Chapter 2

Bayesian learning of latent variable models

Juha Karhunen, Antti Honkela, Tapani Raiko, Alexander Ilin, Koen Van Leemput, Jaakko Luttinen, Matti Tornio, Markus Harva

2.1 Bayesian modeling and variational learning

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 compared alternative popular algorithms. After this we consider variational Bayesian learning of nonlinear state-space models, which are applied to model predictive control. This is followed by extensions of probabilistic principal component analysis (PCA) to binary PCA, missing values and achieving robustness in the presence of outliers. We then consider time series modeling in bioinformatics to learn gene regulatory relationships from time series expression data, as well as climate data analysis using Gaussian processes. We have also applied Bayesian methods to the astronomical data analysis problem of estimating time delays in gravitational lensing, as well as to medical image computing, focusing there on model-based segmentation and registration of magnetic resonance images of the brain. In most of these topics, we used variational approximations.

2.2 Algorithmic improvements for variational inference

Natural conjugate gradient

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.

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.

Transformation of latent variables

Variational methods have been used for learning linear latent variable models in which observed data vectors $\mathbf{x}(t)$ are modeled as linear combination of latent variables $\mathbf{s}(t)$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \boldsymbol{\mu} + \mathbf{n}(t), \quad t = 1, \ldots, N. \tag{2.3}$$

The latent variables are assigned some prior distributions, such as zero-mean Gaussian priors with uncorrelated components in the basic factor analysis model. When VB learning is used, the true posterior probability density function (pdf) of the unknown variables is

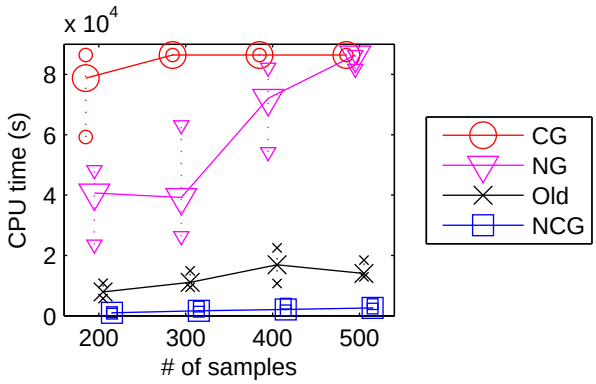


Figure 2.1: Convergence speed of the natural conjugate gradient (NCG), the natural gradient (NG) and the conjugate gradient (CG) methods as well as the heuristic algorithm (Old) with different data sizes. The lines show median times with 25 % and 75 % quantiles shown by the smaller marks. The times were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

approximated using a tractable pdf factorized as follows:

$$p\left(\boldsymbol{\mu}, \mathbf{A}, \mathbf{s}(1), ..., \mathbf{s}(N) \mid \{\mathbf{x}(t)\}\right) \approx q(\boldsymbol{\mu})q(\mathbf{A})q(\mathbf{s}(1)) \ldots q(\mathbf{s}(N)) \,.$$

This form of the posterior approximation q ignores the strong correlations present between the variables, which often causes slow convergence of VB learning.

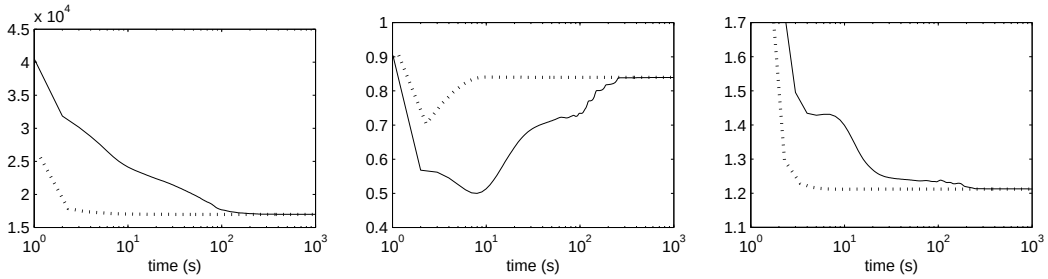
Parameter-expanded VB (PX-VB) methods were recently proposed to address the slow convergence problem [9]. The general idea is to use auxiliary parameters in the original model to reduce the effect of strong couplings between different variables. The auxiliary parameters are optimized during learning, which corresponds to *joint* optimization of different components of the variational approximation of the true posterior. In this way strong functional couplings between the components are reduced and faster convergence is facilitated. One of the main challenges for applying the PX-VB methodology is to use proper reparameterization of the original model.

