TKK Dissertations in Information and Computer Science
Espoo 2008 TKK-ICS-D3

ALGORITHMS FOR APPROXIMATE BAYESIAN INFERENCE WITH APPLICATIONS TO ASTRONOMICAL DATA ANALYSIS

Markus Harva

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 9th of May, 2008, at 12 o'clock noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O. Box 5400
FI-02015 TKK
FINLAND
Tel. +358-9-451 3272
Fax +358-9-451 3277
E-mail: series@ics.tkk.fi

ISBN 978-951-22-9347-6 (Print)
ISBN 978-951-22-9348-3 (Online)
ISSN 1797-5050 (Print)
ISSN 1797-5069 (Online)
URL: http://lib.tkk.fi/Diss/2008/isbn9789512293483/

Multiprint Oy
Espoo 2008

Harva, M. (2008): **Algorithms for approximate Bayesian inference with applications to astronomical data analysis**. Doctoral thesis, Helsinki University of Technology, Dissertations in Information and Computer Science, TKK-ICS-D3, Espoo, Finland.

Keywords: machine learning, data analysis, Bayesian inference, variational methods, blind source separation, nonnegative factor analysis, heteroscedasticity, predictive uncertainty, delay estimation, elliptical galaxies, gravitational lenses

ABSTRACT

Bayesian inference is a theoretically well-founded and conceptually simple approach to data analysis. The computations in practical problems are anything but simple though, and thus approximations are almost always a necessity. The topic of this thesis is approximate Bayesian inference and its applications in three intertwined problem domains.

Variational Bayesian learning is one type of approximate inference. Its main advantage is its computational efficiency compared to the much applied sampling based methods. Its main disadvantage, on the other hand, is the large amount of analytical work required to derive the necessary components for the algorithm. One part of this thesis reports on an effort to automate variational Bayesian learning of a certain class of models.

The second part of the thesis is concerned with heteroscedastic modelling which is synonymous to variance modelling. Heteroscedastic models are particularly suitable for the Bayesian treatment as many of the traditional estimation methods do not produce satisfactory results for them. In the thesis, variance models and algorithms for estimating them are studied in two different contexts: in source separation and in regression.

Astronomical applications constitute the third part of the thesis. Two problems are posed. One is concerned with the separation of stellar subpopulation spectra from observed galaxy spectra; the other is concerned with estimating the time-delays in gravitational lensing. Solutions to both of these problems are presented, which heavily rely on the machinery of approximate inference.

ABSTRAKTI

Bayesiläinen päättely on teoreettisesti hyvin perusteltu ja käsitteellisesti yksinkertainen lähestymistapa data-analyysiin. Käytännön ongelmien täsmällinen laskennallinen käsittely on kuitenkin usein haastavaa ja siksi approksimaatiot ovat lähes aina tarpeen. Tämän väitöskirjan aihe on approksimatiivinen bayesiläinen päättely ja sen sovellukset kolmessa toisiinsa liittyvässä ongelmakokonaisuudessa.

Variaatio-Bayes on yksi approksimatiivisen päättelyn muoto. Se on laskennallisesti huomattavasti tehokkaampi menetelmä kuin paljon käytetyt otantaan perustuvat menetelmät, mutta vaatii käyttäjältään enemmän analyyttistä työtä tarvittavien päivityskaavojen johdossa. Tämän väitöskirjan ensimmäisessä osassa käsitellään variaatio-Bayes-menetelmän automatisointia tietylle malliluokalle.

Väitöskirjan toisessa osassa tutkitaan heteroskedastisia eli erivarianssisia malleja. Tällaisten mallien käsittely bayesiläisessä viitekehyksessä on erityisen perusteltua, koska monet perinteiset estimointitekniikat eivät tuota niille tyydyttäviä tuloksia. Väitöskirjassa heteroskedastista mallinnusta tarkastellaan kahdesta näkökulmasta: toisaalta lähteen erottelun ja toisaalta regression kannalta.

Astronomiset sovellukset muodostavat väitöskirjan kolmannen osan. Toinen kahdesta tarkasteltavasta ongelmasta käsittelee erilaisten tähtipopulaatioiden erottelua havaituista galaksien spektreistä; toinen ongelma puolestaan koskee gravitaatiolinssien viive-estimointia. Työssä esitetään ratkaisut näihin ongelmiin nojautuen approksimatiivisen päättelyn menetelmiin.

Acknowledgements

This thesis work has been carried out at the Adaptive Informatics Research Centre (AIRC) of Helsinki University of Technology. The main source of funding has been the Helsinki Graduate School in Computer Science and Engineering (Hecse). The Finnish Foundation for Advancement of Technology (TES) has also supported this work with personal grants, which are thankfully acknowledged.

I'm grateful to several people at the AIRC: to Prof. Juha Karhunen for guidance and for the possibility to be part of the Bayes group; to Dr Harri Valpola for his work that sparked the research on approximate inference in the lab; to Dr Antti Honkela, Dr Alexander Ilin, Dr Tapani Raiko, and Jaakko Väyrynen for many stimulating conversations and fun moments in and out the lab; and to very many other colleagues for contributing to the benign, humorous atmosphere that reigns at the research centre.

I also thank my colleagues and coauthors at the University of Birmingham. I'm especially grateful to Drs Ata Kabán and Somak Raychaudhury for the possibility to visit their departments and for the fruitful and enjoyable collaboration with them.

Finally I wish to thank the pre-examiners of this thesis, Drs Simon Rogers and Mark Plumbley, for their valuable feedback, and Prof. Manfred Opper for agreeing to be the opponent in the defence.

Otaniemi, April 2008

Markus Harva

Contents

Publications of the thesis

This thesis consists of a summary part and the following seven publications.

I T. Raiko, H. Valpola, M. Harva, and J. Karhunen. Building blocks for variational Bayesian learning of latent variable models. *Journal of Machine Learning Research*, 8(Jan):155–201, 2007.

II M. Harva, T. Raiko, A. Honkela, H. Valpola, and J. Karhunen. Bayes Blocks: An implementation of the variational Bayesian building blocks framework. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence*, pages 259–266. Edinburgh, Scotland, 2005.

III H. Valpola, M. Harva, and J. Karhunen. Hierarchical models of variance sources. *Signal Processing*, 84(2):267–282, 2004.

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

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

VI 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 ultraviolet-optical spectra. *Monthly Notices of the Royal Astronomical Society*, 366(1):321–338, 2006.

VII M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. *Neurocomputing*, 2007. To appear.

