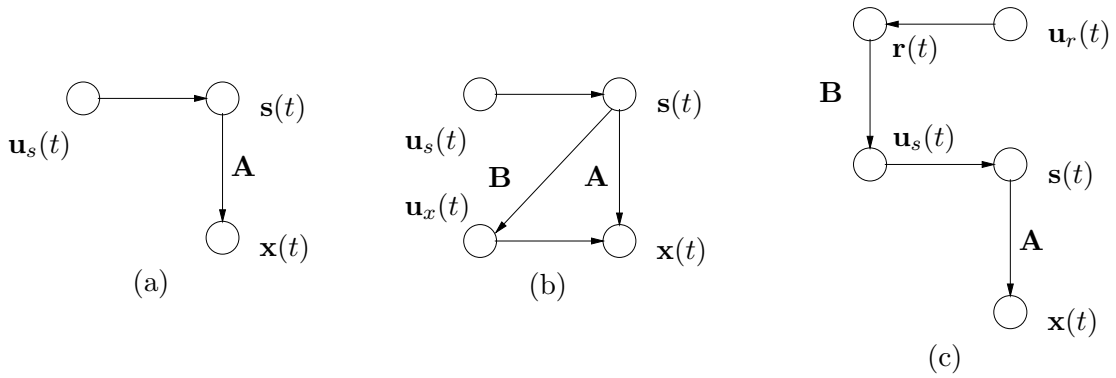


Figure 2.2: Various model structures utilizing variance nodes. Observations are denoted by $\mathbf{x}$, linear mappings by $\mathbf{A}$ and $\mathbf{B}$, sources by $\mathbf{s}$ and $\mathbf{r}$, and variance nodes by $\mathbf{u}$.

2.4 Nonlinear BSS and ICA

A fundamental difficulty in the nonlinear blind source separation (BSS) problem and even more so in the nonlinear independent component analysis (ICA) problem is that they provide non-unique solutions without extra constraints, which are often implemented by using a suitable regularization. Our approach to nonlinear BSS uses Bayesian inference methods for estimating the best statistical parameters, under almost unconstrained models in which priors can be easily added.

We have applied variational Bayesian learning to nonlinear factor analysis (FA) and BSS where the generative mapping from sources to data is not restricted to be linear. The general form of the model is

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_f) + \mathbf{n}(t) \tag{2.3}$$

This can be viewed as a model about how the observations were generated from the sources. The vectors $\mathbf{x}(t)$ are observations at time t, $\mathbf{s}(t)$ are the sources, and $\mathbf{n}(t)$ the noise. The function $\mathbf{f}(\cdot)$ is a mapping from source space to observation space parametrized by $\boldsymbol{\theta}_f$.

In an earlier work [13] we have used multi-layer perceptron (MLP) network with tanh-nonlinearities to model the mapping $\mathbf{f}$:

$$\mathbf{f}(\mathbf{s}; \mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}) = \mathbf{B} \tanh(\mathbf{A}\mathbf{s} + \mathbf{a}) + \mathbf{b}\,. \tag{2.4}$$

The mapping $\mathbf{f}$ is thus parameterized by the matrices $\mathbf{A}$ and $\mathbf{B}$ and bias vectors $\mathbf{a}$ and $\mathbf{b}$. MLP networks are well suited for nonlinear FA and BSS. First, they are universal function approximators which means that any type of nonlinearity can be modeled by them in principle. Second, it is easy to model smooth, nearly linear mappings with them. This makes it possible to learn high dimensional nonlinear representations in practice.

An important special case of general nonlinear mixtures in (2.3) is a post-nonlinear (PNL) mixing model. There linear mixing is followed by component-wise nonlinearities acting on each output independently of the others:

$$x_i(t) = f_i\left[\mathbf{a}_i^T \mathbf{s}(t)\right] + n_i(t) \qquad i = 1, \ldots, n \tag{2.5}$$

Such models are plausible in applications where linearly mixed signals are measured by sensors with nonlinear distortions f_i. The nonlinearities f_i can also be modelled by MLP networks.

Identification of models (2.3) or (2.5) assuming Gaussianity of sources $\mathbf{s}(t)$ helps to find a compact representation of the observed data $\mathbf{x}(t)$. Nonlinear BSS can be achieved by performing a linear rotation of the found sources using, for example, a linear ICA technique.

The paper [12] presents our recent developments on nonlinear FA and BSS. A more accurate linearization increases stability of the algorithm in cases with a large number of sources when the posterior variances of the last weak sources are typically large. A hierarchical nonlinear factor analysis (HNFA) model using the building blocks presented in Section 2.3 is applicable to larger problems than the MLP based method, as the computational complexity is linear with respect to the number of sources. Estimating the PNL factor analysis model in (2.5) using variational Bayesian learning helps achieve separation of signals in very challenging BSS problems.

2.5 Nonlinear state-space models

In many cases, measurements originate from a dynamical system and form a time series. In such instances, it is often useful to model the dynamics in addition to the instantaneous observations. We have used rather general nonlinear models for both the data (observations) and dynamics of the sources (latent variables) [8]. This results in a state-space model where the sources can be interpreted as the internal state of the underlying generative process.

The general form of our nonlinear model for the generative mapping from the source (latent variable) vector $\mathbf{s}(t)$ to the data (observation) vector $\mathbf{x}(t)$ at time t is the same as in Eq. (2.3):

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_f) + \mathbf{n}(t)\,. \tag{2.6}$$

The dynamics of the sources can be modelled by another nonlinear mapping, which leads to a source model [8]

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1), \boldsymbol{\theta}_g) + \mathbf{m}(t)\,, \tag{2.7}$$

where $\mathbf{s}(t)$ are the sources (states) at time t, $\mathbf{m}$ is the Gaussian noise, and $\mathbf{g}(\cdot)$ is a vector containing as its elements the nonlinear functions modelling the dynamics.

As for the static models presented in Sec. 2.4, the nonlinear functions are modelled by MLP networks. The mapping $\mathbf{f}$ has the same functional form (2.4). Since the states in dynamical systems are often slowly changing, the MLP network for mapping $\mathbf{g}$ models the change in the value of the source:

$$\mathbf{g}(\mathbf{s}(\mathbf{t}-\mathbf{1})) = \mathbf{s}(t-1) + \mathbf{D}\tanh[\mathbf{C}\mathbf{s}(t-1) + \mathbf{c}] + \mathbf{d}\,. \tag{2.8}$$

The dynamic mapping $\mathbf{g}$ is thus parameterized by the matrices $\mathbf{C}$ and $\mathbf{D}$ and bias vectors $\mathbf{c}$ and $\mathbf{d}$.

Estimation of the arising state-space model is rather involved, and it is discussed in detail in our earlier paper [8]. An important advantage of the proposed nonlinear state-space method (NSSM) is its ability to learn a high-dimensional latent source space. We have also reasonably solved computational and over-fitting problems which have been major obstacles in developing this kind of unsupervised methods thus far. Potential applications for our method include prediction and process monitoring, control and identification. MATLAB software packages are available for both the static model (2.3)-(2.4) (under the name nonlinear factor analysis) and the dynamic model (2.7)-(2.8) (under the name nonlinear dynamical factor analysis) on the home page of our Bayes group [14].

Time series prediction

Traditionally, time series prediction is done using models based directly on the past observations of the time series. Perhaps the two most important classes of neural network based solutions used for nonlinear prediction are feedforward autoregressive neural networks and recurrent autoregressive moving average neural networks [15]. However, instead of modelling the system based on past observations, it is also possible to model the same information in a more compact form with a state-space model.

We have used the nonlinear state-space model and method [8] described in the beginning of this section to model a time series. The primary goal in the paper [16] was to apply our NSSM method and software [14] to the time series prediction task as a black box tool. The details of this application are given in [16].

We applied the NSSM method to the prediction of the nonlinear scalar time series provided by the organizers of the ESTSP'07 symposium. The original time series containing 875 samples is shown in Figure 2.3. It seems to be strongly periodic with a period of

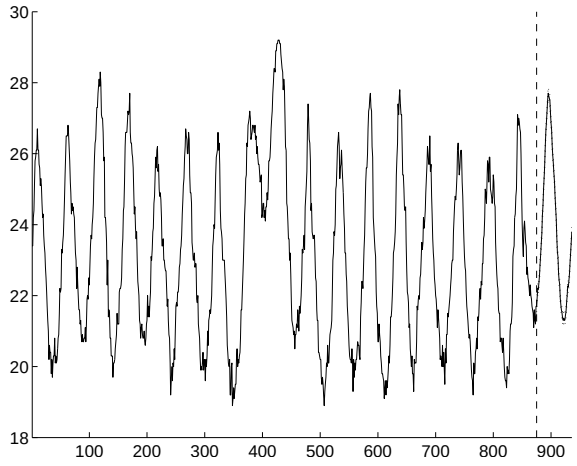


Figure 2.3: The original time series and the predicted 61 next time steps.

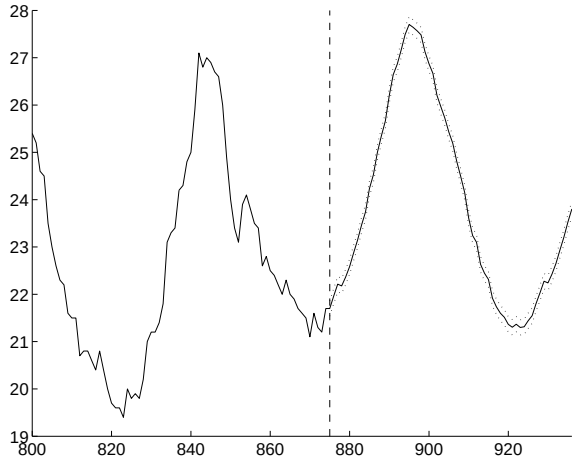


Figure 2.4: Bottom: The original time series starting from time instant 800 and the predicted 61 next time steps.

approximately 52 samples. Figure 2.3 shows also the predicted 61 next time steps, and Figure 2.4 in more detail the original time series starting from time instant 800 and the predicted 61 next time steps. The dotted lines in both figures represent pseudo 95 % confidence intervals. These intervals are, however, smaller than in reality as the variance caused by the innovation is ignored [16].

Improving state inference

The problem of state inference involves finding the source vectors $\mathbf{s}(t-1)$ given the data and the model. While this is an easier problem than finding both the model and the sources, it is more time critical, since it must often be computed in real-time. While the algorithm in [8] can be used for inference, it is very slow because of the slow flow of information through the time series. Standard algorithms based on extensions of the Kalman smoother work rather well in general, but may fail to converge when estimating the states over a long gap or when used together with learning the model.

When updates are done locally, information spreads around slowly because the states of different time slices affect each other only between updates. It is possible to predict this interaction by a suitable approximation. In [17], we derived a novel update algorithm

for the posterior mean of the states by replacing partial derivatives of the cost function with respect to state means $\overline{\mathbf{s}}(t)$ by (approximated) total derivatives

$$\frac{d\mathcal{C}_{\mathrm{KL}}}{d\overline{\mathbf{s}}(t)} = \sum_{\tau=1}^{T} \frac{\partial \mathcal{C}_{\mathrm{KL}}}{\partial \overline{\mathbf{s}}(\tau)} \frac{\partial \overline{\mathbf{s}}(\tau)}{\partial \overline{\mathbf{s}}(t)}. \tag{2.9}$$

They can be computed efficiently using the chain rule and dynamic programming, given that we can approximate the terms $\partial\overline{\mathbf{s}}(t)/\partial\overline{\mathbf{s}}(t-1)$ and $\partial\overline{\mathbf{s}}(t)/\partial\overline{\mathbf{s}}(t+1)$.

This is how we approximated the required partial derivatives. The posterior distribution of the state $\mathbf{s}(t)$ can be factored into three potentials, one from $\mathbf{s}(t-1)$ (the past), one from $\mathbf{s}(t+1)$ (the future), and one from $\mathbf{x}(t)$ (the observation). We linearized the nonlinear mappings so that the three potentials become Gaussian. Then also the posterior of $\mathbf{s}(t)$ becomes Gaussian with a mean that is the weighted average of the means of the three potentials, where the weights are the inverse (co)variances of the potentials. A change in the mean of a potential results in a change of the mean of the posterior inversely proportional to their (co)variances.

Experimental comparison in [17] showed that the proposed algorithm worked reliably and fast. The algorithms from the Kalman family (IEKS and IUKS) were fast, too, but they also suffered from stability problems when gaps of 30 consequtive missing observations were introduced into the data. Basic particle smoother performed very poorly compared to the iterative algorithms. It should be noted that many different schemes exists to improve the performance of particle filters.

Stochastic nonlinear model-predictive control

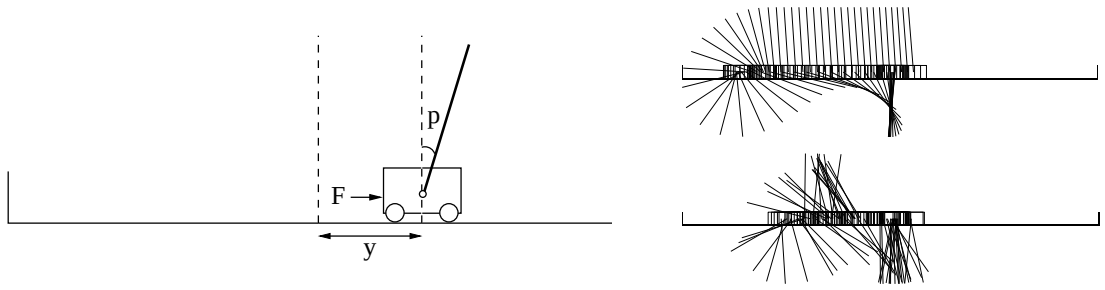


Figure 2.5: Left: The cart-pole system. The goal is to swing the pole to an upward position and stabilize it without hitting the walls. The cart can be controlled by applying a force to it. Top left: The pole is successfully swinged up by moving first to the left and then right. Bottom right: Our controller works quite reliably even in the presence of serious observation noise.