In our recent conference paper [10], we present a similar idea in the context of VB learning of factor analysis models. There we use auxiliary parameters $\mathbf{b}$ and $\mathbf{R}$ which translate and rotate the latent variables:

$$\begin{aligned} \mathbf{s}(t) &\leftarrow \mathbf{s}(t) - \mathbf{b} & \boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} + \mathbf{A}\mathbf{b} \\ \mathbf{s}(t) &\leftarrow \mathbf{R}\mathbf{s}(t) & \mathbf{A} &\leftarrow \mathbf{A}\mathbf{R}^{-1} \,. \end{aligned}$$

The optimal parameters $\mathbf{b}$ and $\mathbf{R}$ which minimize the misfit between the posterior pdf and its approximation can then be computed analytically. This corresponds to joint optimization of factors $q(\mathbf{s}(t))$. In our paper, we show that the proposed transformations essentially perform centering and whitening of the hidden factors taking into account their posterior uncertainties.

We tested the effect of the proposed transformations by applying the VB PCA model to an artificial dataset consisting of $N = 200$ samples of normally distributed 50-dimensional vectors $\mathbf{x}(t)$. Figure 2.2 shows the minimized VB cost and the root mean squared error (RMSE) computed on the training and test sets during learning. The curves indicate that the method first overfits providing a solution with an unreasonably small RMSE. Later, learning proceeds toward a better solution yielding smaller test RMSE. Note that using



(a) VB cost (b) Training RMSE (c) Test RMSE

Figure 2.2: Convergence of VB PCA tested on artificial data. The dotted and solid curves represent the results with and without the proposed transformations, respectively.

the proposed transformations reduced the overfitting effect at the beginning of learning, which led to faster convergence to the optimal solution.

2.3 Nonlinear state-space models for model-predictive control

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

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_f) + \mathbf{n}(t)\,. \tag{2.4}$$

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

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1), \boldsymbol{\theta}_g) + \mathbf{m}(t)\,, \tag{2.5}$$

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.

The nonlinear functions are modelled by MLP networks. 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:

$$\mathbf{g}(\mathbf{s}(\mathbf{t}-\mathbf{1})) = \mathbf{s}(t-1) + \mathbf{D}\tanh[\mathbf{C}\mathbf{s}(t-1) + \mathbf{c}] + \mathbf{d}\,. \tag{2.6}$$

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 package is available under the name nonlinear dynamical factor analysis on the home page of our Bayes group [11].

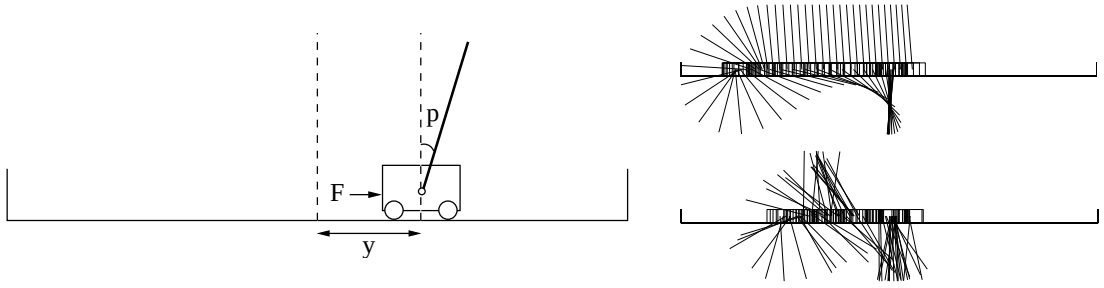


Figure 2.3: Left: The cart-pole system. The goal is to swing the pole to an upward position and stabilize it without hitting the walls. The cart can be controlled by applying a force to it. Top left: The pole is successfully swinged up by moving first to the left and then right. Bottom right: Our controller works quite reliably even in the presence of serious observation noise.