List of abbreviations

ARCH	Autoregressive conditional heteroscedasticity
BSS	Blind source separation
DCF	Discrete correlation function
LNDCF	Locally normalised DCF
EM	Expectation maximisation
EP	Expectation propagation
FA	Factor analysis
ICA	Independent component analysis
iid	Independent identically distributed
KL	Kullback–Leibler (divergence)
MAP	Maximum a posteriori
MCMC	Markov chain Monte Carlo
MDL	Minimum description length
MEG	Magnetoencephalography
ML	Maximum likelihood
MLP	Multilayer perceptron (network)
NLPD	Negative (average) log predictive density
PCA	Principal component analysis
pdf	Probability density function
PP	Predictive perplexity
PSRF	Potential scale reduction factor
RFA	Rectified factor analysis
SNR	Signal to noise ratio
std	Standard deviation
VB	Variational Bayes

List of symbols

$\langle \cdot \rangle$	The expectation operator
$\mathbf{A}, \mathbf{B}, \ldots$	Matrices
$A, B, \ldots$	Propositions
$\mathbf{a}, \mathbf{b}, \ldots$	Vectors
$a, b, \ldots$	Scalars
$\mathcal{C}_{\mathrm{VB}}(q, p)$	The variational Bayesian cost function
$\mathrm{cut}(x)$	The cut (or rectification) function
$\mathcal{D}_{\mathrm{KL}}(q, p)$	The Kullback–Leibler divergence between the two distributions q and p
$\mathrm{diag}(\mathbf{x})$	A diagonal matrix with the elements of vector $\mathbf{x}$ on the main diagonal
$\mathrm{erfc}(x)$	The complementary error function
$\exp(\mathbf{x})$	Exponential function applied component-wise to the vector $\mathbf{x}$
$\mathcal{G}(x\|\alpha, \beta)$	The Gamma distribution with shape α and inverse scale β
M, M_i	The model
$\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	The Gaussian or normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{N}^R(x\|m, v)$	The rectified Gaussian distribution with location parameter m and scale parameter v
$p(x)$	The probability of event x, or the probability density function of x
$q(x)$	Approximate probability density function
$\mathcal{U}(x\|a, b)$	The uniform distribution on the interval $[a, b]$
$u(x)$	The unit step function

Chapter 1

Introduction

1.1 Background

This thesis belongs to the field of machine learning, a broad subfield of computer science, which is concerned with "the study of algorithms that improve automatically through experience" (Mitchell, 1997). Although machine learning can be applied in diverse settings, ranging from robotics to game playing, much of it deals with data analysis. That is the focus in this thesis too.

Many fields of endeavour involve some form of data analysis. Perhaps the main characteristic of the kind of data analysis practised in the context of machine learning is that in that case the algorithms tend to be more data driven and less model dependent. Of course the data do not speak for themselves. There always has to be some set of ideas guiding the analysis, but in machine learning these ideas are usually quite general.

There are several approaches to machine learning, one of which is Bayesian probability theory. It has the appealing property that with certain assumptions it can be shown to be the optimal procedure for logical reasoning under uncertainty. In Bayesian inference, probabilities measure degrees of subjective beliefs, and the theory formalises how these beliefs are to be revised when faced with new information.

The theory of Bayesian inference is concise, and its implementation is straightforward in principle. Unfortunately the exact treatment of any realistic and useful model would require the computation of integrals for which no closed form solutions exist and which are too high dimensional to be tackled with the tools of standard numerical integration. Approximations are thus necessary in almost every real application of Bayesian inference.

The subject of this thesis is approximate Bayesian inference and its applications in three intertwined problem domains.

Variational Bayesian learning is one method for approximate inference. Its main advantage is its computational efficiency; it can be applied to large problems for which sampling-based approaches would be hopelessly inefficient. The price to pay for the reduced computational load is added analytical complexity. Considerable amount of work is required to derive the necessary components of the algorithm for any particular model, which can be a hindrance in the exploration to find the most suitable model for one's problem. One part of this thesis reports on an effort to completely automate variational Bayesian learning of models belonging to a certain, rather general class.

The second part of the thesis is concerned with heteroscedastic modelling. Heteroscedasticity means, shortly put, nonstationarity of variance. This phenomenon is commonplace in many applications, finance being perhaps the most studied one, but its modelling is often neglected as it causes computational difficulties. Or rather, many of the traditional estimation methods simply do not produce satisfactory results for heteroscedastic models—a problem not present with Bayesian computations. In this thesis, heteroscedastic modelling is approached from two different angles. In the source separation setting, it makes sense to look for co-occurring variance fluctuations to aid the modelling of higher order dependencies between the sources which would otherwise go unnoticed. In the regression setting, the uncertainty in the target variable might sometimes be the only thing that can be predicted, the conditional average being mostly meaningless due to high variance of the target.

Astronomical applications constitute the third part of the thesis. The same phenomenon is witnessed in astronomy as in modern society in general: vast amounts of data is, or is becoming, available. In astronomy the process is largely driven by virtual observatory, the Internet archiving of astronomical data. Computational efficiency of the methods used for analysis is in this light a necessity. On the other hand, the extreme opposite of the above applies in certain areas of astronomy, where one can have only a handful of datapoints available for the study of the object of interest. Each datapoint thus becomes extremely valuable both in monetary as well as in information terms, and so it is highly desirable to use methods that squeeze even the last drop of knowledge from that little data one has. In this thesis, astronomical applications with both abundant data and scarce data are encountered, and solutions to them are presented using the tools of approximate Bayesian inference.

1.2 Contents of the publications and author's contributions

Publications I and **II** concern Bayes Blocks, the framework that automates the necessary mathematical derivations for variational Bayesian learning of a flexible class of models. Publication I lays out the theory and the design rationale behind the framework whereas Publication II introduces the accompanying software package that implements it. Bayes Blocks has been a group effort, initiated by Dr Harri Valpola. The present author's contribution consists of having derived and implemented extensions to enable the use of rectification nonlinearities, rectified Gaussian variables, and mixtures-of-Gaussians in the models as well as having implemented some software related features to the library. In Publication I, the main writing responsibility was on Dr Tapani Raiko. The present author participated in the writing and performed some of the experiments. The present author bore the main responsibility of Publication II. The coauthors helped in writing it.

Publications III and **IV** both discuss heteroscedastic modelling, although from different viewpoints. In Publication III, heteroscedastic modelling is studied in the unsupervised learning context. The standard noisy ICA model is extended to several directions by relaxing either the assumption of uncorrelated noise in the observations or the independence of the sources. The paper was written by Dr Harri Valpola. The present author implemented the methods and conducted the experiments. Publication IV shifts the focus to supervised learning. In the paper, nonlinear heteroscedastic regression is studied and an approach based on variational EM is presented.