In [18], we studied such a system combining variational Bayesian learning of an unknown dynamical system with nonlinear model-predictive control. For being able to control the dynamical system, control inputs are added to the nonlinear state-space model. Then we can use stochastic nonlinear model-predictive control, which is based on optimising control signals based on maximising a utility function.

Figure 2.5 shows simulations with a cart-pole swing-up task. The results confirm that selecting actions based on a state-space model instead of the observation directly has many benefits: First, it is more resistant to noise because it implicitly involves filtering. Second, the observations (without history) do not always carry enough information about the system state. Third, when nonlinear dynamics are modelled by a function approximator such

as an multilayer perceptron network, a state-space model can find such a representation of the state that it is more suitable for the approximation and thus more predictable.

Continuous-time modeling

In [19], we have outlined an extension of the discrete-time variational Bayesian NSSM of [8] to continuous-time systems and presented preliminary experimental results with the method. Evaluation of the method with larger and more realistic examples is a very important item of further work. The main differences between continuous-time and discrete-time variational NSSMs are the different method needed to evaluate the predictions of the states and the different form of the dynamical noise or innovation.

2.6 Non-negative blind source separation

In linear factor analysis (FA) [20], the observations are modeled as noisy linear combinations of a set of underlying sources or factors. When the level of noise is low, FA reduces to principal component analysis (PCA). Both FA and PCA are insensitive to orthogonal rotations, and, as such, cannot be used for blind source separation except in special cases. There are several ways to solve the rotation indeterminacy. One approach is to assume the sources independent, which in low noise leads to independent component analysis. Another approach, the one discussed in this section, is to constrain the sources to be non-negative.

Non-negativity constraints in linear factor models have received a great deal of interest in a number of problem domains. In the variational Bayesian framework, positivity of the factors can be achieved by putting a non-negatively supported prior on them. The rectified Gaussian distribution is particularly convenient, as it is conjugate to the Gaussian likelihood arising in the FA model. Unfortunately, this solution has a technical limitation: the location parameter of the prior has to be fixed to zero; otherwise the potentials of both the location and the scale parameter become awkward.

To evade the above mentioned problems, the model is reformulated using rectification nonlinearities. This can be expressed in the form of Eq. (2.4) using the following nonlinearity

$$\mathbf{f}(\mathbf{s};\, \mathbf{A}) = \mathbf{A}\, \mathbf{cut}(\mathbf{s}) \tag{2.10}$$

where **cut** is the componentwise rectification (or cut) function such that $[\mathbf{cut}(\mathbf{s})]_i = \max(s_i, 0)$. In [21], a variational learning procedure was derived for the proposed model and it was shown that it indeed overcomes the problems that exist with the related approaches (see Figure 2.6 for a controlled experiment). In Section 2.10 an application of the method to the analysis of galaxy spectra is presented. There the underlying sources were such that the zero-location rectified Gaussian prior was highly inappropriate, which motivated the development of the proposed approach.

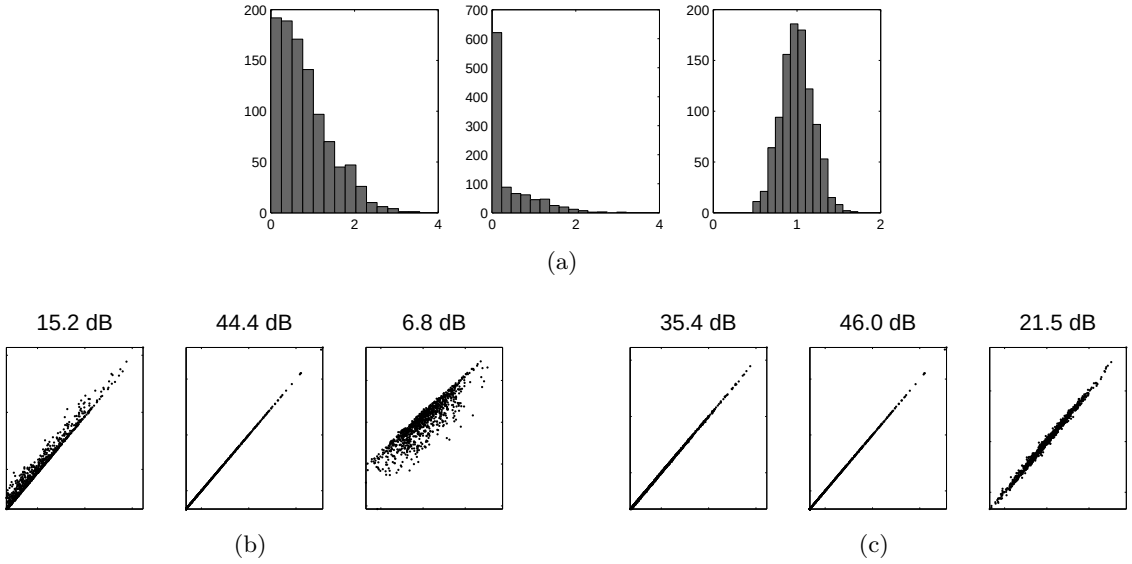


Figure 2.6: (a) The histograms of the true sources to be recovered. (b) and (c) The estimated sources plotted against the true sources with the signal-to-noise ratios printed above each plot. In (b), rectified Gaussian priors have been used for the sources. In (c), the proposed approach employing rectification nonlinearities has been used.

2.7 PCA in the presence of missing values

Principal component analysis (PCA) is a classical data analysis technique. Some algorithms for PCA scale better than others to problems with high dimensionality. They also differ in the ability to handle missing values in the data. In our recent papers [22, 23], a case is studied where the data are high-dimensional and a majority of the values are missing.

In the case of very sparse data, overfitting becomes a severe problem even in simple linear models such as PCA. Regularization can be provided using the Bayesian approach by introducing prior for the model parameters. The PCA model can then be identified using, for example, maximum a posteriori estimates (regularized PCA) or variational Bayesian (VB) learning. We study both approaches in the papers [22, 23].

The proposed learning algorithm is based on speeding up a simple principal subspace rule in which the model parameters are updated as

$$\theta_i \leftarrow \theta_i - \gamma \left(\frac{\partial^2 C}{\partial \theta_i^2} \right)^{-\alpha} \frac{\partial C}{\partial \theta_i}, \tag{2.11}$$

where α is a control parameter that allows the learning algorithm to vary from the standard gradient descent ($\alpha = 0$) to the diagonal Newton's method ($\alpha = 1$). These learning rules can be used for standard PCA learning and extended to regularized PCA and variational Bayesian (VB) PCA.

The algorithms were tested on the Netflix problem (`http://www.netflixprize.com/`), which is a task of predicting preferences (or producing personal recommendations) by using other people's preferences. The Netflix problem consists of movie ratings given by 480189 customers to 17770 movies. There are 100480507 ratings from 1 to 5 given, and the task is to predict 2817131 other ratings among the same group of customers and movies. 1408395 of the ratings are reserved for validation. Thus, 98.8% of the values are missing. We tried to find 15 principal components from the data using a number of methods. The results confirm that the proposed speed-up procedure is much faster than any of the compared methods, and that VB-PCA method provides more accurate predictions for new data than traditional PCA or simple regularized PCA (see Fig. 2.7).

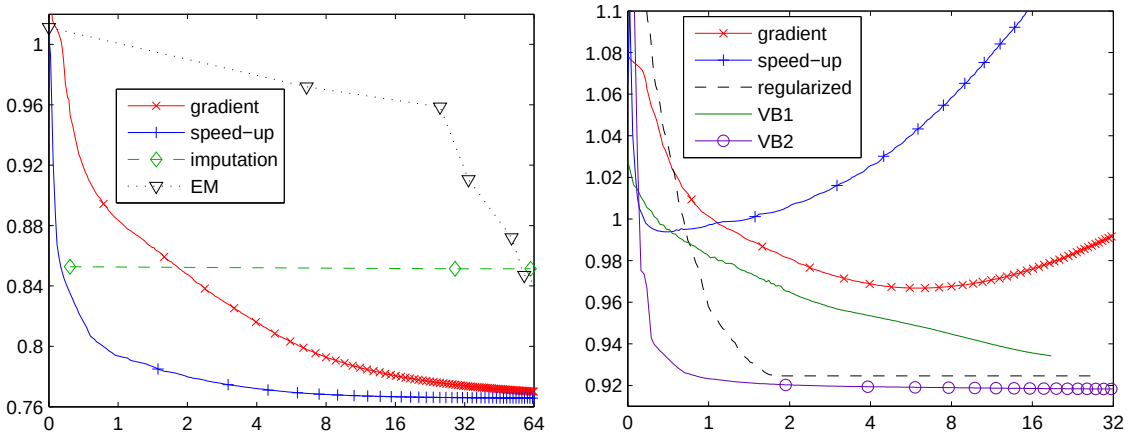


Figure 2.7: *Left:* Training error against computation time in hours in the Netflix problem for unregularized PCA algorithm based on gradient descent and the proposed speed-up. Two alternative methods are shown for comparison. *Right:* The error on test data for the two versions of unregularized PCA, regularized PCA and two variants of variational Bayesian PCA. The time scale is linear below 1 and logarithmic above 1.

2.8 Predictive uncertainty

In standard regression, we seek to predict the value of a response variable based on a set of explanatory variables. Here, the term *predictive uncertainty* is used to refer to a task similar to regression with the exception that we predict not only the mean outcome of the response variable, but also the uncertainty related to its value. For example, consider predicting the concentration of an air pollutant in a city, based on meteorological conditions measured some time in advance. In this task it is the extreme events, namely those occasions when the concentration of the air pollutant rises over a certain threshold, that are interesting. If the conditional distribution of the response variable is not tightly concentrated around its mean value, the mean value by itself will be a poor indicator of the extreme events occurring, and hence predictions based on those alone might lead to policies with ill consequences.

In [26], a method for predictive uncertainty is presented. The method is based on conditioning the scale parameter of the noise process on the explanatory variables and then using MLP networks to model both the location and the scale of the output distribution. The model can be summarised as

$$\begin{aligned} y_t &\sim N(f(\mathbf{x}_t, \boldsymbol{\theta}_y), e^{-u_t}) \\ u_t &\sim N(f(\mathbf{x}_t, \boldsymbol{\theta}_u), \tau^{-1}) \end{aligned} \tag{2.12}$$

Above, y_t is the response variable and $\mathbf{x}_t$ is the vector of explanatory variables. The function f, representing the MLP network, has essentially the same form as in Eq. (2.4). When the latent variable u_t is marginalised out of the model the predictive distribution for y_t becomes super-Gaussian. The extent to which this happens depends on the uncertainty in u_t as measured by the precision parameter τ which is adapted in the learning process. This adaptive nongaussianity of the predictive distribution is highly desirable as then the uncertainty in the scale parameter can be accommodated by making the predictive distribution more robust.

The problem with heteroscedastic models is that learning them using simple methods can be difficult as overfitting becomes a serious concern. Variational Bayesian (VB) methods can, however, largely avoid these problems. Unfortunately, VB methods for nonlinear models, such as that in Eq. (2.12), become involved both in analytic as well as in computational terms. Therefore the learning algorithm in [26] is based on the slightly weaker approximation technique, the variational EM algorithm, and only the "important" parameters have distributional estimates. These parameters include the latent variables u_t, the precision parameter, and the second layer weights of the MLPs. The rest of the parameters, that is, the first layer weights of the MLPs, have point estimates only.

The method summarized in this section was applied to all four datasets in the 'Predictive uncertainty in environmental modelling' competition held at World Congress on Computational Intelligence 2006. The datasets varied in dimensionality from one input variable to 120 variables. The proposed method performed well with all the datasets where heteroscedasticity was an important component being the overall winner of the competition.

2.9 Relational models

In the previous sections, we have studied models belonging to two categories: static and dynamic. In static modeling, each observation or data sample is independent of the others. In dynamic models, the dependencies between consecutive observations are modeled. A generalization of both types of models is that the relations are described in the data itself, that is, each observation might have a different structure.

Logical hidden Markov models

Many real-world sequences such as protein secondary structures or shell logs exhibit rich internal structures. In [24], we have proposed logical hidden Markov models as one solution. They deal with logical sequences, that is, sequences over an alphabet of logical atoms. This comes at the expense of a more complex model selection problem. Indeed, different abstraction levels have to be explored. Logical hidden Markov models (LOHMMs) upgrade traditional hidden Markov models to deal with sequences of structured symbols in the form of logical atoms, rather than characters. Our recent paper [24] formally introduces LOHMMs and presents solutions to the three central inference problems for LOHMMs: evaluation, most likely hidden state sequence, and parameter estimation. The resulting representation and algorithms are experimentally evaluated on problems from the domain of bioinformatics (see Figure 2.8).