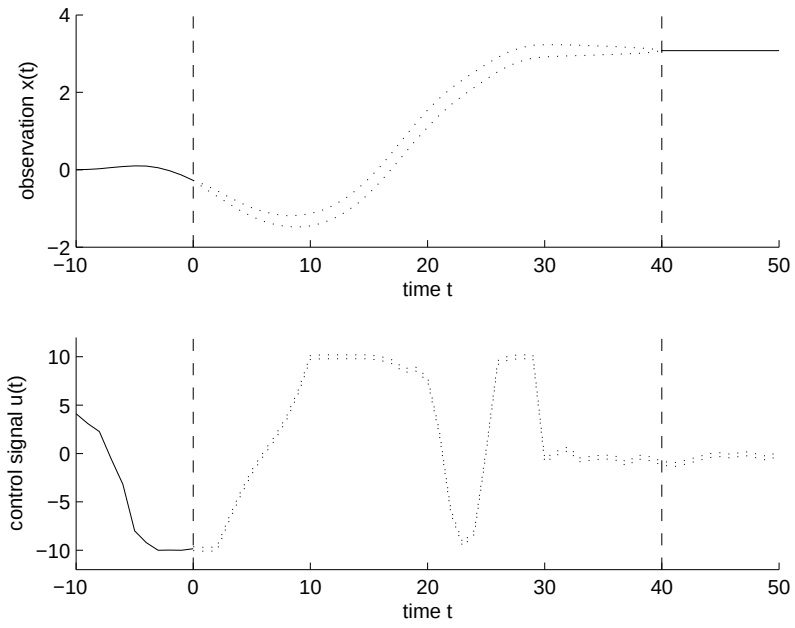


Figure 2.4: Optimistic inference control is a novel way of doing model predictive control, where we assume that the goal state has been reached after some window of uncertainty. The hidden states, observations and control signals are inferred using Bayesian inference methods. This approach bridges the gap between model-predictive control and Bayesian inference and thus algorithic developments on one side can be applied on the other side. The inferred observations and control signals are plotted with confidence intervals. The current time is $t_0 = 0$ and after time $t_0 + T_c = 40$, the observation $\mathbf{x}(t)$ is assumed to be at the desired level $\mathbf{r}(t)$.

In [15], 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 as part of the hidden state. Then we can use stochastic nonlinear model-predictive control, which is based on optimising control signals based on maximising a utility function. Figure 2.3 shows a simulation with an alternative method for model-predictive control.

The results with a simulated cart-pole swing-up task 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.

2.4 Extensions of probabilistic PCA

PCA of large-scale datasets with many 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 [16, 17], 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 (MAPPCA) or variational Bayesian (VBPCA) learning.

In [16, 17], we study different approaches to PCA for incomplete data. We show that faster convergence can be achieved using the following rule for the model parameters:

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

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 MAPPCA and VBPCA.

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 used different variants of PCA in order to predict the test ratings in the Netflix data set. The obtained results are shown in Figure 2.5. The best accuracy was obtained using VB PCA with a simplified form of the posterior approximation (VBPCAd in Figure 2.5). That method was also able to provide reasonable estimates of the uncertainties of the predictions.

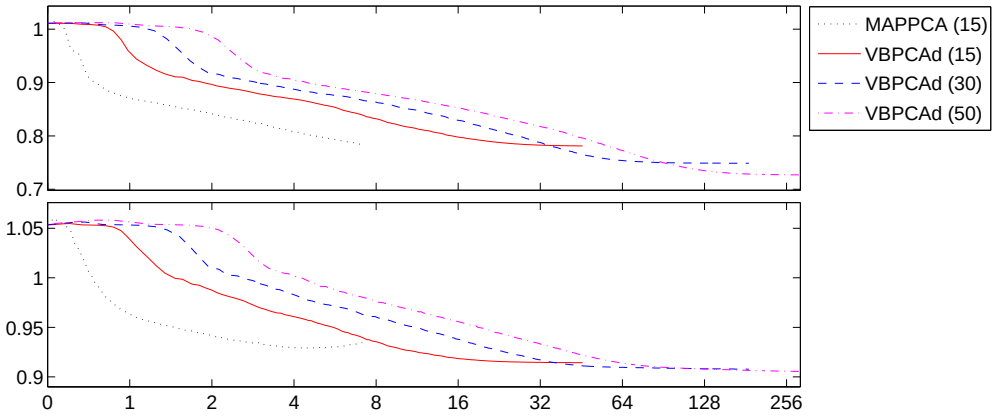


Figure 2.5: Root mean squared errors for the Netflix data (y-axis) plotted against the processor time in hours. The upper plot shows the training error while the lower plot shows the error for the probing data provided by Netflix. The time scale is linear from 0 to 1 and logarithmic above 1.

Binary PCA for collaborative filtering

In [18], we proposed an algorithm for binary principal component analysis that scales well to very high dimensional and very sparse data. Binary PCA finds components from data assuming Bernoulli distributions for the observations. The probabilistic approach allows for straightforward treatment of missing values.