Publications V, **VI**, and **VII** are related to astronomical applications. Publication V presents a model and a learning algorithm for nonnegative factor analysis termed as rectified factor analysis (RFA). The present author derived and implemented the method, performed all the experiments in that paper, and had the primary responsibility in writing it. Dr Ata Kabán helped in the writing. Publication VI is concerned with the application of RFA to finding stellar subpopulations from a set of observed galaxy spectra. Dr Louisa Nolan has the main writing credits. She also performed the comparison experiments with the astrophysical stellar population model whereas the present author made all the experiments with RFA and helped in writing the paper. Publication VII presents a method for delay estimation in the case when the signals are irregularly sampled. The problem is closely related to delay estimation in gravitational lensing systems which serves as the main motivation for the work, although the method is of general applicability. The present author derived and implemented the method, performed all the experiments, and, for most parts,

wrote the article. Dr Somak Raychaudhury gathered the astronomical data from various sources, and helped in interpreting the results.

Chapter 2

Bayesian probability theory

Probability theory can be seen as an extension of logic applicable when there is uncertainty in the premises. It formalises the process of updating one's beliefs when one observes new data. The Bayesian formulation of probability theory is particularly appealing as it addresses all aspects of statistical inference in a single concise theoretical framework. The basic theory is, indeed, delineated without much effort and that is the subject matter in this chapter. It is the practical implementation of Bayesian inference that is hard, calling for elaborate approximation methods. These are discussed in the next chapter.

2.1 Fundamentals of Bayesian probability

Bayesian probability theory can be derived from many starting points. One of the more intuitive axiomatic systems was formulated by Cox (1946) (see Jaynes (2003) for a detailed discussion). The essence of Cox's axioms is that inference must be rational and consistent. The rules of probability that follow from these principles are the sum rule and the product rule, and from those one can further derive the marginalisation principle and the Bayes's theorem—which is at the core of all Bayesian inference. Other axiomatic systems leading to Bayesian inference include decision theory (Bernardo and Smith, 2000) and the Dutch book arguments.[1]

Before going into the fundamental rules of probability theory, the essential notation is introduced. All probabilities are conditional on some prior knowledge. Given the prior information I, the probability of the propo-

[1] A gambler who is not Bayesian is subject to a Dutch book, i.e. sure to lose money (Lehman, 1955).

sition A is denoted as $p(A|I)$. The probability of the logical conjunction of the two propositions A and B is denoted as $p(A, B|I)$. Sometimes it is convenient to drop out the prior information I to facilitate shorter notation, but even then there is always the underlying assumption that the probabilities are conditional on some prior knowledge. Inference can never take place in vacuum.[2]

2.1.1 The sum rule and the product rule

It follows from Cox's axioms that probabilities are real numbers between zero and one. Zero represents impossibility and one certainty. Let A and B be propositions, let I be the relevant prior information, and let $\neg A$ denote the logical negation of A. Now the sum and the product rule can be represented concisely as

$$\begin{aligned} p(\neg A|I) &= 1 - p(A|I) \\ p(A, B|I) &= p(A|B, I)p(B|I)\,. \end{aligned}$$

The above two equations are all that there is to probability theory at its most fundamental level.

2.1.2 The Bayes's theorem and the marginalisation principle

The Bayes's theorem follows directly from the product rule by writing the probability of the product both possible ways: $p(A, B|I) = p(A|B, I)p(B|I)$ $= p(B|A, I)p(A|I)$. By dividing with $p(B|I)$ we get the Bayes's theorem

$$p(A|B, I) = \frac{p(B|A, I)p(A|I)}{p(B|I)}\,.$$

The marginalisation principle now follows from the sum rule applied to $p(A|B, I)$ which implies

$$\begin{aligned} p(B|I) &= p(B|A, I)p(A|I) + p(B|\neg A, I)p(\neg A|I) \\ &= p(A, B|I) + p(\neg A, B|I)\,. \end{aligned}$$

Above, $p(B|I)$ is called the *marginal* probability of B. Equipped with the Bayes's theorem and the marginalisation principle, we notice that the necessary ingredients for computing the *inverse* probability $p(A|B, I)$ are the *likelihood* $p(B|A, I)$ and the *prior* $p(A|I)$ of A. Often the term *posterior*

[2]For an elaborate discussion of prior assumptions (and lack thereof) in regression, see Wolpert (1996).

probability is used to describe the probability $p(A|B, I)$ as it is the result of combining the prior information I with the additional knowledge B about A.

2.1.3 The continuous case

In the above rules, only propositions are considered, but the extension to discrete variables is immediate. The generalisation to continuous variables is also identical, with the exception that the probabilities are replaced by probability densities and the sums by integrals. The symbol p is somewhat overloaded here as it represents both probabilities and probability densities. In practice, this cannot lead to confusion. For continuous (possibly vector valued) a and b the Bayes's theorem reads

$$p(a|b, I) = \frac{p(b|a, I)p(a|I)}{p(b|I)} = \frac{p(b|a, I)p(a|I)}{\int p(b|a, I)p(a|I)\, da}\,.$$

Of course any combination of discrete and continuous variables can be considered as well. In that case, the expression consists of an appropriate mixture of probabilities and probability densities, and summations and integrals. A rigorous derivation of the rules of Bayesian probability for the continuous case can be found in Bernardo and Smith (2000).

2.2 Uses of Bayesian probability

The above rules are sufficient to answer any question we might ask in the context of probabilistic modelling. But what are the questions most often asked? Below some common scenarios are discussed which cover a large portion of applications of Bayesian inference.

The following notation will be used: $\mathbf{X}$ denotes the data, or the observations made on which the inferences are based, M denotes the model, or the overall assumption about the given problem, and $\boldsymbol{\theta}$ denotes the parameters of the model.

2.2.1 Parameter inference

Assuming that the parameters $\boldsymbol{\theta}$, or a subset of them $\boldsymbol{\theta}_s$, are interesting as such, then we simply want to update our beliefs regarding them based on our model M and the observed data $\mathbf{X}$. This is a matter of an application

of Bayes's theorem:

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}{p(\mathbf{X}|M)}\,. \tag{2.1}$$

If we are only interested in the marginal distribution of $\boldsymbol{\theta}_s$, we can integrate the other parameters out using the marginalisation principle.

The difficulties in making posterior inferences usually begin already at the outset, in computing the posterior, as the normaliser $p(\mathbf{X}|M)$ cannot often be expressed in closed form. The integrals needed for computing the marginal distributions are also commonly intractable.