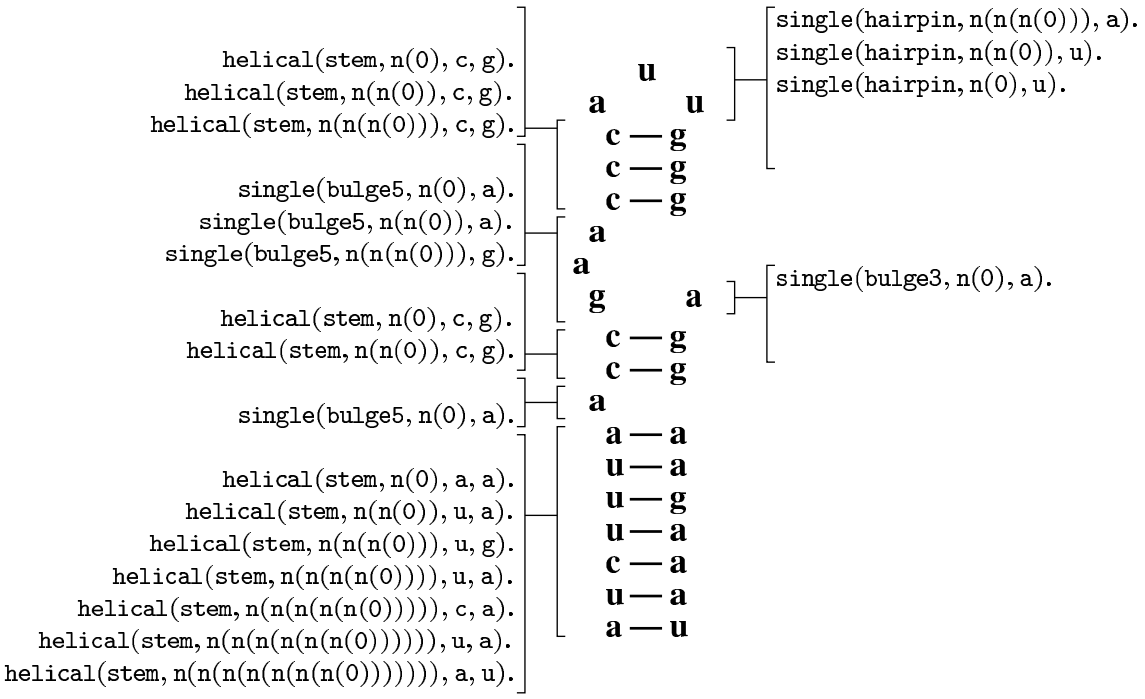


Figure 2.8: Representation of mRNA signal structures as a sequence of logical atoms to be analyzed with a logical hidden Markov model.

Higher order statistics in play-out analysis

A second relational study involves game playing. There is a class of board games called connection games for which traditional artificial intelligence approach does not provide a good computer player. For such games, it is an interesting option to play out the game

from the current state to the end many times randomly. Play-outs provide statistics that can be used for selecting the best move. In [25], we introduce a method that selects relevant patterns of moves to collect higher order statistics. Play-out analysis avoids the horizon effect of regular game-tree search. The proposed method is especially effective when the game can be decomposed into a number of subgames. Preliminary experiments on the board games of Hex and Y are reported in [25].

2.10 Applications to astronomy

Two astronomical applications are discussed in this section: analysis of galaxy spectra and estimation of time delays in gravitational lensing.

Analysis of galaxy spectra

We have applied rectified factor analysis [21] described in Section 2.6 to the analysis of real stellar population spectra of elliptical galaxies. Ellipticals are the oldest galactic systems in the local universe and are well studied in physics. The hypothesis that some of these old galactic systems may actually contain young components is relatively new. Hence, we have investigated whether a set of stellar population spectra can be decomposed and explained in terms of a small set of unobserved spectral prototypes in a data driven but yet physically meaningful manner. The positivity constraint is important in this modelling application, as negative values of flux would not be physically interpretable.

Using a set of 21 real stellar population spectra, we found that they can indeed be decomposed to prototypical spectra, especially to a young and old component [27]. Figure 2.9 shows one spectrum and its decomposition to these two components. The right subfigure shows the ages of the galaxies, known from a detailed astrophysical analysis, plotted against the first weight of the mixing matrix. The plot clearly shows that the first component corresponds to a galaxy containing a significant young stellar population.

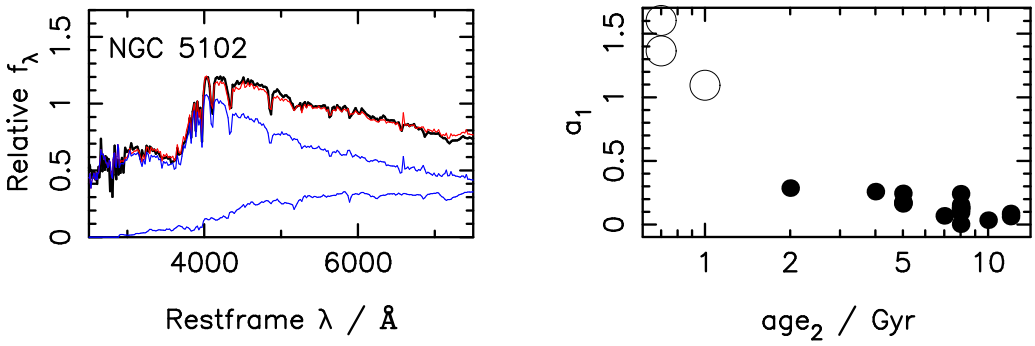


Figure 2.9: Left: the spectrum of a galaxy with its decomposition to a young and old component. Right: the age of the dominating stellar population against the mixing coefficient of the young component.

Estimation of time delays in gravitational lensing

Gravitational lensing occurs when the light coming from a distant bright source is bent by the gravitational potential of an intermediate galaxy such that several images of the source are observed (see the left panel of Figure 2.10 for an example system). Relativistic effects and the different lengths of the paths affect the time it takes for the photons originating from the source to travel to the observer. This is perceived as a delay in the intensity variations between the images (see the right panel of Figure 2.10). The significance of estimating the delays in such systems stems from the early observation that they can be used in determining important cosmological quantities [28].

The delay estimation problem is difficult for various reasons. The main challenge is the uneven sampling rate, as the sampling times are determined by factors one cannot control such as observing conditions and scheduling. The signal-to-noise ratio in the

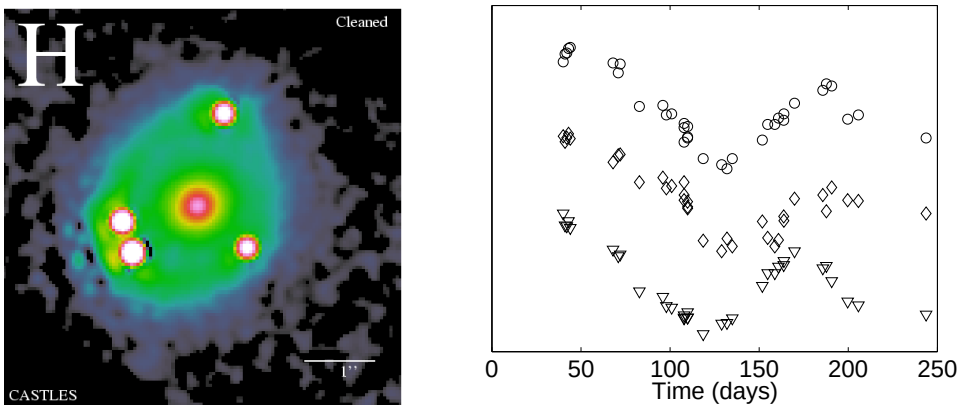


Figure 2.10: Left: The four images of PG1115+080. Right: The corresponding intensity measurements (the two images closest to each other are merged).

observations is often poor too, although this varies somewhat between datasets. Classical delay estimation methods usually rely on the cross-correlation function which is easy to evaluate between regularly sampled signals. The obvious way to attack the problem with unevenly sampled signals would then be to interpolate them appropriately to obtain evenly sampled signals and then apply the cross correlation method. But with all the gaps and the noise in the data, the interpolation can introduce spurious features to the data which make the cross-correlation analysis go awry [29].

In [30, 31], a method for estimating the delay between irregularly sampled signals is presented. Since interpolation on the gappy and noisy data can be venturesome, that is avoided. Instead the two observed signals, $x_1(t)$ and $x_2(t)$, are postulated to have been emitted from the same latent source signal $s(t)$, the observation times being determined by the actual sampling times and the delay. The source is then assumed to follow the Wiener process: $s(t_{i+1}) - s(t_i) \sim N(0, [(t_{i+1} - t_i)\,\sigma]^2)$. This prior encodes the notion of "slow variability" into the model which is an assumption implicitly present in many of the other methods as well. The model is estimated using exact marginalization, which leads to a specific type of Kalman-filter, combined with the Metropolis-Hastings algorithm.

We have used the proposed method to determine the delays in several gravitational lensing systems. Controlled comparisons againts other methods cannot, however, be done with real data as the true delays are unknown to us. Instead, artificial data, where the ground truth is known, must be used. Figure 2.11 shows the performance of several methods in an artificial setting.

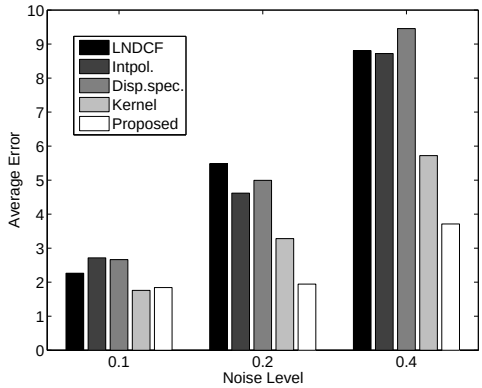


Figure 2.11: Average errors of the methods for three groups of datasets.

References

[1] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[2] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[3] H. Lappalainen and J. Miskin. Ensemble learning. In M. Girolami, editor, *Advances in Independent Component Analysis*, Springer, 2000, pages 75–92.

[4] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. Jordan, editor, *Learning in Graphical Models*, MIT Press, 1999, pages 105–161.

[5] A. Honkela and H. Valpola. Variational learning and bits-back coding: an information-theoretic view to Bayesian learning. *IEEE Transactions on Neural Networks*, 15(4):267–282, 2004.

[6] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[7] A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In *Proc. of the 14th Int. Conf. on Neural Information Processing (ICONIP 2007)*, Kitakyushu, Japan, November 2007.

[8] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[9] H. Valpola, A. Honkela, M. Harva, A. Ilin, T. Raiko, and T. Östman. Bayes Blocks software library. `http://www.cis.hut.fi/projects/bayes/software/`, 2003.

[10] T. Raiko, H. Valpola, M. Harva, and J. Karhunen. Building blocks for variational Bayesian learning of latent variable models. *Journal of Machine Learning Research*, Vol. 8, pp. 155-201, January 2007.

[11] A. Honkela, M. Harva, T. Raiko, H. Valpola, and J. Karhunen. Bayes Blocks: A Python toolbox for variational Bayesian learning. *NIPS2006 Workshop on Machine Learning Open Source Software*, Whistler, B.C., Canada, 2006.

[12] A. Honkela, H. Valpola, A. Ilin and J. Karhunen. Blind separation of nonlinear mixtures by variational Bayesian learning. *Digital Signal Processing*, Vol. 17, No 2, pp. 914–934, 2007.

[13] H. Lappalainen and A. Honkela. Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Mark Girolami, editor, *Advances in Independent Component Analysis*, pages 93–121. Springer-Verlag, Berlin, 2000.

[14] Home page of our Bayes group: `http://www.cis.hut.fi/projects/bayes/`.

[15] A. Trapletti, *On Neural Networks as Statistical Time Series Models.* PhD Thesis, Technische Universität Wien, 2000.

[16] M. Tornio, A. Honkela, and J. Karhunen. Time series prediction with variational Bayesian nonlinear state-space models. In *Proc. European Symp. on Time Series Prediction (ESTSP'07)*, pages 11–19, Espoo, Finland, February 2007.

[17] T. Raiko, M. Tornio, A. Honkela, and J. Karhunen. State inference in variational Bayesian nonlinear state-space models. In *Proc. of the 6th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA 2006)*, pages 222–229, Charleston, South Carolina, USA, March 2006.

[18] M. Tornio and T. Raiko. Variational Bayesian approach for nonlinear identification and control. In *Proc. of the IFAC Workshop on Nonlinear Model Predictive Control for Fast Systems, NMPC FS06*, pp. 41-46, Grenoble, France, October 9-11, 2006.

[19] A. Honkela, M. Tornio, and T. Raiko. Variational Bayes for continuous-time nonlinear state-space models. In *NIPS2006 Workshop on Dynamical Systems, Stochastic Processes and Bayesian Inference*, Whistler, B.C., Canada, 2006.

[20] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[21] M. Harva and A. Kabán. Variational learning for rectified factor analysis. *Signal Processing*, 87(3):509–527, 2007.

[22] T. Raiko, A. Ilin and J. Karhunen. Principal component analysis for large scale problems with lots of missing values. In *Proc. of the 18th European Conf. on Machine Learning (ECML 2007)*, pages 691–698, Warsaw, Poland, September 2007.

[23] T. Raiko, A. Ilin, and J. Karhunen. Principal component analysis for sparse high-dimensional data. In *Proc. of the 14th Int. Conf. on Neural Information Processing (ICONIP 2007)*, Kitakyushu, Japan, November 2007.

[24] K. Kersting, L. De Raedt, and T. Raiko. Logical hidden Markov models. *Journal of Artificial Intelligence Research (JAIR)*, Vol. 25, pp. 425–456, April 2006.

[25] T. Raiko. Higher order statistics in play-out analysis. *Proc. of the Scandinavian Conf. on Artificial Intelligence, SCAI2006*, pp. 189–195, Espoo, Finland, October 25-27, 2006.

[26] M. Harva. A variational EM approach to predictive uncertainty. *Neural Networks*, 20(4):550–558, 2007.

[27] L. Nolan, M. Harva, A. Kabán, and S. Raychaudhury. A data-driven Bayesian approach for finding young stellar populations in early-type galaxies from their UV-optical spectra. *Monthly Notices of the Royal Astronomical Society*, 366(1):321–338, 2006.

[28] S. Refsdal. On the possibility of determining Hubble's parameter and the masses of galaxies from the gravitational lens effect. *Monthly Notices of the Royal Astronomical Society*, 128:307–310, 1964.