We applied the proposed method to the same collaborative filtering problem prepared by Netflix. The collected ratings can be represented in the form of a matrix $\mathbf{X}$ in which each column contains ratings given by one user and each row contains ratings given to one movie. As a preprocessing step, the ratings were encoded with binary values, according to the following scheme:

$$
\begin{aligned}
1 &\rightarrow 0000 \\
2 &\rightarrow 0001 \\
3 &\rightarrow 0011 \\
4 &\rightarrow 0111 \\
5 &\rightarrow 1111
\end{aligned}
$$

With this scheme, each element in the data tells whether a rating is greater or smaller than a particular threshold.

We model the probability of each element x_{ij} of $\mathbf{X}$ to be 1 using the following formula:

$$P(x_{ij} = 1) = \sigma(\mathbf{a}_i^T \mathbf{s}_j) \tag{2.7}$$

where $\mathbf{a}_i$ and $\mathbf{s}_j$ are parameter vectors (both contain c elements) corresponding to the i-th movie and j-th user, respectively. The parameters $\mathbf{a}_i$ and $\mathbf{s}_j$ are assigned Gaussian priors and they are estimated from on the available ratings using the MAP method.

The results with the proposed binary PCA algorithm are slighlty worse than the ones obtained with PCA. However, by blending the two approaches, we were able to improve our previously best results obtained with PCA alone [16, 17]. Figure 2.6 shows the predictions of binary PCA against traditional PCA on a smaller MovieLens data set. The difference between the predictions suggests that the two methods model the data differently and blending them can improve the overall prediction performance.

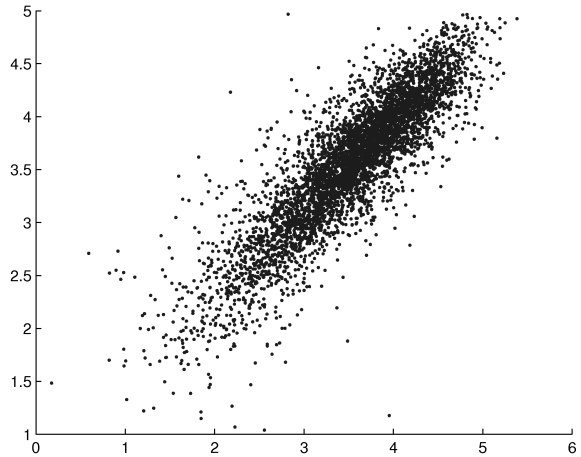


Figure 2.6: Predictions on a test set from the MovieLens data using PCA (x-axis) and the binary PCA model (y-axis). Note that PCA gives predictions outside the allowed range 1 to 5, whereas the predictions of binary PCA fall between 1 and 5 by construct.

Robust PCA for incomplete data

Standard PCA is known to be sensitive to outliers in the data because it is based on minimisation of a quadratic criterion such as the mean-square representation error. Thus, corrupted or atypical observations may cause the failure of PCA, especially for data sets with missing values. A standard way to cope with this problem is replacing the quadratic cost function of PCA a function which grows more slowly.

In [19], we present a new robust PCA model based on the Student-t distribution and show how it can be identified for data sets with missing values. We make the assumption that the outliers can arise independently in each sensor (i.e. for each dimension of a data vector). This assumption is different to the previously introduced techniques [21] and it turns out to be important for modeling incomplete data sets. The proposed model can improve the quality of the principal subspace estimation and provide better reconstructions of missing values. The model can also be used to remove outliers by estimating the true values of their corrupted components from the uncorrupted ones.

We tested the robust PCA model on the Helsinki Testbed data set which at the moment of our studies contained many atypical measurements and missing values. The model was used to estimate four principal components of the temperature measurements from 79 stations in Southern Finland. Figure 2.7 presents the reconstruction of the data using our robust PCA model for four different stations. The reconstructions look very reasonable with most of the outliers being removed.

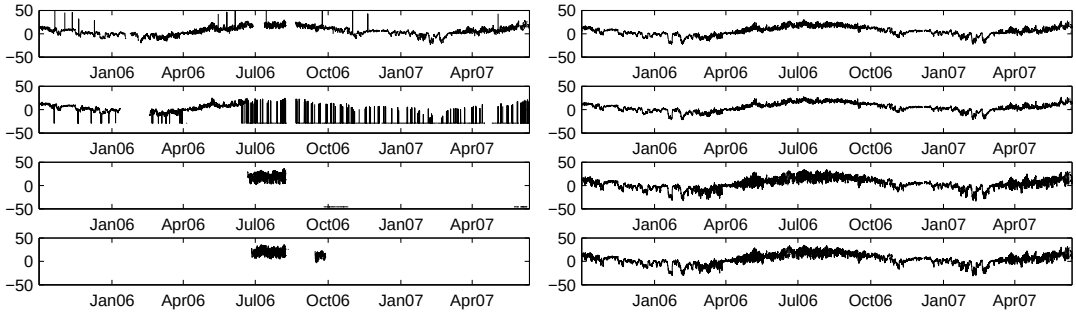


Figure 2.7: Four example signals from the Helsinki Testbed dataset and their reconstructions using the proposed robust PCA.