2.2.2 Predictive inference

In some applications, especially in the field of machine learning, the parameters of the model are *not* interesting as such, but only as a device for making predictions. The multi-layer perceptron is a fine example, where indeed it is difficult to assign meaning to the values of the weights in the network—the model merely serves as a black box for predictions. To compute the posterior predictive density, $p(\mathbf{x}_{\text{new}}|\mathbf{X}, M)$, one computes the posterior of the parameters and then integrates over it:

$$p(\mathbf{x}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{x}_{\text{new}}|\boldsymbol{\theta}, \mathbf{X}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)\,\mathrm{d}\boldsymbol{\theta}\,.$$

Again, the straightforward principle can be difficult to implement in practice due to the common complication of the integral being tricky to evaluate.

2.2.3 Model inference

In most cases, the problem at hand is not so well understood that there would be no doubt about the correctness of the chosen model. There might be various model candidates, $M_1, \ldots, M_n$, from which the most appropriate should be chosen. The posterior probabilities of the competing models can again be computed using the Bayes's theorem:

$$p(M_k|\mathbf{X}, I) = \frac{p(\mathbf{X}|M_k, I)p(M_k|I)}{\sum_{i=1}^{n} p(\mathbf{X}|M_i, I)p(M_i|I)}\,. \tag{2.2}$$

Above, I denotes the higher level assumption that we are confining our inferences to the model family $\{M_i\}_{i=1}^n$. It is important to acknowledge that the model probabilities are conditional on this information. Then it becomes clear that the model posterior tells nothing about the goodness of models outside the chosen family. Conditional on I, a model M_k might have

a probability near unity, but with other assumptions J, perhaps postulating a larger model family, the probability of the very same model could be negligible.

Equation (2.2) unveils the significance of the normaliser $p(\mathbf{X}|M)$. It is an essential ingredient in computing the posterior probability distribution over a set of models. If the prior distribution over the competing models is uniform, the relation between the normaliser and the posterior probability of a given model becomes particularly simple—they are directly proportional to each other in that case. Hence $p(\mathbf{X}|M)$ is sometimes referred to as the *model evidence*.

In making predictions, one might consider averaging not only over the parameters of a particular model, but also over the model family. This is called Bayesian model averaging (Hoeting et al., 1999) and is deemed the correct way of computing predictive distributions. Often in practice, however, one of the models so dominates the posterior distribution, that the averaged predictive distribution is almost equivalent to that of the dominant model.

2.3 On constructing models

Computational complications aside, the above rules are all that is needed to answer our inferential questions, assuming, of course, that we have an appropriate model or model family chosen. Constructing a suitable model is then a problem in its own right. Some of the usual techniques are discussed in this section.

2.3.1 Conjugate priors

A prior distribution is said to be conjugate to a likelihood if the posterior distribution has the same form as the prior (Gelman et al., 1995). To put this more formally, a family of prior distributions $\mathcal{P} = \{p(\theta)\}$ is conjugate to a family of sampling distributions $\mathcal{F} = \{p(y|\theta)\}$ if

$$p(\theta|y) \in \mathcal{P} \text{ for all } p(y|\theta) \in \mathcal{F} \text{ and } p(\theta) \in \mathcal{P}\,.$$

For example, consider the Gaussian distribution parametrised by its mean μ and precision (inverse variance) τ:

$$p(x|\mu,\tau) = \mathcal{N}\left(x|\mu,\tau^{-1}\right)\,.$$

Now the prior distribution conjugate to the likelihood of τ, assuming μ fixed, is the Gamma distribution as then the posterior of τ will also be a Gamma distribution.

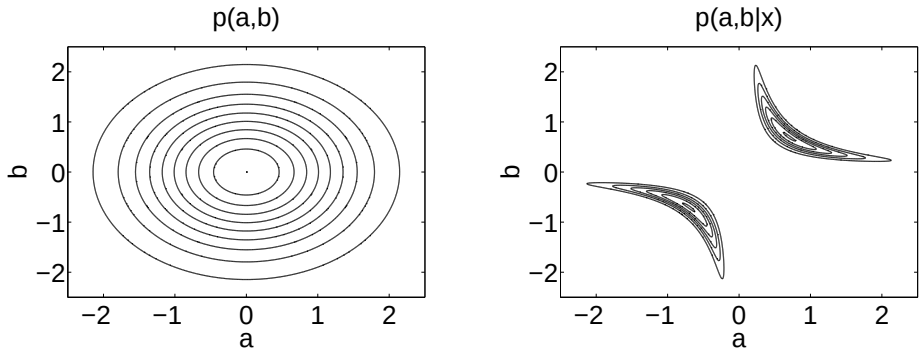


Figure 2.1: The prior and the posterior in the example problem when x is observed to be 0.5.

The ubiquitous use of conjugate priors stems from their convenience in computations because only the parameters of the prior distribution need to be updated in the prior-to-posterior analysis. The models where the prior of the parameter vector as a whole would be conjugate to the likelihood are unfortunately limited, but often it is sufficient that conjugacy applies only conditionally in a multiparameter model. Take the following probabilistic model as an example:

$$\begin{aligned} p(x|a,b) &= \mathcal{N}(x|ab, 0.01) \\ p(a) &= \mathcal{N}(a|0,1) \\ p(b) &= \mathcal{N}(b|0,1)\ . \end{aligned}$$

Above, x is assumed to be observed and it is modelled as the product of two unobserved variables a and b. If we look at the parameter vector as a whole, $\boldsymbol{\theta} = (a, b)$, it is obvious that the prior, which is a bivariate Gaussian with zero mean and identity covariance, is not conjugate to the likelihood as the double-boomerang-shaped posterior (see Figure 2.1) is indeed not a Gaussian. But the posterior for each parameter by itself, assuming the other fixed, is Gaussian, and hence the priors can be said to be conditionally conjugate to the likelihood.

Conditional conjugacy[3] plays an especially important role with some of the approximate methods such as Gibbs's sampling and variational Bayes (discussed in the next chapter).

[3]The term *conditional conjugacy* does not seem to be widely adopted, but at least Gelman (2006) uses it in the exact same meaning as here.

2.3.2 Exponential families

For a probability distribution belonging to an *exponential family*, a conjugate prior always exists (Gelman et al., 1995). That is the foremost reason why the concept of exponential families is of importance and is worth a short review here. A conditional distribution is said to be in an exponential family, if it has the following form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^T u(\mathbf{x}) + f(\mathbf{x}) + g(\boldsymbol{\theta}))\,. \tag{2.3}$$

Above, $\boldsymbol{\theta}$ is the natural parameter, $u(\mathbf{x})$ the sufficient statistic, and $g(\boldsymbol{\theta})$ the normaliser. The choice of the functions u, f, and g define the family, and the choice of the parameters $\boldsymbol{\theta}$ pick one particular distribution from that family. A model where all conditional distributions belong to an exponential family, and where the prior-likelihood relations are (conditionally) conjugate, is called a *conjugate-exponential* model (Ghahramani and Beal, 2001).