[29] J. C. Cuevas-Tello, P. Tino, and S. Raychaudhury. How accurate are the time delay estimates in gravitational lensing? *Astronomy & Astrophysics*, 454:695–706, 2006.

[30] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. In *Proc. Int. Workshop on Machine Learning for Signal Processing (MLSP'06)*, pages 111–116. Maynooth, Ireland, 2006.

[31] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. *Neurocomputing*, 2008. To appear.

Chapter 3

Independent component analysis and blind source separation

Erkki Oja, Juha Karhunen, Alexander Ilin, Antti Honkela, Karthikesh Raju, Tomas Ukkonen, Zhirong Yang, Zhijian Yuan

3.1 Introduction

What is Independent Component Analysis and Blind Source Separation? Independent Component Analysis (ICA) is a computational technique for revealing hidden factors that underlie sets of measurements or signals. ICA assumes a statistical model whereby the observed multivariate data, typically given as a large database of samples, are assumed to be linear or nonlinear mixtures of some unknown latent variables. The mixing coefficients are also unknown. The latent variables are nongaussian and mutually independent, and they are called the independent components of the observed data. By ICA, these independent components, also called sources or factors, can be found. Thus ICA can be seen as an extension to Principal Component Analysis and Factor Analysis. ICA is a much richer technique, however, capable of finding the sources when these classical methods fail completely.

In many cases, the measurements are given as a set of parallel signals or time series. Typical examples are mixtures of simultaneous sounds or human voices that have been picked up by several microphones, brain signal measurements from multiple EEG sensors, several radio signals arriving at a portable phone, or multiple parallel time series obtained from some industrial process. The term blind source separation is used to characterize this problem. Also other criteria than independence can be used for finding the sources.

Our contributions in ICA research. In our ICA research group, the research stems from some early work on on-line PCA, nonlinear PCA, and separation, that we were involved with in the 80's and early 90's. Since mid-90's, our ICA group grew considerably. This earlier work has been reported in the previous Triennial and Biennial reports of our laboratory from 1994 to 2005. A notable achievement from that period was the textbook "Independent Component Analysis" (Wiley, May 2001) by A. Hyvärinen, J. Karhunen, and E. Oja. It has been very well received in the research community; according to the latest publisher's report, over 5000 copies had been sold by August, 2007. The book has been extensively cited in the ICA literature and seems to have evolved into the standard text on the subject worldwide. In Google Scholar, the number of hits (in early 2008) is over 2300. In 2005, the Japanese translation of the book appeared (Tokyo Denki University Press), and in 2007, the Chinese translation (Publishing House of Electronics Industry).

Another tangible contribution has been the public domain FastICA software package (`http://www.cis.hut.fi/projects/ica/fastica/`). This is one of the few most popular ICA algorithms used by the practitioners and a standard benchmark in algorithmic comparisons in ICA literature.

In the reporting period 2006 - 2007, ICA/BSS research stayed as one of the core projects in the laboratory, with the pure ICA theory somewhat waning and being replaced by several new directions. Chapter 3 starts by introducing some theoretical advances on the FastICA algorithm undertaken during the reporting period, followed by a number of extensions of ICA and BSS. The first one is the method of independent subspaces with decoupled dynamics, that can be used to model complex dynamical phenomena. The second extension is related to Canonical Correlation Analysis, and the third one is nonnegative separation by the new Projective Nonnegative Matrix Factorization (P-NMF) principle. An application of ICA to telecommunications is also covered. Then the Denoising Source Separation (DSS) algorithm is applied to climate data analysis. This is an interesting and potentially very useful application that will be under intensive research in the future in the group.

Another way to formulate the BSS problem is Bayesian analysis. This is covered in the separate Chapter 2.

3.2 Convergence and finite-sample behaviour of the FastICA algorithm

Erkki Oja

In Independent Component Analysis, a set of original source signals are retrieved from their mixtures based on the assumption of their mutual statistical independence. The simplest case for ICA is the instantaneous linear noiseless mixing model. In this case, the mixing process can be expressed as

$$\mathbf{X} = \mathbf{AS}, \tag{3.1}$$

where $\mathbf{X}$ is a $d \times N$ data matrix. Its rows are the observed mixed signals, thus d is the number of mixed signals and N is their length or the number of samples in each signal. Similarly, the unknown $d \times N$ matrix $\mathbf{S}$ includes samples of the original source signals. $\mathbf{A}$ is an unknown regular $d \times d$ mixing matrix. It is assumed square because the number of mixtures and sources can always be made equal in this simple model.

In spite of the success of ICA in solving even large-scale real world problems, some theoretical questions remain partly open. One of the most central questions is the theoretical accuracy of the developed algorithms. Mostly the methods are compared through empirical studies, which may demonstrate the efficacy in various situations. However, the general validity cannot be proven like this. A natural question is, whether there exists some theoretical limit for separation performance, and whether it is possible to reach it.

Sometimes the algorithms can be shown to converge in theory to the correct solution giving the original sources, under the assumption that the sample size N is *infinite*. In [1], the FastICA algorithm was analyzed from this point of view. A central factor in the algorithm is a nonlinear function that is the gradient of the ICA cost function. It may be a polynomial, e.g. a cubic function in the case of kurtosis maximization/minimization, but it can be some other suitable nonlinearity as well. According to [1], let us present an example of convergence when the nonlinearity is the third power, and the 2×2 case is considered for the mixing matrix $\mathbf{A}$ in model (3.1).

In the theoretical analysis a linear transformation was made first, so that the correct solution for the separation matrix $\mathbf{W}$ (essentially the inverse of matrix $\mathbf{A}$) is a unit matrix or a variant (permutation and/or sign change). Thus the four matrix elements of $\mathbf{W}$ convergence to zero or to ± 1. The FastICA algorithm boils down to an iteration $w_{t+1} = f(w_t)$ for all the four elements of the separation matrix. The curve in Figure 3.2 shows the iteration function $f(.)$ governing this convergence. It is easy to see that close to the stable points, the convergence is very fast, because the iteration function is very flat.

In practice, however, the assumption of infinite sample size is unrealistic. For *finite* data sets, what typically happens is that the sources are not completely unmixed but some traces of the other sources remain in them even after the algorithm has converged. This means that the obtained demixing matrix $\widehat{\mathbf{W}}$ is not exactly the inverse of $\mathbf{A}$, and the matrix of estimated sources $\mathbf{Y} = \widehat{\mathbf{W}}\mathbf{X} = \widehat{\mathbf{W}}\mathbf{AS}$ is only approximately equal to $\mathbf{S}$. A natural measure of error is the deviation of the so-called gain matrix $\mathbf{G} = \widehat{\mathbf{W}}\mathbf{A}$ from the identity matrix, i.e., the variances of its elements.

The well-known lower limit for the variance of a parameter vector in estimation theory is the Cramér-Rao lower bound (CRB). In [2], the CRB for the demixing matrix of the FastICA algorithm was derived. The result depends on the score functions of the sources,

$$\psi_k(s) = -\frac{d}{ds}\log p_k(s) = -\frac{p_k'(s)}{p_k(s)} \tag{3.2}$$

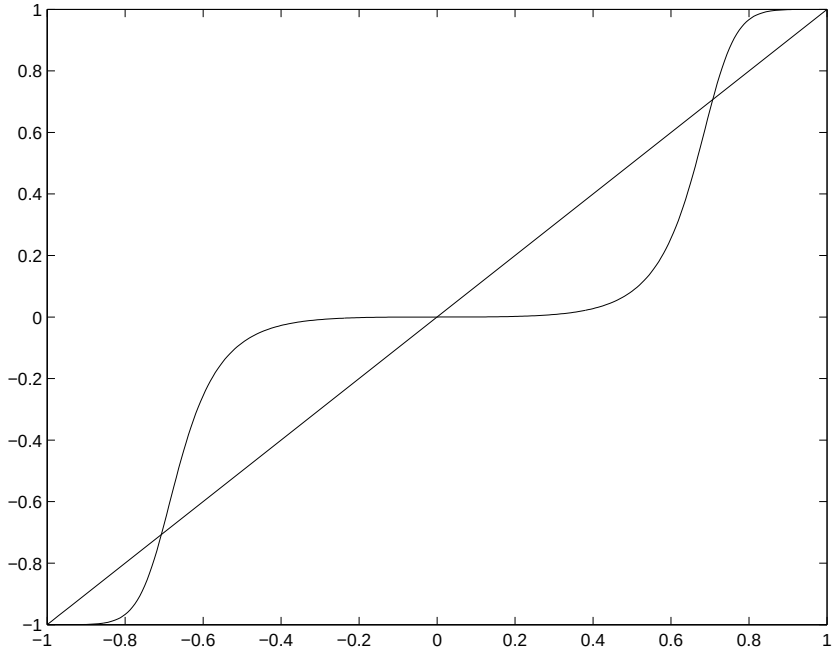


Figure 3.1: Shape of the iteration function for separation matrix elements, kurtosis case

where $p_k(s)$ is the probability density function of the k-th source. Let

$$\kappa_k = \mathrm{E}\left[\psi_k^2(s_k)\right]. \tag{3.3}$$

Then, assuming that the correct score function is used as the nonlinearity in the FastICA algorithm, the asymptotic variances of the off-diagonal elements (k, ℓ) of matrix $\mathbf{G}$ for the one-unit and symmetrical FastICA algorithm, respectively, read

$$V_{k\ell}^{1U-opt} = \frac{1}{N}\frac{1}{\kappa_k - 1} \tag{3.4}$$

$$V_{k\ell}^{SYM-opt} = \frac{1}{N}\frac{\kappa_k + \kappa_\ell - 2 + (\kappa_\ell - 1)^2}{(\kappa_k + \kappa_\ell - 2)^2}, \tag{3.5}$$

while the CRB reads

$$\mathrm{CRB}(\mathbf{G}_{k\ell}) = \frac{1}{N}\frac{\kappa_k}{\kappa_k\kappa_\ell - 1}. \tag{3.6}$$

Comparison of these results implies that the algorithm FastICA is nearly statistically efficient in two situations:

(1) One-unit version FastICA with the optimum nonlinearity is asymptotically efficient for $\kappa_k \to \infty$, regardless of the value of κ_ℓ.

(2) Symmetric FastICA is nearly efficient for κ_i lying in a neighborhood of 1^+, provided that all independent components have the same probability distribution function, and the nonlinearity is equal to the joint score function.

The work was continued to find a version of the FastICA that would be asymptotically efficient, i.e. able to attain the CRB. This can be achieved in the orthogonalization stage of the FastICA algorithm: instead of requiring strict orthogonalization, this condition is relaxed to allow small deviations from orthogonality, controlled by a set of free parameters. These parameters can be optimized so that the exact CRB is reached by the new algorithm, given that the correct score functions are used as nonlinearities.

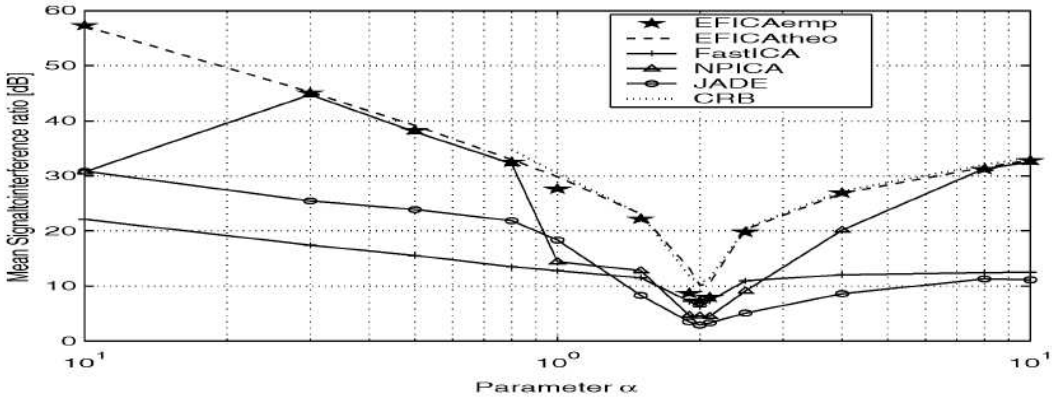


Figure 3.2: The Mean Signal-to Inference Ratio of EFICA, compared to CRB and some other ICA algorithms

The new efficient FastICA algorithm, dubbed EFICA, requires two phases because the score functions have to estimated first. Once they have been estimated, the new approximative orthogonalization scheme is run for a number of steps to reach the optimal solution. Figure 3.2 shows the efficiency of EFICA. To make meaningful comparisons, 13 source signals were artificially generated, each having a generalized gamma density $GG(\alpha)$ (where the value $\alpha = 2$ corresponds to the Gaussian density). The α values ranged from 0.1 to 10 and their places are marked by asterisks in the figure. The Mean Signal-to-Inference Ratio (SIR), both theoretical and experimental, obtained by EFICA is shown in the image (uppermost curve). It is very close to the Cramér-Rao Bound attainable in this situation, and far better than the Mean SIR attained by some other algorithms such as plain FastICA, NPICA, or JADE.

References

[1] Oja, E. and Yuan, Z.: The FastICA algorithm revisited – convergence analysis. *IEEE Trans. on Neural Networks 17*, no. 6, pp. 1370 - 1381 (2006).

[2] Tichavský, P., Koldovský, Z. and Oja, E.: Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *IEEE Trans. on Signal Processing 54*, no. 4, pp. 1189 - 1203 (2006).

[3] Koldovský, Z., Tichavský, P., and Oja, E.: Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Trans. on Neural Networks 17*, no. 5, pp. 1265 - 1277 (2006).