2.5 Time-series modelling in bioinformatics

Bayesian methods are well-suited for analysis of molecular biology data as the data sets practically always consist of very few samples with a high noise level. We have studied models of gene transcription regulation based on time series gene expression data in collaboration with the Machine Learning and Optimisation group at the University of Manchester. This is a very challenging modelling task as the time series are very short, typically at most a dozen time points.

In [22], we have developed a method of modelling single input motif systems, where a single transcription factor regulates a number of genes. This is achieved by imposing a Gaussian process prior on the latent regulator (transcription factor protein) activity, which under a linear ODE transcription model leads to a joint Gaussian process model for all observable gene expression values. The model can further be extended by incorporating the transcription factor expression levels through a translation model. It is also possible to consider nonlinear models by using approximate inference. A sample model of p53 activation is illustrated in Fig. 2.8.

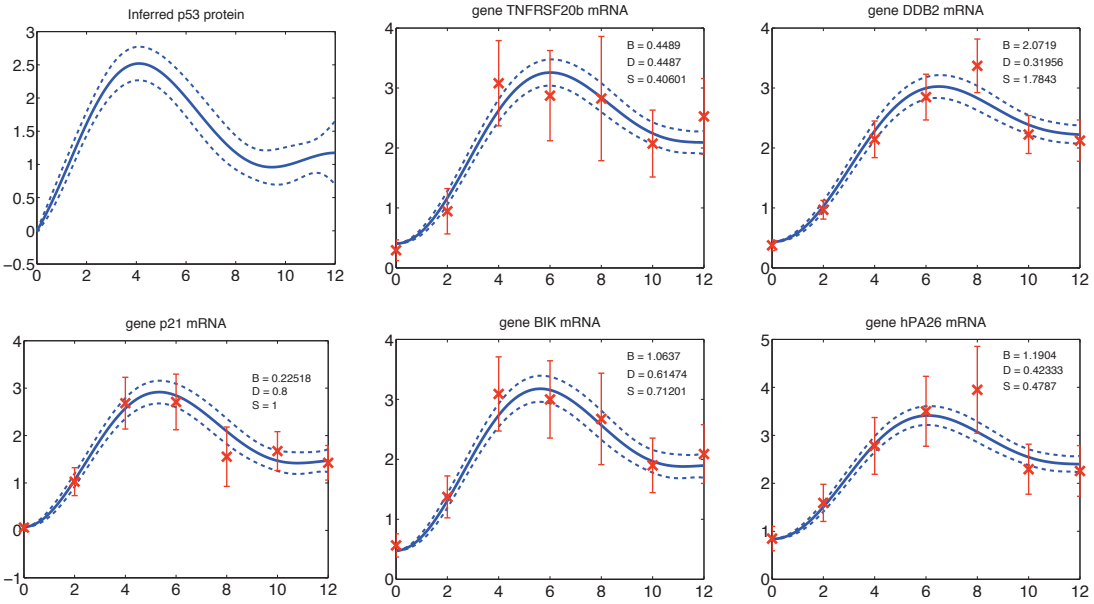


Figure 2.8: An inferred model of transcription factor p53 activation based on five known target genes. Red marks denote observed gene expression values while blue curves are inferred by the model along with 2 standard deviation error bars.

We have applied the model to genome-wide ranking of potential target genes of transcription factors. In experiments with key regulators of *Drosophila* mesoderm and muscle development, this has lead to extremely promising results in terms of enrichment of differential expression in loss-of-function mutants as well as ChIP-chip binding near the predicted target genes [23].