Most of the commonly used distributions belong to an exponential family. For example, the Gaussian distribution

$$p(x|\mu,\sigma^2) = \mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2\sigma^2}(x-\mu)^2\right)$$

can be presented in the form of Equation (2.3), if the natural parameter is set to $\boldsymbol{\theta} = (\mu/\sigma^2, 1/\sigma^2)$ and the sufficient statistic to $u(x) = (x, -x^2/2)$.

2.3.3 Latent variable models

Latent variables are formally defined as variables that are not directly observed. Using this definition, all unobserved quantities in a model would be entitled to be called latent. Often it is also assumed that the latent variables somehow break the dependencies between the observed variables, e.g. given the latent variables the observed variables would be independent. This, however, is not always the case.

Introducing latent variables to an otherwise equivalent model often simplifies the model and consequently makes the model estimation easier. Consider, for example, the mixture-of-Gaussians model for an iid sample of one dimensional observations $\mathbf{X} = [x_1, \ldots, x_N]$:

$$p(\mathbf{X}|\mathbf{m},\mathbf{v},\boldsymbol{\pi}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x_i|m_k, v_k\right)\,.$$

Above, the unobserved quantities $\mathbf{m}$, $\mathbf{v}$, and $\boldsymbol{\pi}$ would not usually be called latent variables; they would rather be called parameters of the model as all

the data samples are conditioned on them. An equivalent mixture model can be formulated in terms of (true) latent variables λ_i that indicate from which mixture component the corresponding data samples are generated. This simplifies the conditional model to

$$p(\mathbf{X}|\mathbf{m}, \mathbf{v}, \lambda_i) = \prod_{i=1}^{N} \mathcal{N}\left(x_i|m_{\lambda_i}, v_{\lambda_i}\right) .$$

The latter model is easy to estimate using the EM-algorithm (see the next chapter). It also has the additional benefit of providing the probabilities of the allocations λ_i for each data sample.

2.3.4 Markov blanket

One useful concept, which is not so much related to construction of models as it is to Bayesian modelling in general, is that of Markov blanket (Pearl, 1988). In a probabilistic model, the Markov blanket of a variable consists of its parents, its children, and its children's parents (so called coparents). The terms parent, child, and coparent, in turn, have intuitive meaning when the probabilistic model is viewed as a graph. In this formalism the variables of a model are represented as nodes and the logical dependencies as edges.

Let us make this concrete by an example. Consider a probabilistic model over variables $A, B, \ldots, M$ where the joint probability distribution factors as

$$\begin{aligned} p(A, B, \ldots, M) = {} & p(A|C)p(B|C)p(C|F, G)p(D|G, H)p(E|H)p(F|I) \\ & \times p(G|J, K)p(H)p(I)p(J|L)p(K|M)p(L)p(M) . \end{aligned} \tag{2.4}$$

A graphical representation of this model is shown in Figure 2.2. The node G's parents are J and K, its children are C and D, and its coparents are F and H. The Markov blanket is then the set $\{J, K, C, D, F, H\}$.

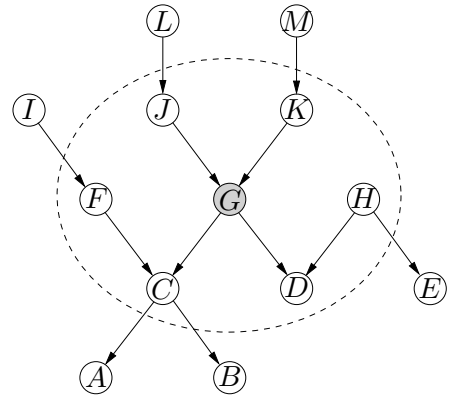


Figure 2.2: The model in Eq. (2.4) represented as a graph. The dashed line marks the set of nodes belonging to the Markov blanket of G.

In predicting a node's state, only the states of the nodes in its Markov blanket are needed, the state of the rest of the model being irrelevant. The importance of this property is later seen in Chapter 4 where a variational Bayesian message passing scheme is discussed.

Chapter 3

Approximate Bayesian inference

As the integrals appearing in Bayesian computations are seldom tractable, approximations are almost always needed. Approximate methods come in many varieties, ranging from simple approaches of reducing the posterior distribution to a point estimate, to complex techniques involving variational calculus or Monte Carlo methods. In this chapter a review of approximate Bayesian inference is given, the emphasis being on those methods that are used in this thesis.

3.1 Deterministic methods

One way to categorise the different approximation schemes is to divide them into deterministic and stochastic methods. As the name implies, a deterministic method always gives the same solution if the initial conditions are kept the same. In this section some of the deterministic methods are reviewed.

3.1.1 Maximum likelihood and maximum a posteriori

The simplest technique to approximate a posterior distribution is to reduce it to a single representative point. The maximum a posteriori (MAP) method does this by finding the parameter values that maximise the posterior density (2.1):

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

The problematic normaliser, $p(\mathbf{X}|M)$, need not be computed as it does not affect the extrema of the probability density function. Also, if the joint density is composed of a product of many simple terms, as is often the case, taking its logarithm will yield an expression involving a sum of those simple terms, making subsequent computations convenient. Once the expression to be maximised is written down, any suitable optimiser can be used to find the optimum.

Nonlinear programming is, of course, plagued with many nontrivial problems, such as nonglobal optima and slow convergence, that as practical issues need to be dealt with when searching for maximum a posteriori estimates. But those aside, there also exist fundamental problems that would not vanish even if there were a perfectly reliable and efficient method for solving the optimisation task. The problem is this: high probability density is an unreliable indicator of where most of the probability mass lies, and the regions of high density but low mass often represent overfitted models. This is especially true in heteroscedastic modelling, as will be demonstrated in some detail in Chapter 5.

Maximum likelihood (ML) estimation differs from maximum a posteriori in that the prior information is ignored and the maximisation is done over the mere likelihood:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta}, M)\,.$$

The same effect is achieved with MAP if one uses a prior which is essentially flat in the region of high likelihood (a prior like this is called *vague*). When the likelihood being under scrutiny has been obtained by marginalising out some of the parameters of a larger model, then maximum likelihood estimation on the marginal likelihood is referred to as type-II maximum likelihood.

Neither MAP nor ML provide means for estimating the model order. By making the model more expressive, one can always obtain a larger likelihood. In practice, the model order selection is often done using cross validation, i.e., by leaving out part of the data while fitting the model and then evaluating the model performance on the left-out data.