3.3 Independent subspaces with decoupled dynamics

Alexander Ilin

Independent subspace models extend the general source separation problem by allowing groups (subspaces) $\mathbf{s}_k$ of sources:

$$\mathbf{x}(t) = \sum_{k=1}^{K} \mathbf{A}_k \mathbf{s}_k(t)\,. \tag{3.7}$$

The sources within one group $\mathbf{s}_k$ are assumed dependent while signals from different groups are mutually independent. Similarly to classical BSS, subspaces can be separated exploiting non-Gaussianity or temporal structures of the mixed signals. The technique presented in [2] uses a first-order nonlinear model to model the dynamics of each subspace:

$$\mathbf{s}_k(t) = \mathbf{g}_k(\mathbf{s}_k(t-1)) + \mathbf{m}_k(t)\,, \quad k = 1, \ldots, K\,, \tag{3.8}$$

Both the de-mixing transformation and the nonlinearities $\mathbf{g}_k$ governing the dynamics are estimated simulateously by minimizing the mean prediction error of the subspace dynamical models (3.8). The optimization procedure can be implemented using the algorithmic structure of denoising source separation [1].

The algorithm was tested on artificially generated data containing linear mixtures of two independent Lorenz processes with different parameters, a harmonic oscillator and two white Gaussian noise signals (see Fig. 3.3). The algorithm is able to separate the three subspaces using only the information about their dimensionalities.

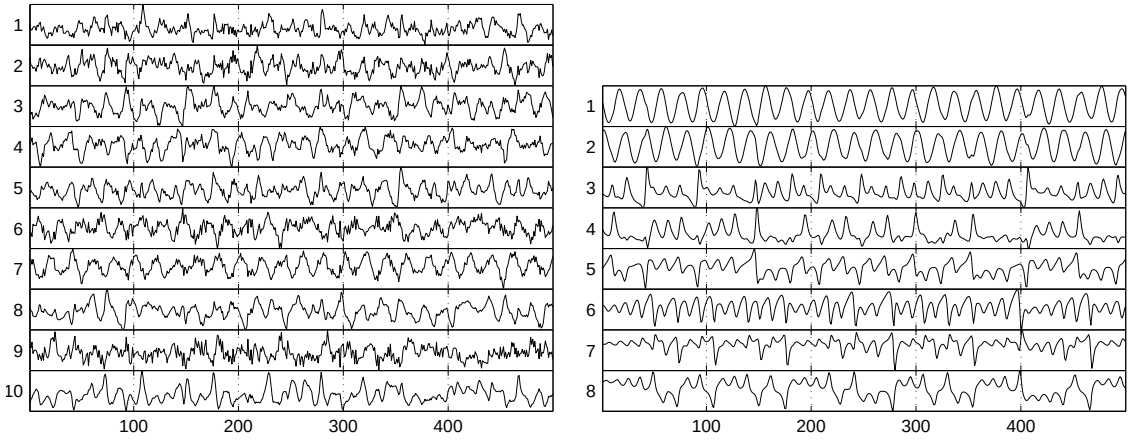


Figure 3.3: Left: Artificially generated linear mixtures of three dynamical processes and white noise signals. Right: Sources extracted by the technique extracting subspaces (signals 1–2, 3–5 and 6–9) with decoupled dynamics.

References

[1] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

[2] A. Ilin. Independent dynamics subspace analysis. In *Proc. of the 14th European Symposium on Artificial Neural Networks (ESANN 2006)*, pp. 345–350, April 2006.

3.4 Extending ICA for two related data sets

Juha Karhunen, Tomas Ukkonen

Standard linear principal component analysis (PCA) [2, 1] and independent component analysis (ICA) [1] are both based on the same type of simple linear latent variable model for the observed data vector $\mathbf{x}(t)$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{n} s_i(t)\mathbf{a}_i \tag{3.9}$$

In this model, the data vector $\mathbf{x}(t)$ is expressed as a linear transformation of the coefficient vector $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_n(t)]^T$. The column vectors $\mathbf{a}_i$, $i = 1, 2, \ldots, n$, of the transformation matrix $\mathbf{A}$ comprise the basis vectors of PCA or ICA, and the components $s_i(t)$ of the source vector $\mathbf{s}(t)$ are respectively principal or independent components corresponding to the data vector $\mathbf{x}(t)$. For simplicity, we assume that both the data vector $\mathbf{x}(t)$ and the source vector $\mathbf{s}(t)$ are zero mean n-vectors, and that the basis matrix $\mathbf{A}$ is a full-rank constant $n \times n$ matrix.

In PCA, the basis vectors $\mathbf{a}_i$ are required to be mutually orthogonal, and the coefficients $s_i(t)$ to have maximal variances (power) in the expansion (3.9) [2, 1]. While in ICA the basis vectors $\mathbf{a}_i$ are generally non-orthogonal, and the expansion (3.9) is determined under certain ambiguities from the strong but often meaningful condition that the coefficients $s_i(t)$ must be mutually statistically independent or as independent as possible [1].

Canonical correlation analysis (CCA) [2] is a generalization of PCA for two data sets whose data vectors are denoted by $\mathbf{x}$ and $\mathbf{y}$. CCA seeks for the linear combinations of the components of the vectors $\mathbf{x}$ and $\mathbf{y}$ which are maximally correlated. In this work, we have considered a similar expansion as (3.9) for both $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad \mathbf{y} = \mathbf{B}\mathbf{t} \tag{3.10}$$

We then try to find in a similar manner as in ICA the maximally independent and dependent components from $\mathbf{x}$ and $\mathbf{y}$ by using higher-order statistics. As a result, we get an ICA style counterpart for canonical correlation analysis.

These ideas are introduced in [3], and discussed in more detail in the journal paper [4]. The methods introduced in these papers are somewhat heuristic, buut seem to work adequately both for artificially generated data and in a difficult cryptographic problem. We also consider in these papers practical measures for statistical dependence or independence of two random variables.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001, 481+xxii pages.

[2] A. Rencher, *Methods of Multivariate Analysis, 2nd ed.* Wiley, 2002.

[3] J. Karhunen and T. Ukkonen. Generalizing independent component analysis for two related data sets. In *Proc. of the IEEE 2006 Int. Conf. on Neural Networks / 2006 IEEE World Congress on Computational Intelligence (IJCNN2006/WCCI2006)*, Vancouver, Canada, July 2006, pp. 1822-1829.

[4] J. Karhunen and T. Ukkonen, Extending ICA for finding jointly dependent components from two related data sets. *Neurocomputing*, Vol. 70, Issues 16-18, October 2007, pp. 2969-2769.

3.5 ICA in CDMA communications

Karthikesh Raju, Tapani Ristaniemi, Juha Karhunen, Erkki Oja

In wireless communication systems, like mobile phones, an essential issue is division of the common transmission medium among several users. A primary goal is to enable each user of the system to communicate reliably despite the fact that the other users occupy the same resources, possibly simultaneously. As the number of users in the system grows, it becomes necessary to use the common resources as efficiently as possible.

During the last years, various systems based on CDMA (Code Division Multiple Access) techniques [1, 2] have become popular, because they offer several advantages over the more traditional FDMA and TDMA schemes based on the use of non-overlapping frequency or time slots assigned to each user. Their capacity is larger, and it degrades gradually with increasing number of simultaneous users who can be asynchronous. On the other hand, CDMA systems require more advanced signal processing methods, and correct reception of CDMA signals is more difficult because of several disturbing phenomena [1, 2] such as multipath propagation, possibly fading channels, various types of interferences, time delays, and different powers of users.

Direct sequence CDMA data model can be cast in the form of a linear independent component analysis (ICA) or blind source separation (BSS) data model [3]. However, the situation is not completely blind, because there is some prior information available. In particular, the transmitted symbols have a finite number of possible values, and the spreading code of the desired user is known.

In this project, we have applied independent component analysis and denoising source separation (DSS) to blind suppression of various interfering signals appearing in direct sequence CDMA communication systems. The standard choice in communications for suppressing such interfering signals is the well-known RAKE detection method [2]. RAKE utilizes available prior information, but it does not take into account the statistical independence of the interfering and desired signal. On the other hand, ICA utilizes this independence, but it does not make use of the prior information. Hence it is advisable to combine the ICA and RAKE methods for improving the quality of interference cancellation.

In the journal paper [4], various schemes combining ICA and RAKE are introduced and studied for different types of interfering jammer signals under different scenarios. By using ICA as a preprocessing tool before applying the conventional RAKE detector, some improvement in the performance is achieved, depending on the signal-to-interference ratio, signal-to-noise ratio, and other conditions [4].

All these ICA-RAKE detection methods use the FastICA algorithm [3] for separating the interfering jammer signal and the desired signal. In the case of multipath propagation, it is meaningful to examine other temporal separation methods, too. We have also applied denoising source separation [5] to interference cancellation. This is a semi-blind approach which uses the spreading code of the desired user but does not require training sequences. The results of the DSS-based interference cancellation scheme show improvements over conventional detection.

All the results achieved in this project have been collected and presented in the monograph type doctoral thesis [6].

References

[1] S. Verdu, *Multiuser Detection*. Cambridge Univ. Press, 1998.

[2] J. Proakis, *Digital Communications*. McGraw-Hill, 3rd edition, 1995.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001, 481+xxii pages.

[4] K. Raju, T. Ristaniemi, J. Karhunen, and E. Oja, Jammer cancellation in DS-CDMA arrays using independent component analysis. *IEEE Trans. on Wireless Communications*, Vol. 5, No. 1, January 2006, pp. 77–82.

[5] J. Särelä and H. Valpola, Denoising source separation. *J. of Machine Learning Research*, Vol. 6, 2005, pp. 233–272.

[6] K. Raju, *Blind Source Separation for Interference Cancellation in CDMA Systems.* PhD Thesis, Helsinki Univ. of Technology, 2006. Published as Report D16, Laboratory of Computer and Information Science.

3.6 Non-negative projections

Zhirong Yang, Jorma Laaksonen, Zhijian Yuan, Erkki Oja

Projecting high-dimensional input data into a lower-dimensional subspace is a fundamental research topic in signal processing, machine learning and pattern recognition. Non-negative projections are desirable in many real-world applications where the original data are non-negative, consisting for example of digital images or various spectra. It was pointed out by Lee and Seung [1] that the positivity or non-negativity of a linear expansion is a very powerful constraint, that seems to lead to sparse representations for the data. Their method, *non-negative matrix factorization (NMF)*, minimizes the difference between the data matrix $\mathbf{X}$ and its non-negative decomposition $\mathbf{WH}$. The difference can be measured by the Frobenius matrix norm or the Kullback-Leibler divergence.

Yuan and Oja [2] proposed the *projective non-negative matrix factorization (P-NMF)* method which replaces $\mathbf{H}$ in NMF with $\mathbf{W}^T\mathbf{X}$. This actually combines the objective of principal component analysis (PCA) with the non-negativity constraint. The P-NMF algorithm has been applied to facial image processing [4] using a popular database, FERET [3]. Figure (3.4) visualizes the basis images learned by NMF and P-NMF. The empirical results indicate that P-NMF is able to produce more spatially localized, part-based representations of visual patterns.

Another attractive feature of the NMF and P-NMF methods is that their multiplicative update rules do not involve human-specified parameters such as the learning rate. Thus the analysis results are completely data driven. In [5] we have studied how to construct multiplicative update rules for non-negative projections based on Oja's iterative learning rule. Our method integrates the multiplicative normalization factor into the original additive update rule as an additional term which generally has a roughly opposite direction. As a consequence, the modified additive learning rule can easily be converted to its multiplicative version, which maintains the non-negativity after each iteration. With this technique, almost identical results to P-NMF can be obtained by imposing the non-negativity constraint on linear Hebbian networks.

The derivation of our approach provides a sound interpretation of learning non-negative projection matrices based on iterative multiplicative updates—a kind of Hebbian learning with normalization. A convergence analysis is provided by interpreting the multiplicative

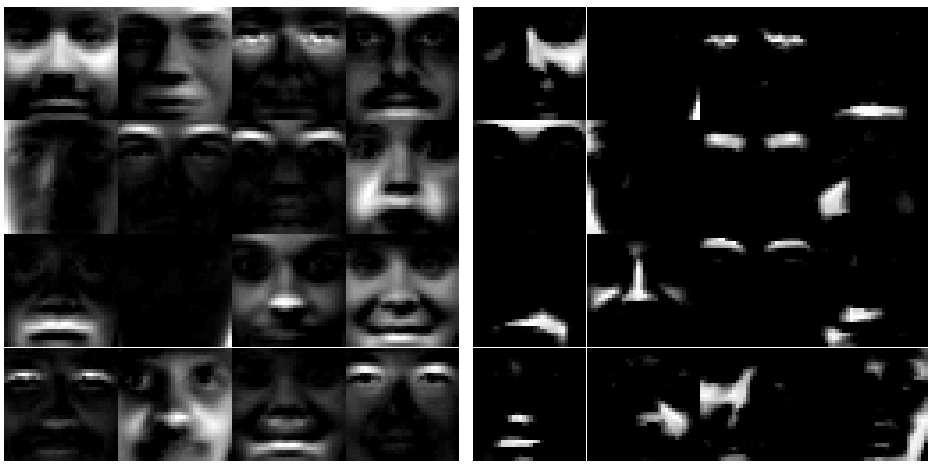


Figure 3.4: NMF (left) and P-NMF (right) bases of 16 dimensions.

updates as a special case of natural gradient learning. Furthermore, our non-negative variant of *linear discriminant analysis (LDA)* can serve as a feature selector. Its kernel extension can reveal an underlying factor in the data and be used as a sample selector.

References

[1] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[2] Zhijian Yuan and Erkki Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In *Proc. of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, pages 333–342, Joensuu, Finland, June 2005.

[3] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, October 2000.

[4] Zhirong Yang, Zhijian Yuan, and Jorma Laaksonen. Projective non-negative matrix factorization with applications to facial image processing. *International Journal on Pattern Recognition and Artificial Intelligence*, 21(8):1353–1362, December 2007.