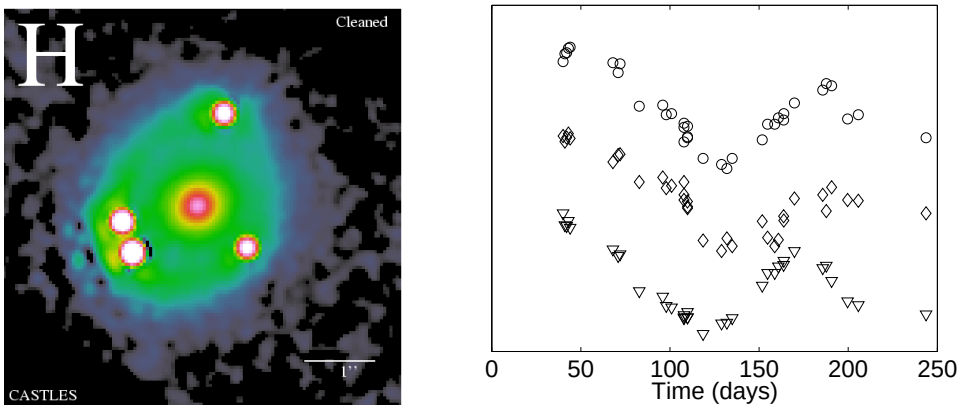


Figure 2.9: Left: The four images of PG1115+080. Right: The corresponding intensity measurements (the two images closest to each other are merged).

2.6 Estimation of time delays in gravitational lensing in astronomy

Most of the research topics contained in Markus Harva's doctoral thesis [24] which appeared in 2008 have already been described in our earlier biennial reports under their chapters on Bayesian learning of latent variable models. However, the journal paper [27] on estimation of time delays in gravitational lensing was published in 2008, and therefore we discuss that work here.

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 2.9 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 2.9). The significance of estimating the delays in such systems stems from the early observation that they can be used in determining important cosmological quantities [25].

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 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 [26].

In [27], 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 against 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 2.10 shows the performance of several methods in an artificial setting.

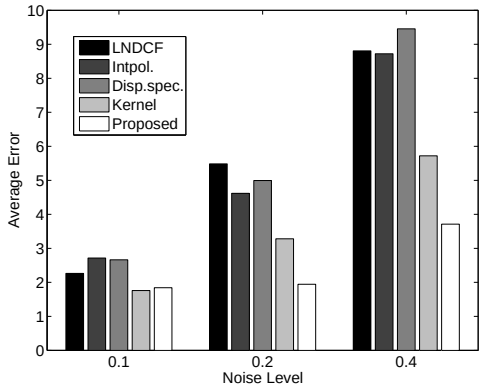


Figure 2.10: Average errors of the methods for three groups of datasets.

2.7 Automated segmentation of brain MR images

Many studies in basis neuroscience and neurological and psychiatric diseases benefit from fully-automated techniques that are able to reliably assign a neuroanatomical label to each voxel in magnetic resonance (MR) images of the brain. In order to cope with the complex anatomy of the human brain, the large overlap in intensity characteristics between structures of interest, and the dependency of MR intensities on the acquisition sequence used, state-of-the-art brain MR labeling techniques rely on prior information extracted from a collection of manually labeled training datasets. Typically, this prior information is represented in the form of *probabilistic atlases*, constructed by first aligning the training datasets together using linear spatial transformations, and then calculating the probability of each voxel being occupied by a particular structure as the relative frequency that structure occurred at that voxel across the training datasets.

While these "average" atlases are intuitive and straightforward to compute, they are not necessarily the best way to extract population-wise statistics from the training data. Atlases built from a limited number of training images tend to generalize poorly to subjects not included in the training database, necessitating heuristic approaches such as spatially blurring atlases used in automated segmentation algorithms. Another problem is that such atlases do not include non-linear deformations aligning corresponding structures across subjects, although this would be a natural way to model anatomical variations.

In [31], we took a critical look at the generative model implicitly underlying probabilistic brain atlases, and proposed to generalize it using tetrahedral mesh-based representations endowed with explicit deformation models. We demonstrated how Bayesian inference can be used to automatically learn the optimal properties of the resulting atlases from a set of manual example segmentations in MR images of training subjects. The learning involves maximizing the probability with which an atlas model would generate the example segmentations, or, equivalently, minimizing the number of bits needed to encode them. This procedure automatically yields sparse atlas representations that explicitly avoid overfitting to the training data, and are therefore better at predicting the neuroanatomy in new subjects than conventional probabilistic atlases [31]. An example of an optimal mesh-based atlas, built from manual annotations of 36 neuroanatomical structures in four individuals, is shown in figure 2.11.