3.1.2 Laplace's method

Laplace's method (Tierney and Kadane, 1986) is more an add-on on top of ML or MAP estimation than an estimation technique in its own right. Once the maximum of the posterior (or likelihood) has been found, a second order Taylor expansion is used to approximate the logarithm of the distribution

at that point. This yields a Gaussian that is then used as an approximation to the posterior.

While Laplace's approximation is often easy to compute, its quality largely depends on whether the posterior, or at least one of its significant modes, is close to a Gaussian. And since the first step in the computation of Laplace's approximation is finding the MAP estimate, the problems of MAP are inherited as such. When, however, MAP or ML is sufficient for a problem, Laplace's method provides a simple way to obtain credible intervals for the parameters. The method can also be used to compute the evidence; it can been obtained as the ratio between the unnormalised posterior and the Gaussian approximation.

3.1.3 Variational Bayes

The central idea in variational Bayesian learning (Jordan et al., 1999), or variational Bayes (VB) for short, is to fit a simpler, tractable distribution to the posterior by variational methods. The details of the most common variant of VB, sometimes referred to as ensemble learning (MacKay, 1995; Lappalainen and Miskin, 2000), are described below.

Let the true posterior distribution of the parameters $\boldsymbol{\theta}$ be $p(\boldsymbol{\theta}|\mathbf{X}, M)$. As usual, $\mathbf{X}$ denotes the data and M the modelling assumptions. The VB approximation is the distribution $q(\boldsymbol{\theta})$ from a suitable family $\mathcal{Q}$, that is closest to $p(\boldsymbol{\theta}|\mathbf{X}, M)$ in the sense of the Kullback-Leibler divergence

$$\mathcal{D}_{\mathrm{KL}}(q, p) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{X}, M)} \,\mathrm{d}\boldsymbol{\theta} \,. \tag{3.1}$$

Computing the VB approximation is then a matter of solving the following variational problem:

$$\begin{cases} \text{Minimise } \mathcal{D}_{\mathrm{KL}}(q, p) \text{ w.r.t. } q \\ \text{subject to } q \in \mathcal{Q} \,. \end{cases}$$

The minimisation of $\mathcal{D}_{\mathrm{KL}}$ is equivalent to the minimisation of the cost functional

$$\mathcal{C}_{\mathrm{VB}}(q, p) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\mathbf{X}, \boldsymbol{\theta}|M)} \,\mathrm{d}\boldsymbol{\theta} \tag{3.2}$$

in which the evaluation of the often intractable evidence term $p(\mathbf{X}|M)$ is avoided. The equations (3.1) and (3.2) are connected by

$$\mathcal{C}_{\mathrm{VB}}(q, p) = \mathcal{D}_{\mathrm{KL}}(q, p) - \log p(\mathbf{X}|M) \,. \tag{3.3}$$

From the Gibbs's inequality we know that $\mathcal{D}_{\mathrm{KL}}$ is always nonnegative and so we arrive at the inequality

$$\mathcal{C}_{\mathrm{VB}}(q, p) \geq -\log p(\mathbf{X}|M)\,, \tag{3.4}$$

which shows that the negative of the VB cost bounds the log-evidence of the model from below, the margin separating them being exactly the KL-divergence between the approximate posterior and the true posterior.

Choosing a suitable distribution family $\mathcal{Q}$ is at the core of VB. Indeed, if $\mathcal{Q}$ is set to the class of all distributions, it is easy to see that the optimal "approximation" is the original posterior distribution. But the reason to search for an approximation was, to begin with, that the true posterior is difficult to handle. Hence, $\mathcal{Q}$ should be restricted to distributions that are tractable. Usually one assumes that $\mathcal{Q}$ consists of factorial distributions of the form:

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{N} q(\boldsymbol{\theta}_i)\,.$$

Often the model being studied hints toward a sensible factoring. In variational Bayesian PCA (Bishop, 1999), for example, it is sufficient to split $\boldsymbol{\theta}$ in two parts: one containing the factors and the other the mixing matrix. In some cases the factoring needs to be taken to its extreme and assume a fully factorial q.

The factored posterior also suggests a straightforward way of solving the variational problem. We can update the factors $q(\boldsymbol{\theta}_i)$ one at a time, optimising the relevant part of $\mathcal{C}_{\mathrm{VB}}$ w.r.t. $q(\boldsymbol{\theta}_i)$ while keeping all the other factors fixed. This leads to the so called variational Bayesian EM algorithm (VBEM, Algorithm 1). That name is a bit of a misnomer, however, since there are really no separate, qualitatively different E- and M-steps involved as opposed to the standard EM algorithm or to the variational EM algorithm (discussed in the next section).

Algorithm 1 The variational Bayesian EM algorithm

Initialise $q(\boldsymbol{\theta}_i)$, $\forall i$ to some appropriate distributions
while the change in $\mathcal{C}_{\mathrm{VB}}(q, p) > \epsilon$ **do**
  for $i = 1$ to N **do**
    $q(\boldsymbol{\theta}_i) \leftarrow \arg\min_{q(\boldsymbol{\theta}_i)} \mathcal{C}_{\mathrm{VB}}\big(q(\boldsymbol{\theta}_i) \prod_{k \neq i} q(\boldsymbol{\theta}_k),\ p\big)$
  end for
end while

Why gauge the misfit between q and p with (3.1) and not with some other measure? The reason is, more than anything else, that the particular measure produces tractable algorithms. The KL-divergence the other way around, $\mathcal{D}_{\mathrm{KL}}(p, q)$ that is, would be a natural candidate, as it measures, with

certain assumptions, the expected loss of reporting the probability distribution q instead of the true beliefs encoded in p (Bernardo and Smith, 2000). The shortcoming of $\mathcal{D}_{\mathrm{KL}}(p, q)$ is that it involves integration over p rather than q, which renders its use intractable. Some authors use the term exclusive divergence for $\mathcal{D}_{\mathrm{KL}}(q, p)$ and the term inclusive divergence for $\mathcal{D}_{\mathrm{KL}}(p, q)$ (Winn and Bishop, 2005). The former can produce approximations that exclude parts of the posterior. For example, in a model having symmetries, represented by equivalent modes in the posterior, the exclusive divergence is perfectly content in approximating only one of the modes. The inclusive divergence, on the other hand, tries to capture all of the modes which can result in an approximation covering large portions of the parameter space where the exact posterior has negligible density.