[5] Zhirong Yang and Jorma Laaksonen. Multiplicative updates for non-negative projections. *Neurocomputing*, 71(1-3):363–373, 2007.

3.7 Climate data analysis with DSS

Alexander Ilin, Harri Valpola, Erkki Oja

An important task for which statistical methods are used in climate research is seeking physically meaningful interpretations of observed climate variability, for example, identification of 'modes' in the observational record. Statistical techniques which are widely used in this task include principal component analysis (PCA) or empirical orthogonal functions (EOFs), extended EOFs, and Hilbert EOFs [1]. Although EOFs have probably been the most popular tool for an efficient representation of climate records, EOF representation may be intuitively meaningless in a meteorological sense [2]. Therefore several techniques of rotated PCA/EOF have been proposed to ensure easier interpretation of the results. The rotation is realized using a linear transformation of principal components such that a suitably chosen criterion of "simple structure" is optimized. The objective is to find a data representation allowing for compact scientific explanation of a variable with a smaller number of principal components. Different assumptions on simplicity yield different rotation techniques.

We extend the concept of rotated PCA by introducing the concept of "interesting structure". In our case, the goal of exploratory analysis is to find signals with some specific structures of interest. They may for example manifest themselves mostly in specific variables, which exhibit prominent variability in a specific timescale etc. An example of such analysis can be extracting clear trends or quasi-oscillations from climate records. The procedure for obtaining suitable rotations of EOFs can be based on the general algorithmic structure of denoising source separation (DSS) [3].

In our initial studies, we tested the effectiveness of the proposed methodology to discover climate phenomena which are well-known in climatology, using very little information about their properties. One of the most prominent results is the extraction of the El Niño–Southern Oscillation phenomenon, using only a very generic assumption of its prominent variability in the interannual timescale (see Figs. 3.5-3.6) [4]. Other prominent signals found in this analysis might correspond to significant climate phenomena as well; for example, the second signal with prominent interannual variability somewhat resembles the derivative of the El Niño index (see Fig. 3.5).

Several other techniques for studying prominent climate variations have been introduced in our papers [4, 5]. Analysis which separates prominent quasi-oscillations in climate records by their frequency contents gives a meaningful representation of the slow climate variability as combination of trends, interannual oscillations, the annual cycle and slowly changing seasonal variations [4]. The technique presented in [5] can be used for studying slow variability present in fast weather fluctuations.

The results of the climate research were presented at the Fifth Conference on Artificial Intelligence Applications to Environmental Science as part of the 87th Annual Meeting of the American Meteorological Society (best student presentation) [6] and at the 10th International Meeting on Statistical Climatology.

References

[1] H. von Storch, and W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, U.K, 1999.

[2] M. B. Richman. Rotation of principal components. *Journal of Climatology*, 6:293–335, 1986.

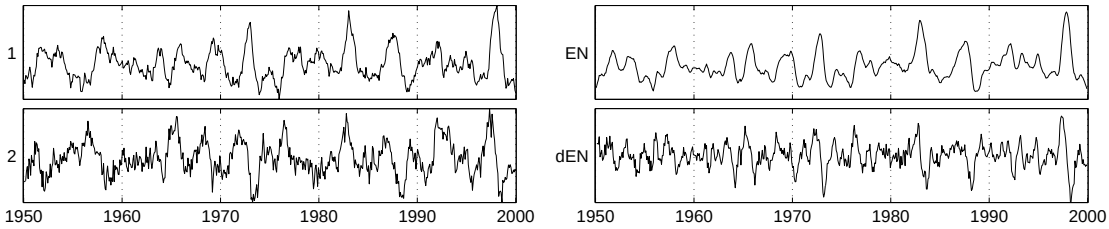


Figure 3.5: Left: The time courses of the two interannual phenomena found in global temperature, air pressure and precipitation data using DSS. Right: The index used in climatology to measure the strength of El Niño (marked as EN) and the derivative of the El Niño index (marked as dEN). The similarity is striking for the upper signals and some common features can be observed in the lower signals.

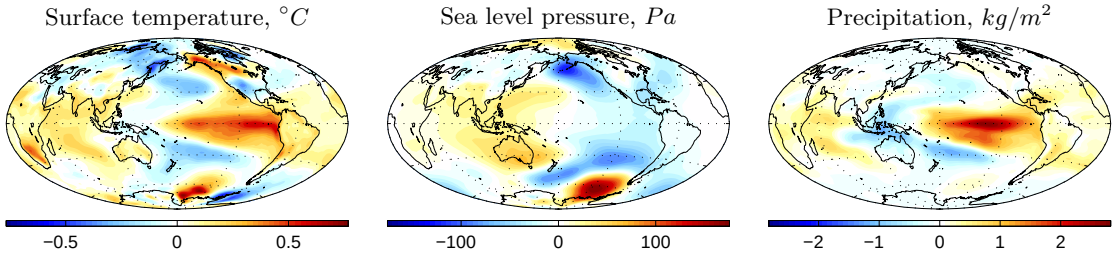


Figure 3.6: Spatial patterns corresponding to the most prominent interannual phenomenon found in climate data. The maps display the regions in which the effect of the phenomenon is most prominent. The maps contain many features traditionally associated with El Niño–Southern Oscillation phenomenon.

[3] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

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

[5] A. Ilin, and H. Valpola, and E. Oja. Extraction of components with structured variance. In *Proc. of the IEEE World Congress on Computational Intelligence (WCCI 2006)*, pp. 10528–10535, Vancouver, BC, Canada, July 2006.

[6] A. Ilin, and H. Valpola, and E. Oja. Finding interesting climate phenomena by exploratory statistical techniques. In *Proc. of the Fifth Conference on Artificial Intelligence Applications to Environmental Science as part of the 87th Annual Meeting of the American Meteorological Society*, San Antonio, TX, USA, January 2007. Best student presentation.

Chapter 4

Modeling of relevance

Samuel Kaski, Jaakko Peltonen, Kai Puolamäki, Janne Sinkkonen, Jarkko Venna, Arto Klami, Jarkko Salojärvi, Eerika Savia

4.1 Introduction

We develop statistical machine learning methods for extracting useful regularities from large, high-dimensional data sets. The key concept is *modeling of relevance*: data are usually full of patterns but the extracted ones should obviously be relevant to the analyst. An explicit definition of what is relevant is usually not known, and relevance needs to be inferred indirectly.

We have developed methods that use the structure of data in constraining which kinds of regularities are considered relevant. The structure here means several data sources or data sets. To make the task more concere, we have divided the ways of using the structure of data into three subtypes:

- *Relevance through data fusion* can mean two principal things: *dependency mining* and *supervised mining*, which are applicable in different settings. In both, several sources are combined with the goal of identifying relevant *aspects*, features or feature combinations, of data.

  In *dependency mining* or exploration, the aim is to decompose variation in each data source into source-specific and shared components. The within-source variation is assumed irrelevant, "noise", and only the shared effects are relevant. An example is measurement of several noisy signals from a common source, when characteristics of the noise are not known. More examples are given in Sections 5 and 6.

  While dependency mining is symmetric, in *supervised mining* a supervising auxiliary data set supervises the mining of primary data. Otherwise the methods are similar. If the supervising set consists of class labels of the primary data samples, the setup is *supervised unsupervised learning*. Our earlier research topic *learning metrics* was one suitable method for supervised unsupervised learning.

- *Relevant subtask learning* is a new research topic we introduced for addressing the problem of having too little representative or known-to-be-relevant training data. Given that other, partly or wholly irrelevant data sets are available, the relevant small data set is used as a "query" to retrieve more relevant data. At the same time, a model is built using all relevant data.

  This work can be seen as a special kind of asymmetric multi-task learning, or as combining information retrieval with multi-task learning.

- For *modeling of networks* we develop scalable models capable of dealing with uncertainty in network data. Networks are the simplest kinds of relational data, where the relations give hints of relevance.

These two general topics are useful in most of the modeling tasks above:

- *Discriminative generative modeling* describes how to use rigorous statistical modeling machinery for learning what is relevant to classes, and for making inference.

- *Information visualization* is a central subproblem in exploratory analysis and mining. We have introduced new very competitive nonlinear projection methods particularly suitable for projection to small dimensions for visualization.

4.2 Relevance through data fusion

Unsupervised data exploration or mining is defined as search for systematic properties, statistical structures or patterns from data. The findings need to be *relevant* as well, and typically relevance has been defined implicitly by selecting which kinds of patterns to find, which distance measures or features to use, and which model family to use. In general, relevance is defined by bringing in prior knowledge or assumptions to the task.

We have introduced methods for bringing in the prior information in a data-driven way, by choosing additional data sources and defining relevance through statistical dependencies between the sources. The underlying assumption is that aspects of data that are visible in one source only are "noise", whereas aspects visible in several sources describe the common thing of which all sources have different views. This will become clearer in the detailed descriptions below.

A straightforward way of finding the shared view is to build representations of data from each source, by maximizing the statistical dependency of the representations of different sources. We have developed both theory and practical methods for this task, and applied the methods in particular in neuro- and bioinformatics (Chapters 6 and 5).

Probabilistic models for detecting dependencies

Above the general approach was formulated in terms of maximizing a chosen dependency measure for mappings, that is, representations of the observations. We have previously introduced various methods for this task, including *associative clustering* and a linear projection method maximizing a non-parametric estimate of mutual information.

Recently we have studied an alternative formulation for the same task. One of the central problems in data analysis is overlearning, which means that models estimated with small data sets do not generalize well to new observations. One common solution to overlearning is to apply *Bayesian analysis* that allows treating uncertainties and choosing model complexity in a justified manner. Prior information can be rigorously incorporated to improve learning from small data sets, it is straightforward to extend models by changing distributional assumptions, and it is easy to construct larger models by combining submodels, at least in principle.

Bayesian tools can be applied to probabilistic models providing a generative description of the observed data. The methods for detecting dependencies between data sets are not, however, formulated as such models, and hence the Bayesian approach has not been possible for this task. We have introduced new theory on how such models can be built [3], and presented example models derived from the theory.

The proposed model family consists of latent variable models (see Fig. 4.1), where the observed data of each source is assumed to be an additive composition of two sources: one that is shared with the other data sources, and one that is specific to that particular data source. We have shown [3] that such models can extract the statistical dependencies to the shared latent source if a particular requirement is satisfied: the part of the model describing the data-source-specific variation in the observed data should be accurate enough.

Based on this basic principle we have re-derived an earlier probabilistic interpretation of canonical correlation analysis [1], and provided two novel models. In [3] a Bayesian clustering model for detecting dependencies is solved with variational Bayes approximation. The model is illustrated graphically in Figure 4.1. In [2] a Bayesian version of canonical correlation analysis is introduced, this time using Gibbs sampling for inference. Besides introducing a way of analyzing small-sample data to CCA the method lifts a critical restriction of classical CCA: the requirement of global linear dependency. It is overcome by

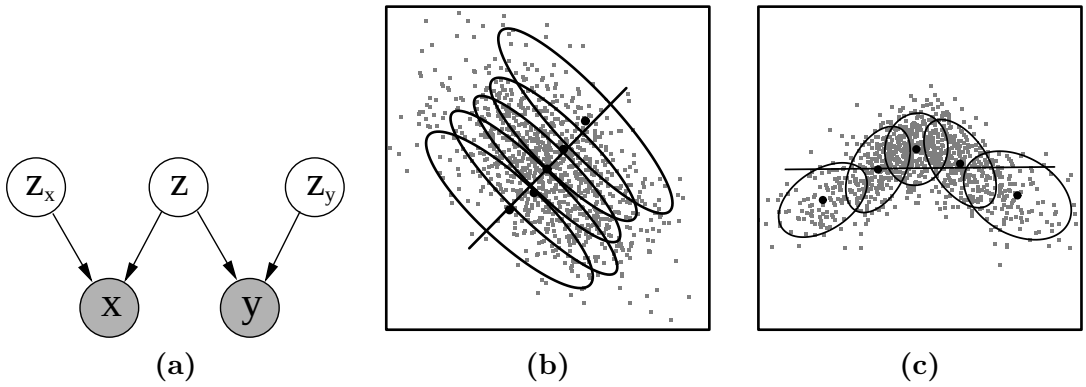


Figure 4.1: **(a):** A general latent-variable model structure for detecting statistical dependencies between **x** and **y**. **(b-c):** Illustration of a clustering version of the general model. The two panels show scatter-plots of two data sets having co-occurring samples. The lines depict linear dependency found by canonical correlation analysis. The clusters found by the clustering model have aligned according to the dependency, while still capturing nonlinear structure of the data in panel **(c)**.

introducing a Dirichlet process mixture model, allowing different kinds of dependencies in different parts of the data space.

Dependency with class variables

A common case of two data sources is class labels coupled with feature vectors. Standard classifiers use the dependencies between the sources to predict class labels; other applications include visualization, discriminative clustering or discriminative feature extraction. These tasks use the labels to guide unsupervised analysis of the features; we call them *supervised unsupervised learning*.

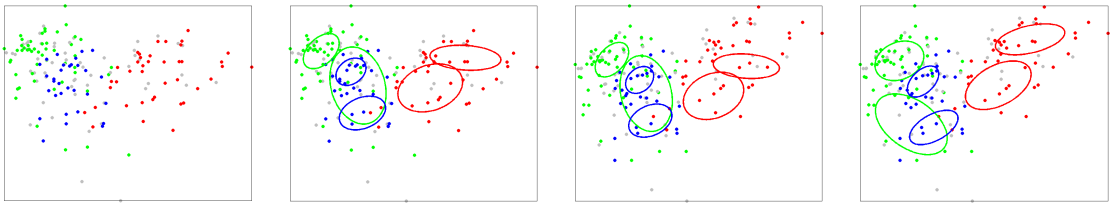
Recently we have studied a particular application of supervised unsupervised learning: fast learning of a class-discriminative subspace of data features. The subspace is defined by a linear transformation, and the features in the class-discriminative subspace are *discriminative components* of data. The subspace is useful for visualization, dimensionality reduction, feature extraction, and for learning a regularized distance metric.