In subsequent work aiming at automatically delineating the subregions of the hippocampus from very high resolution MR images [32, 36, 35], we supplemented the prior distribution provided by a mesh-based atlas, which models the generation of images where each voxel is assigned a unique neuroanatomical label, with a parametric likelihood distribution that predicts how such label images translate into MR images, where each voxel has an intensity. Together these distributions form a complete generative model of MR images that we then used to obtain fully automated structural measurements in a Bayesian fashion, using concepts from our earlier work [28, 29]. In particular, we estimated how the position of the nodes of the atlas mesh are optimally warped onto an image under study, while simultaneously inferring the parameters of the likelihood distribution. Figure 2.12 shows an example of a fully-automated segmentation of the subregions of the hippocampus computed using this approach.

Additional joint work in brain MR analysis we contributed to during the years 2008-2009 include group-wise segmentation of collections of images for which no manual training data is available [38, 41], non-parametric Bayesian whole-brain parcellation [39, 40] and information theoretical image alignment [37], as well as a number of clinical research papers [30, 33, 34].

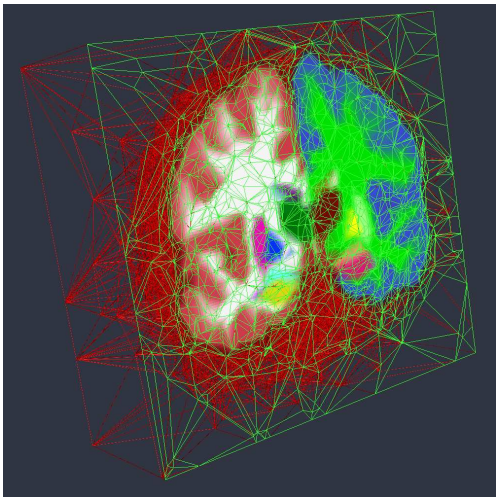


Figure 2.11: Optimal tetrahedral mesh-based atlas built from manual annotations of 36 neuroanatomical structures in 4 subjects. The prior probabilities for the different structures have been color-coded for visualization purposes.

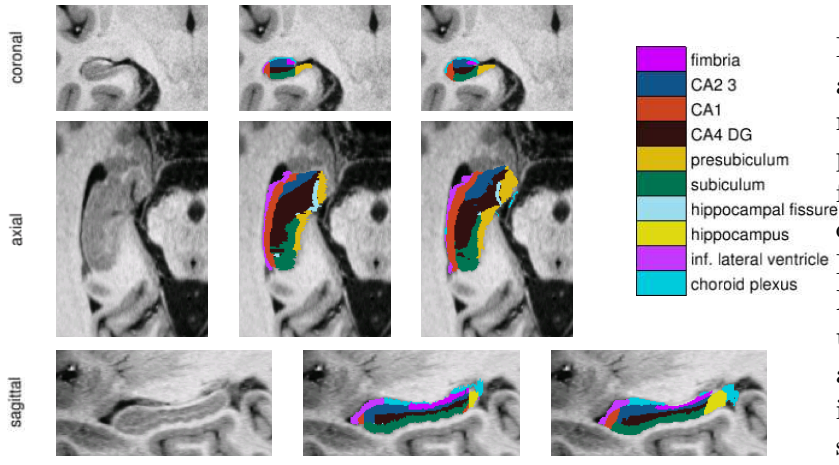


Figure 2.12: Fully automated segmentation of hippocampal subfields from ultra-high resolution MR scans. From left to right: MR data, manual delineations, and corresponding automated segmentations.

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] Y. Qi, T. S. Jaakkola. Parameter expanded variational Bayesian methods. In *Advances in Neural Information Processing Systems 19*, pp. 1097–1104, Cambridge, MA, 2007.

[10] J. Luttinen, A. Ilin, and Tapani Raiko. Transformations for variational factor analysis to speed up learning. In *Proc. of the 14th European Symposium on Artificial Neural Networks (ESANN 2009)*, pp. 77–82, Bruges, Belgium, April 2009.

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

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

[13] 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.

[14] 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.

[15] T. Raiko and M. Tornio. Variational Bayesian learning of nonlinear hidden state-space models for model predictive control. In *Neurocomputing*, volume 72, issues 16–18, pages 3704–3712, October, 2009.

[16] 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, pp. 566–575, 2008.

[17] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. Tech. report TKK-ICS-R6, Helsinki University of Technology, TKK reports in information and computer science, Espoo, Finland, 2008.

[18] L. Kozma, A. Ilin, and Tapani Raiko. Binary principal component analysis in the Netflix collaborative filtering task. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, Grenoble, France, September 2009.

[19] J. Luttinen, A. Ilin, and Juha Karhunen. Bayesian robust PCA for incomplete data. In *Proc. of the 8th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2009)*, pp. 66–73, Paraty, Brazil, March 2009.