Variational Bayes has its roots in statistical mechanics and especially in the mean field theory (Parisi, 1998; MacKay, 2003) where variational methods are used to approximate the free energy of a physical system. In the machine learning literature, one of VB's earliest appearances has been in the disguise of the minimum description length (MDL) principle (Hinton and van Camp, 1993). There is, indeed, a close connection between VB and MDL. For example, Honkela and Valpola (2004) explain several phenomena in VB learning, such as model pruning and overfitting, by interpreting the modelling problem in the framework of the information theoretic MDL principle.

Compared to the method described in this section, there is a rather different approach to variational Bayesian learning, used for example by Jaakkola and Jordan (1997) and Girolami (2001). There a variational bound is also optimised, but the bound is for the posterior distribution instead of the marginal likelihood.

3.1.4 Variational EM algorithm

The variational EM algorithm (Neal and Hinton, 1999) is very similar to VBEM except that the parameters $\boldsymbol{\theta}$ are divided into two sets, $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$, which are treated asymmetrically. For $\boldsymbol{\psi}$, VEM proceeds like VBEM, revising the approximation $q(\boldsymbol{\psi})$ at each iteration, but for $\boldsymbol{\xi}$ only a point estimate is sought by maximising

$$\int q(\boldsymbol{\psi}) \log p(\mathbf{X}, \boldsymbol{\psi} | \boldsymbol{\xi}, M) \, \mathrm{d}\boldsymbol{\psi} \,. \tag{3.5}$$

This is equivalent to optimising $\mathcal{C}_{\mathrm{VB}}$ w.r.t. $\boldsymbol{\xi}$. Hence the variational EM algorithm can be described as an alternating optimisation of $\mathcal{C}_{\mathrm{VB}}$ w.r.t. $q(\boldsymbol{\psi})$ and $\boldsymbol{\xi}$, which are the E-step and M-step of the algorithm (Algorithm 2).

Computing the E-step involves variational calculus whereas computing the M-step involves ordinary optimisation, usually in a real vector space.

Algorithm 2 The variational EM algorithm

Initialise $q(\boldsymbol{\psi}_i)$, $\forall i$ to some appropriate distributions
Initialise $\boldsymbol{\xi}$ to some appropriate values
while the change in $\mathcal{C}_{\mathrm{VB}}(q, p(\cdot|\boldsymbol{\xi})) > \epsilon$ **do**
  for $i = 1$ to N **do**
    $q(\boldsymbol{\psi}_i) \leftarrow \arg\min_{q(\boldsymbol{\psi}_i)} \mathcal{C}_{\mathrm{VB}}\big(q(\boldsymbol{\psi}_i) \prod_{k \neq i} q(\boldsymbol{\psi}_k),\ p\big)$
  end for
  $\boldsymbol{\xi} \leftarrow \arg\min_{\boldsymbol{\xi}} \mathcal{C}_{\mathrm{VB}}(q, p(\cdot|\boldsymbol{\xi}))$
end while

In contrast to VB, where a lower bound for the model evidence $p(\mathbf{X}|M)$ is obtained, VEM yields a lower bound for the marginal likelihood $p(\mathbf{X}|\boldsymbol{\xi}, M)$ which depends on the values of the parameters $\boldsymbol{\xi}$. To do model comparison then, one needs to account for the additional model complexity due to the parameters $\boldsymbol{\xi}$. This can be done in one of the information criteria frameworks.[1] They, however, have the downside of making quite specific assumptions about the model. Even when these assumptions are not fulfilled, as is often the case, one can still apply the criteria, but then the procedure is no longer as sound as it would be if one were to compare the model evidences.

The ordinary EM algorithm (Dempster et al., 1977) is of course a special case of VEM with no constraints on $q(\boldsymbol{\psi})$ (meaning that $q(\boldsymbol{\psi})$ is equated with $p(\boldsymbol{\psi}|\mathbf{X}, \boldsymbol{\xi}, M)$).

3.1.5 Other deterministic methods

There are a host of other deterministic algorithms for approximate Bayesian inference, the most notable, and the one that is often referred to as an alternative to VB, being expectation propagation (Minka, 2001). In EP, the factors in the probabilistic model are approximated, one at a time, and then used to refine the approximation for the whole posterior distribution. This involves local minimisation of $\mathcal{D}_{\mathrm{KL}}(p, q)$, but is not to be confused with global minimisation of it. When EP converges—it does not always,

[1] For example, the Bayesian information criterion (Schwarz, 1978) is computed as

$$\mathrm{BIC} = -2 \log L + k \log n$$

where L is the maximum of the likelihood, k is the number of parameters, and n is the number of samples. The model with the lowest BIC is to be preferred to the other candidates.

although double-loop algorithms exist that do (see e.g. Heskes and Zoeter, 2002)—the stable point can be shown to correspond to a local minimum of the Bethe free energy (Minka, 2001).

The notable difference to VB is that in EP the inclusive divergence as opposed to the exclusive divergence is used. It is part of the machine learning folk lore that this leads to better modelling of the posterior. Some factual evidence is starting to accumulate as well. For example, Winther and Petersen (2007) study Bayesian ICA and show that the expectation consistent approximation (Opper and Winther, 2005), computed via expectation propagation, is indeed better than VB in modelling the posterior and subsequently leads to more accurate separation of the sources.

3.2 Stochastic methods

As a motivating example, let us consider computing the estimate of a quantity f which depends on some parameters $\boldsymbol{\theta}$. Our knowledge of $\boldsymbol{\theta}$ comes from observed data $\mathbf{X}$ in the form of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$. The Bayes estimate of f is then

$$\langle f(\boldsymbol{\theta}) \rangle = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}) \, \mathrm{d}\boldsymbol{\theta} \,. \tag{3.6}$$

Assuming that we cannot analytically evaluate the integral, but can obtain independent samples, $\{\boldsymbol{\theta}_i\}_{i=1}^N$, from the posterior, we can approximate (3.6) by Monte Carlo integration:

$$\langle f(\boldsymbol{\theta}) \rangle \approx \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\theta}_i) \,. \tag{3.7}$$

Another object of interest to us might be the predictive distribution

$$p(\mathbf{x}_{\mathrm{new}}|\mathbf{X}) = \int p(\mathbf{x}_{\mathrm{new}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}) \, \mathrm{d}\boldsymbol{\theta} \,. \tag{3.8}$$

Again, if we have a sample from the posterior, we can compute an approximation

$$p(\mathbf{x}_{\mathrm{new}}|\mathbf{X}) \approx \frac{1}{N} \sum_{i=1}^{N} p(\mathbf{x}_{\mathrm{new}}|\boldsymbol{\theta}_i) \,. \tag{3.9}$$

A great portion of Bayesian computations is covered by Equations (3.6) and (3.8). Being able to draw samples from the posterior would hence largely solve the computational problems in Bayesian analysis. Unfortunately, sampling the posterior is not trivial.