Earlier we had learned such transformations with nonparametric estimation [5] which is accurate but slow; the computational complexity is $O(N^2)$ per iteration; here N is the number of samples. We now introduced a method that learns the linear transformation in a fast, semisupervised way [4], by optimizing a mixture model for classes in the subspace. The new method (Fig. 4.2) is fast ($O(N)$ per iteration) and semi-supervised, that is, can use unlabeled and pairwise-constrained data as well as labeled data.

References

[1] Fransic R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Tech. Rep 688, Department of Statistics, University of California, Berkeley, 2005.

[2] Arto Klami and Samuel Kaski. Local dependent components. In Zoubin Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning*, pages 425–432, 2007.



(a) (b) (c) (d)

Figure 4.2: Sample iterations of optimizing the discriminative subspace. Dots show data in the subspace; ellipses show the shape of mixture model components used to model the distribution in the subspace. There are three classes (red, green, blue) and unlabeled samples (gray dots). **(a):** Initial transformation. **(b):** The mixture model is optimized for the transformation. **(c):** The transformation is optimized for the mixture model. **(d):** The mixture model is optimized for the new transformation. The iteration continues in alternating steps.

[3] Arto Klami and Samuel Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, accepted for publication, 2008.

[4] Jaakko Peltonen, Jacob Goldberger, and Samuel Kaski. Fast Semi-supervised Discriminative Component Analysis. In Konstantinos Diamantaras, Tülay Adali, Ioannis Pitas, Jan Larsen, Theophilos Papadimitriou, and Scott Douglas, editors, *Machine Learning for Signal Processing XVII*, pages 312–317. IEEE, 2007.

[5] Jaakko Peltonen and Samuel Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16: 68–83, 2005.

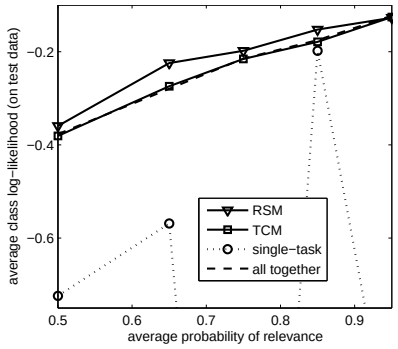


Figure 4.3: Relevant subtask learning model (RSM) outperforms a multi-task method that clusters tasks (TCM) and to two naive methods ("single-task" and "all together"), on news article data. The task was to predict relevance of news articles to a specific reader (the reader-of-interest), using articles rated by other readers as additional sources of information. Average results over 10 generated problems are shown, as a function of one experiment design parameter, the average probability that a sample is relevant to the reader-of-interest.

4.3 Relevant subtask learning

Having too little labeled training data is a common problem in classifier design. The problem is particularly hard for the high-dimensional data in genome-wide studies of modern bioinformatics, but appears also in image classification from few examples, finding of relevant texts, etc.

After realizing that the world is full of other data sets, the problem becomes how to simultaneously learn from a small data set and retrieve useful information from the other data sets. We have recently introduced a learning problem called *relevant subtask learning*, a variant of multi-task learning, which aims to solve the small-data problem by intelligently making use of other, potentially related "background" data sets.

Such potentially related "background" data sets are available for instance in bioinformatics, where there are databases full of data measured for different tasks, conditions or contexts; for texts there is the web. Such data sets are *partially relevant*: they do not come from the exact same distribution as future test data, but their distributions may still contain some useful part. Our research problem is, *can we use the partially relevant data sets to build a better classifier for the test data?*

Learning from one of the data sets is called a "task". Our scenario is then a special kind of *multi-task learning* problem. However, in contrast to typical multi-task learning, our problem is fundamentally asymmetric and more structured; test data fits one task, the "*task-of-interest*," and other tasks may contain *subtasks* relevant for the task-of-interest, but no other task needs to be wholly relevant.

In [1] we introduced a method that uses logistic regression classifiers. The key is to assume that each data set is a mixture of relevant and irrelevant samples. By fitting this model to all data sets, the common model for relevant samples learns from all tasks. We model the irrelevant part with a sufficiently flexible model such that irrelevant samples cannot distort the model for relevant data. A sample application is a news recommender for one user, where classifications from other users are available (Fig. 4.3). The relevant subtask learner outperforms a comparable standard multi-task learning model (related to [2]).

References

[1] Samuel Kaski and Jaakko Peltonen. Learning from relevant tasks only. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 608–615. Springer-Verlag, Berlin, Germany, 2007.

[2] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research*, 8: 35–63, 2007.

4.4 Discriminative generative modeling

The more traditional counterpart to supervised mining is *discriminative learning* where the data set is the same but the task is different. Given paired data $(\mathbf{x}, c)$, the task is to predict c for a test set where only the values of $\mathbf{x}$ are known.

There exist two traditional modeling approaches for predicting c, discriminative and generative. Discriminative models optimize the conditional probability $p(c|\mathbf{x})$ (or some other discriminative criterion) directly. The models are good classifiers since they do not waste resources on modeling those properties of the data that do not affect the value of c, that is, the marginal distribution of $\mathbf{x}$. The alternative approach is generative modeling of the joint distribution $p(c, \mathbf{x})$. Generative models add prior knowledge of the distribution of $\mathbf{x}$ into the task. This facilitates for example inferring missing values, since the model is assumed to generate also the covariates $\mathbf{x}$. The generative models are often additionally simpler to construct, and their parameters offer simple explanations in terms of expected sufficient statistics.

Discriminative Joint Density Models. In discriminative generative modeling we study discriminative inference given a generative model family $p(c, \mathbf{x}, \theta)$. The model family is assumed to be as good as possible but still known to be incorrect, and the objective is to obtain a distribution or point estimate that is optimal for predicting the values of c given $\mathbf{x}$. The Bayesian approach of using the posterior of the generative model family $p(c, \mathbf{x}, \theta)$ is not particularly well justified in this case, and it is known that it does not always generalize well to new data [1].

One way of learning discriminative classifiers is to take a joint density model, and then change the objective function from joint likelihood $\prod_i p(c_i, \mathbf{x}_i|\theta)$ to conditional likelihood $\prod_i p(c_i|\mathbf{x}_i, \theta)$. Earlier, we have presented an EM algorithm for obtaining discriminative point estimates [2]. The point estimate is (asymptotically) consistent for discrimination, given the model family. In [3] we proved that this applies for distributions as well; we derived an axiomatic proof that a *discriminative posterior* is consistent for conditional inference; using the discriminative posterior is standard practice in Bayesian regression, but we show that it is rigorous for model families of joint densities as well.

Compared to pure discriminative models, the benefit of the approach is that prior knowledge about $\mathbf{x}$ is brought in. The models operate in the same parameter space as ordinary discriminative models, but the generative formulation constrains the model manifold. Additionally, the density estimate for $\mathbf{x}$ from the model can be used for inferring missing values in the data [3].

References

[1] Peter D. Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2–3):119–149, 2007.

[2] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Expectation maximization algorithms for conditional likelihoods. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning (ICML-2005)*, pages 753–760, New York, USA, 2005. ACM press.

[3] Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, and Samuel Kaski. Discriminative MCMC. Report E1, Publications in Computer and Information Science, Helsinki University of Technology, 2006.

4.5 Visualization methods

Visualization of mutual similarities of entries in large high-dimensional data sets is a central subproblem in exploratory analysis and mining. It is makes sense to "look at the data" in all stages of data analysis, and reducing the dimensionality to two or three gives a scatterplot visualization.

When the intrinsic dimensionality of the data is higher than the dimensionality of the visualization, as is often the case, the visualization cannot represent the data flawlessly; some properties are necessarily lost or misrepresented. A compromise is unavoidable, but which compromise is the best for visualization? Many existing nonlinear dimensionality reduction methods practically ignore this question altogether, because they are not designed to reduce the dimensionality of the data set lower than is possible without losing information. Some methods choose the compromise implicitly in that they produce the lower-dimensional representation by minimizing a cost function, but the cost function has not been motivated from the point of view of visualization, that is, it is not obvious why a projection that minimizes the cost function should be a good visualization. We have filled this gap by introducing rigorously motivated measures for the quality of a visualization, as well as a nonlinear dimensionality reduction method that optimizes these measures and is therefore specifically designed for optimal visualization.

Visualization as information retrieval

We view visualization as an information retrieval task. Consider an analyst studying a scatterplot of countries, organized according to their welfare indicators. Being interested in Finland, she wants to know which other countries are similar. The visualization helps in this task of retrieving similar items, and quality of retrieval can be measured with standard information retrieval measures *precision* and *recall*. Any information retrieval method needs to make a compromise between these measures, parameterized by the relative cost of false positives and misses. Since a visualizer is an information retrieval device as well, it needs to make the same compromise.

We have adapted the information retrieval measures to visualization by smoothing them and representing them as differences between distributions of points being neighbors. It turns out that the traditional measures are limiting cases of these more general measures. Once the relative cost λ of false positives and misses has been fixed, we can directly optimize the visualization to minimize the retrieval cost. We call the resulting visualization method the Neighborhood Retrieval Visualizer (NeRV) [1].

The NeRV is a further development of our earlier method *local multidimensional scaling* [2], a faster method where the trade-off between precision and recall was heuristic and hence the results were less accurate.

Later we added the Self-Organizing Map to the comparison [3]. The SOM was very good in terms of (smoothed) precision, even producing a slightly better result than NeRV in some cases. In terms of recall the SOM performed poorly.

References

[1] Jarkko Venna and Samuel Kaski. Nonlinear dimensionality reduction as information retrieval. In Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS*07), San Juan, Puerto Rico, March 21-24, 2007.

[2] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19:889–899, 2006.

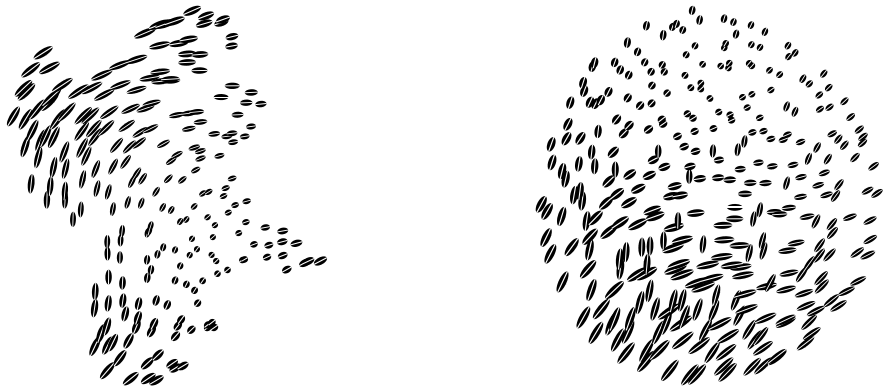


Figure 4.4: Two nonlinear projections of data that lies on the surface of a three-dimensional sphere. One of the input coordinates governs the rotation of the glyphs, the second their scale, and the third their degree of elongation. Hence, points having similar glyphs are close to each other in the input space. On the *left*, precision has been maximized; the sphere has become split open and the glyphs change smoothly, but on the opposite ends of the projection there are similar glyphs that are projected far from each other. On the *right*, recall has been maximized and the sphere has become squashed flat. There are areas where the different kinds of glyphs are close to each other, but there are no areas where similar glyphs are very far from each other.

[3] Kristian Nybo, Jarkko Venna and Samuel Kaski. The Self-Organizing Map as a Visual Neighbor Retrieval Method. In *Proceedings of 6th Int. Workshop on Self-Organizing Maps (WSOM '07)*. Bielefeld University, Bielefeld, Germany, 2007.

4.6 Networks

Machine Learning is in the midst of a "structural data revolution". After many decades of focusing on independent and identically-distributed examples, many researchers are now modelling inter-related entities that are linked together into complex graphs. A major driving force is the explosive growth of heterogeneous data collected on diverse sectors of the society. Example domains include bioinformatics, communication networks, and social network analysis.

Networks are a special case of structural data. Inferring properties of the network nodes, or vertices, from the links, or edges, has become a common data mining problem. Network data are typically not a complete description of reality but come with errors, omissions and uncertainties. Some links may be spurious, for instance due to measurement noise in biological networks, and some potential links may be missing, for instance friendship links of newcomers in social networks. Probabilistic generative models are a tool for modeling and inference under such uncertainty. They treat the links as random events, and give an explicit structure for the observed data and its uncertainty. Compared to non-stochastic methods, they are therefore likely to perform well as long as their assumptions are valid; they may reveal properties of networks that are difficult to observe with non-statistical techniques from the noisy and incomplete data, and they also offer a groundwork for new conceptual developments.

Component models for large networks

Being among the easiest ways to find meaningful structure from discrete data, Latent Dirichlet Allocation (LDA) and related component models have been applied widely. They are simple, computationally fast and scalable, interpretable, and admit flexible nonparametric priors. In the currently popular field of network modeling, relatively little work has taken uncertainty of data seriously in the Bayesian sense, and component models have been introduced to the field only recently. We have developed a component model of networks that finds community-like structures like the earlier methods motivated by physics. With Dirichlet Process priors and an efficient implementation the models are highly scalable.

References

[1] Janne Sinkkonen, Janne Aukia, and Samuel Kaski. Inferring vertex properties from topology in large networks. In *The 5th International Workshop on Mining and Learning with Graphs (MLG'07)*, Florence, Italy, 2007. Universita Degli Studi di Firenze.