[20] J. Zhao, Q. Jiang. Probabilistic PCA for t distributions. *Neurocomputing*, 69:2217–2226, 2006.

[21] C. Archambeau, N. Delannay, M. Verleysen. Robust probabilistic projections. In *Proc. of the 23rd International Conference on Machine Learning (ICML 2006)*, pp. 33-40, New York, NY, USA, 2006.

[22] P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics* 24(16):i70–i75, 2008.

[23] A. Honkela et al. A model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A*, 2010. doi:10.1073/pnas.0914285107

[24] M. Harva. Algorithms for approximate Bayesian inference with applications to astronomical data analysis. *TKK Dissertations in Information and Computer Science*, TKK-ICS-D3, Espoo, Finland, 2008. Available at `http://lib.tkk.fi/Diss/2008/isbn9789512293483/`.

[25] 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.

[26] 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.

[27] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. *Neurocomputing*, 72(1-3):32–38, 2008.

[28] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated Model-based Bias Field Correction of MR Images of the Brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896, 1999

[29] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated Model-based Tissue Classification of MR Images of the Brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, 1999

[30] T. Autti, M. Mannerkoski, J. Hämäläinen, K. Van Leemput, and L. Åberg. JNCL patients show marked brain volume alterations on longitudinal MRI in adolescence. *Journal of Neurology*, 255(8):1226–1230, 2008

[31] K. Van Leemput Encoding Probabilistic Brain Atlases Using Bayesian Inference. *IEEE Transactions on Medical Imaging*, 28(6):822–837, 2009

[32] K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L.L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl. Automated Segmentation of Hippocampal Subfields from Ultra-High Resolution In Vivo MRI. *Hippocampus*, 19(6):549–557, 2009

[33] B. Fischl, A. A. Stevens, N. Rajendran, B. T. T. Yeo, D. N. Greve, K. Van Leemput, J. Polimeni, S. Kakunoori, R. L. Buckner, J. L. Pacheco, D. H. Salat, J. Melcher, M. P. Frosch, B. T. Hyman, P. E. Grant, B. R. Rosen, A. J. W. van der Kouwe, G. C. Wiggins, L. L. Wald, J. C. Augustinack. Predicting the Location of Entorhinal Cortex from MRI. *NeuroImage*, 47(1):8–17, 2009

[34] M. K. Mannerkoski, H. J. Heiskala, K. Van Leemput, L. E. Åberg, R. Raininko, J. Hämäläinen, and T. H. Autti. Children and adolescents with learning and intellectual disabilities and familial need for full-time special-education show regional brain alterations: A voxel-based morphometry study. *Pediatric Research*, 66(3):306-0311, 2009

[35] K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L. L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl. Model-Based Segmentation of Hippocampal Subfields in Ultra-High Resolution In Vivo MRI. *Proceedings of the MICCAI 2008 Workshop on the Computational Anatomy and Physiology of the Hippocampus (CAPH'08)*, pp. 46–55, September 6, 2008, New York, USA

[36] K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L. L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl. Model-Based Segmentation of Hippocampal Subfields in Ultra-High Resolution In Vivo MRI. *Lecture Notes in Computer Science*, 5241:235–243, 2008

[37] M. R. Sabuncu, B. T. T. Yeo, T. Vercauteren, K. Van Leemput, and P. Golland. Asymmetric image-template registration. *Lecture Notes in Computer Science*, 5761:565–573, 2009

[38] T. Riklin Raviv, K. Van Leemput, W. M. Wells, and P. Golland. Joint Segmentation of Image Ensembles via Latent Atlases. *Lecture Notes in Computer Science*, 5761:272–280, 2009

[39] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. Supervised Nonparametric Image Parcellation. *Lecture Notes in Computer Science*, 5762:1075–1083, 2009

[40] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. Nonparametric Mixture Models for Supervised Image Parcellation. *Proceedings of the MICCAI 2009 Workshop on Probabilistic Models for Medical Image Analysis (PMMIA 2009)*, pp. 301-313, September 20, 2009, London, UK

[41] T. Riklin Raviv, B. Menze, K. Van Leemput, B. Stieltjes, M. A. Weber, N. Ayache, W. M. Wells, and P. Golland. Joint Segmentation via Patient-Specific Latent Anatomy Model. *Proceedings of the MICCAI 2009 Workshop on Probabilistic Models for Medical Image Analysis (PMMIA 2009)*, pp. 244-255, September 20, 2009, London, UK