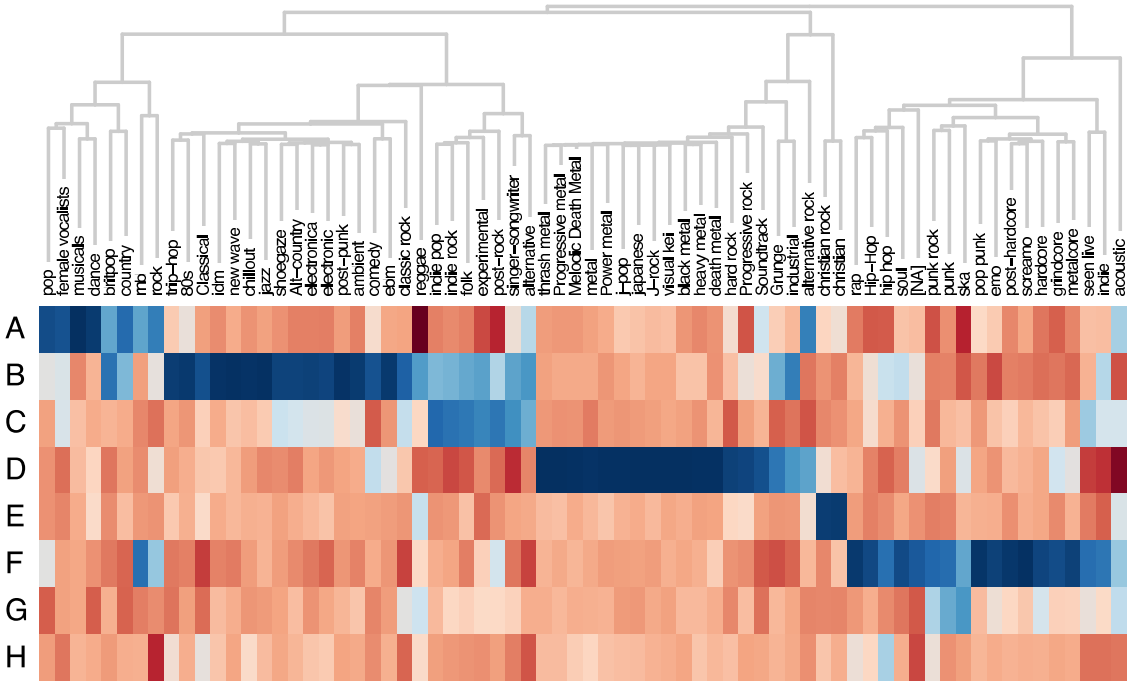


Figure 4.5: Last.fm is an Internet site that learns the musical taste of its members on the basis of examples, and then constructs a personalized, radio-like music feed. The web site also has a richer array of services, including a possibility to announce friendships with other users. The friendship network alone, when divided into components, reveals musical structures, because the music tastes of friends tend to be similar. Here the latent components found by our model were afterwards correlated with user's listening habits; songs are aggregated by tags given to them. Tags are intuitively grouped into genres. The network has 147,000 nodes and 353,000 links, but the running time with an efficient implementation by our collaborators at Xtract Ltd. was just 8.4 hours.

Bioinformatics and Neuroinformatics

Chapter 5

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja, Jaakko Peltonen, Jarkko Venna, Antti Ajanki, Andrey Ermolov, Ilkka Huopaniemi, Arto Klami, Leo Lahti, Jarkko Salojärvi, Abhishek Tripathi

5.1 Introduction

New so-called high-throughput measurement techniques have made possible genome-wide studies of gene function. Gene expression, gene regulation, protein content, protein interaction, and metabolic profiles can be measured and combined with the genetic sequence. The methods are used routinely in modern biology and medicine, and now the current challenge is to extract meaningful findings from the noisy and incomplete data masses, collected into both community resource and private data banks. The data needs to be analyzed, mined, understood, and taken into account in further experiments, which makes data analysis an integral part of biomedical research. Successful genome-wide analyses would allow a completely novel systems-level view into a biological organism.

Combining the different kinds of data produces new systems-level hypotheses about gene function and regulation, and ultimately functioning of biological organisms. We develop probabilistic modeling and statistical data analysis methods to advance this field. Our main novel contributions stem from the cross-breeding of the methodological basic research, in particular on Modeling of Relevance, and collaboration with top groups in Biology and Medicine. We have had long-standing collaboration with Laboratory of Cytomolecular Genetics (Prof. S. Knuutila) and Neuroscience Center (Prof. E. Castrén), University of Helsinki, University of Uppsala (Prof. J. Blomberg), Turku Centre for Biology (Doc. T. Aittokallio), VTT (Prof. M. Oresic), and smaller-scale collaboration with several other groups. During 2007 we started new projects with EBI, UK (A. Brazma) and Finnish CoE in Plant Signal Research, University of Helsinki (Prof. J. Kangasjärvi) with promising results that will be reported in the next biennial report.

In 2006 we started a new conference series in collaboration with Prof. E. Ukkonen and J. Rousu of University of Helsinki. The conference "Probabilistic Modeling and Machine Learning in Structural and Systems Biology" inspired a special issue in a main journal, and yearly conferences in Evry, France, in 2007, and in Belgium in 2008.

References

[1] Juho Rousu, Samuel Kaski, and Esko Ukkonen, editors. *Probabilistic Modeling and Machine Learning in Structural and Systems Biology. Workshop Proceedings; Tuusula, Finland, June 17-18*. Helsinki, Finland, 2006.

[2] Samuel Kaski, Juho Rousu, and Esko Ukkonen. Probabilistic modeling and machine learning in structural and systems biology. *BMC Bioinformatics*, 8(Suppl 2):S1, 2007.

5.2 Translational medicine on metabolical level

Translational medicine is a research field which attempts to more directly bring basic research findings to clinical practice. One of the necessary steps of this process is to translate inferences made on the molecular level, for example about metabolites, in model organisms into inferences about humans. Such translation is extremely challenging and the existing knowledge, if there is any, is currently largely tacit and only known to experts of the specific disease and model organism.

Metabolomics is the study of the set of all metabolites found in a sample tissue. Metabolite concentrations are affected strongly by diseases and drugs, and hence they complement the genomic, proteomic, and transcriptomic measurements in an excellent way, in studies of the biological state of an organism.

We are in the process of developing new computational methods for translational medicine, for mapping between the observed metabolomics data from model organisms and humans. In project TRANSCENDO we apply the methods to studies of the emergence of Type I diabetes, by computing mappings between non-obese diabetic (NOD) mice and children, and between the effects of a disease in several tissues. The project is collaboration within a consortium involving computational systems biology (Matej Oresic, VTT), semantic modelling (Antti Pesonen, VTT), probabilistic modelling (us), and pharmacology and animal models of metabolic disease (Eriika Savontaus, University of Turku).

Metabolomic development in humans

Metabolic development of children and its differences between the genders is not yet well understood. These dynamic changes may, however, affect strongly the susceptibility to diseases and the responses to drugs.

We are studying a metabolomic data set derived from a collection of blood samples collected during the first years of life from boys and girls. We assume that the metabolic profiles are generated by a set of unobserved metabolic states, and we model those states and the data with a Hidden Markov Model (HMM). HMM fits the assumption of latent states very well and is easy to compute and interpret. Moreover, HMM provides a way for probabilistic re-alignment of the time series, which takes into account the individual variation in the dynamics. Simulations have indicated that HMMs can separate the boys' and girls' metabolic states more efficiently apart than traditional linear method; classification accuracy is 73% for HMM, and under 60% for linear methods. Figure 5.1 presents the model structures for girls and boys.

Disease-related dependencies between multiple tissues

A common setting in medical research is that a disease may be mainly located in a specific organ, for example in lungs, but it indirectly affects multiple tissues. Giving drugs to patients induces an analogous setup: the drugs may affect multiple other tissues in addition to the target tissue (and hence disease). We are developing new methods for discovering the disease-related metabolic dependencies between the multiple tissues, with the goal of revealing potential side effects of the diseases and drugs.

In practice, we have metabolomics data from mice belonging to 4 classes: healthy and untreated, sick and untreated, healthy and treated, sick and treated. A fast and straightforward way of digging out disease-related dependencies is to first find disease-related aspects with partial least-squares classifiers, and then dependencies with canonical correlation analyses and more straightforward correlations between contributing metabolites.

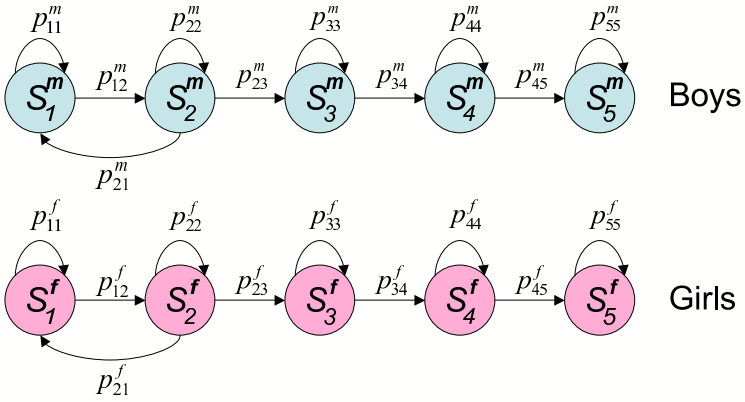


Figure 5.1: HMM models for metabolic states in boys and girls. The nodes represent hidden metabolic states, and the arrows possible transitions. Note that the states form a chain in order to force the models to focus on progressive changes in metabolite concentrations.

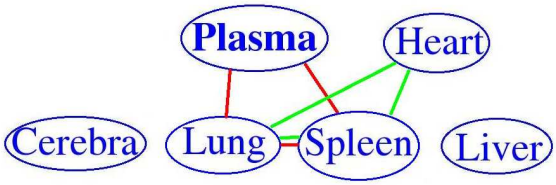


Figure 5.2: Disease-related dependencies between tissues before treatment (red), and after treatment (green). The disease is located in the lungs so the dependencies between lungs and plasma and spleen are logical, but note that after the treatment the dependency with plasma disappears and a dependency to heart emerges. This might be a sign of a side effect of the treatment.

This multivariate approach complements the traditional metabolite-wise linear models. Figure 5.2 shows the dependencies found between tissues before and after drug treatment.

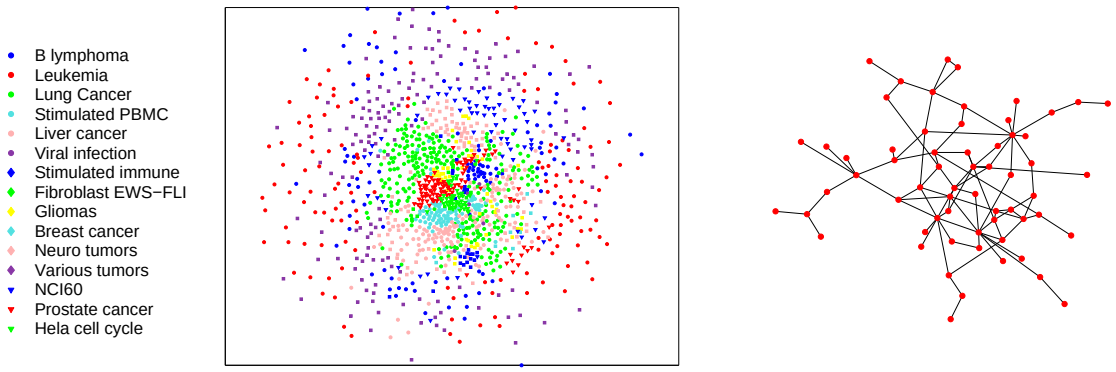


Figure 5.3: *Left:* Sample visualization of a gene expression atlas of cancer samples by curvilinear component analysis. Each dot denotes one microarray; the colors show the cancer class of the sample. *Right:* Part of yeast gene regulatory interaction network visualized by local multidimensional scaling.

5.3 Visualizing gene expression and interaction data

A large community-resource or private gene expression databank consists of numerous data sets submitted by several parties. They may have been measured for different purposes, with different treatments and methods in different laboratories. Several such databanks have been established and they continue to grow. A key challenge is how to best use the databanks to support further research. Currently information in these databanks is accessed using queries on the imperfect meta-data, that is, textual annotations and descriptions. In the future more sophisticated search methods, that take the actual data into account, are needed. Our study [2] aimed at comparing the different methods applicable as a visual interface that reveals similarities of data sets.

We compared several different visualization methods in the task of visualizing a large collection of gene expression arrays. Several new methods have been recently proposed for the estimation of data mani- folds or embeddings, but they have so far not been compared in the task of visualization. In visualizations the dimensionality is constrained, in addition to the data itself, by the presentation medium. It turned out that an older method, curvilinear components analysis, outperforms the new ones in terms of trustworthiness of the projections. Even though the standard preprocessing methods still need to be improved to make measurements of different labs and platforms more commensurable, the good news is that the visualized overview, expression atlas, reveals many of the cancer subsets (Fig. 5.3). Hence, we conclude that dimensionality reduction even from 1339 to 2 can produce a useful interface to gene expression databanks.

Biological high-throughput data sets can also be visualized as graphs that represent the relations between the biological entities. We applied our visualization methods for visualizing gene interaction graphs, and showed that Local Multidimensional Scaling performs very well in this task (Fig. 5.3; [1]).

References

[1] Jarkko Venna and Samuel Kaski. Visualizing Gene Interaction Graphs with Local Multidimensional Scaling In *Proceedings of ESANN'06, 14th European Symposium on Artificial Neural Networks*, pages 557–562, d-side, Evere, Belgium, 2006.

[2] Jarkko Venna and Samuel Kaski. Comparison of visualization methods for an atlas of gene expression data sets *Information Visualization*, 6:139–154, 2007.