The most widely used sampling schemes belong to the family of Markov chain Monte Carlo (MCMC) methods. The common factor in those methods is that one constructs a Markov chain that, if and when it converges to its equilibrium distribution, produces samples from the posterior. Given the current state $\boldsymbol{\theta}$, the next state $\boldsymbol{\theta}^*$ in the chain is drawn from a jumping distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, which, depending on the method, is either specified by the user or is implicitly defined by the studied model. The process of drawing the next state must satisfy a condition called detailed balance. This means that the transitions must be such that they preserve the equilibrium distribution. Some of the most common MCMC methods are discussed in what follows.

3.2.1 Metropolis-Hastings

The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), a near synonym to MCMC, is a method of wide applicability, always implementable when the unnormalised posterior distribution can be evaluated pointwise. Complex likelihoods and priors pose no difficulties to Metropolis-Hastings. The jumping distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ in this method is specified by the user and can be almost anything as long as it satisfies certain general properties. At each step in the algorithm, a candidate sample $\boldsymbol{\theta}^*$ is drawn from q conditional on the sample $\boldsymbol{\theta}$ drawn at the previous step. The condition of detailed balance is satisfied by accepting the candidate with probability

$$r = \frac{p(\mathbf{X}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta})} . \tag{3.10}$$

In the case that the candidate is rejected, the sample is replicated (Algorithm 3).

Algorithm 3 The Metropolis-Hastings algorithm

```
Set θ^0 to a random value
for i = 1 to M do
  Draw θ* from q(θ*|θ^{i-1})
  Compute r as in (3.10)
  Draw s from U(0, 1)
  if s < r then
    Set θ^i = θ*
  else
    Set θ^i = θ^{i-1}
  end if
end for
```

Although there are few theoretical limitations to the jumping distribution

q, in practice it needs to be cleverly constructed. If the candidates suggested by a slovenly chosen jumping distribution always get rejected, one obtains M replicates of the initial value. On the other hand, if the jumping distribution is too cautiously set up, such that it perturbs the previous sample only ever so slightly, the candidates get often accepted, but the chain explores little of the parameter space. In both cases, convergence to the equilibrium distribution remains a goal far to be achieved.

3.2.2 Gibbs sampling

In Gibbs sampling (Gelfand and Smith, 1990) the model parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ are updated cyclicly, in a fashion similar to VBEM. When it is the turn of the particular subset of parameters $\boldsymbol{\theta}_j$ to be updated, the conditional distribution $p(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{\setminus j}, \mathbf{X})$ is computed and then used as the jumping distribution (Algorithm 4). The advantage of Gibbs sampling is that there are no parameters to be tuned. One need not spend time in search of a jumping distribution that would make the sampler converge within a reasonable time, as is the case with Metropolis-Hastings. The disadvantage, however, is the requirement that one must be able to sample from the conditional distributions. This is convenient only for certain model families, rendering Gibbs sampling applicable to a limited class of problems.

Algorithm 4 The Gibbs sampler

Set $\boldsymbol{\theta}^0$ to a random value
for $i = 1$ to M **do**
  Draw $\boldsymbol{\theta}_1^i$ from $p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{i-1}, \boldsymbol{\theta}_3^{i-1}, \ldots, \boldsymbol{\theta}_N^{i-1}, \mathbf{X})$
  Draw $\boldsymbol{\theta}_2^i$ from $p(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^i, \boldsymbol{\theta}_3^{i-1}, \ldots, \boldsymbol{\theta}_N^{i-1}, \mathbf{X})$
  $\vdots$
  Draw $\boldsymbol{\theta}_N^i$ from $p(\boldsymbol{\theta}_N | \boldsymbol{\theta}_1^i, \boldsymbol{\theta}_2^i, \ldots, \boldsymbol{\theta}_{N-1}^i, \mathbf{X})$
end for

3.2.3 Advanced sampling methods

More advanced sampling methods target various weaknesses in the above standard sampling algorithms. To give a few examples: slice sampling (Neal, 2003) alleviates the problems of choosing a good jumping distribution in Metropolis-Hastings; ordered overrelaxation (Neal, 1995) reduces the random walk behaviour in Gibbs sampling; parallel tempering (see e.g. Gregory, 2005) helps in both Metropolis-Hastings and Gibbs sampling when the posterior distribution is multimodal.

A particular situation when one is forced to resort to an advanced sampling method is when the model evidence needs to be computed. Neither Metropolis-Hastings nor Gibbs sampling lend themselves to this task. There are, however, many other methods that do yield the model evidence either as the primary or as the side product of their operation, including thermodynamic integration (Gregory, 2005), annealed importance sampling (Neal, 2001), and nested sampling (Skilling, 2006).

Thermodynamic integration, a method adapted from statistical physics, is one of the better known approaches for the computation of the evidence. It is based on running parallel Markov chains in several "temperatures" using any suitable sampling method. As the temperature varies from hot to cold, the posterior accordingly transforms from the prior to the true posterior. From these parallel runs, the evidence is obtained by integrating over the temperature scale. The computational complexity of thermodynamic integration is at least an order of magnitude higher than in the standard sampling methods, because of the need for running several parallel chains. It can also be difficult to adjust the jumping distributions so that the sampling is efficient at each temperature. Similar complications are often present in the other sampling schemes for the computation of the evidence.

3.2.4 Convergence issues

MCMC methods have one distinct drawback: it is difficult to know if and when the Markov chain has converged to its equilibrium distribution. If it has not, the samples do not come from the posterior distribution and the subsequent analysis is thus rendered unreliable if not meaningless. Several authors have proposed methods for evaluating the convergence. Perhaps the most popular approach is the one by Brooks and Gelman (1998). They suggest several different statistics which they collectively call potential scale reduction factors (PSRF). All the variants of PSRF are based on running parallel Markov chains starting from different initial conditions and then comparing the within sequence statistics to the total sequence statistics. Let us denote the n samples from m parallel chains as θ_{jt} ($j = 1, \ldots, m$, $t = 1, \ldots, n$). Then the particular PSRF that is based on the s:th moment is defined as

$$\hat{R}_s = \frac{\frac{1}{mn-1} \sum_{j=1}^{m} \sum_{t=1}^{n} |\theta_{jt} - \bar{\theta}_{\cdot\cdot}|^s}{\frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{t=1}^{n} |\theta_{jt} - \bar{\theta}_{j\cdot}|^s}, \tag{3.11}$$

where $\bar{\theta}_{j\cdot} = \frac{1}{n} \sum_{t=1}^{n} \theta_{jt}$ and $\bar{\theta}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}_{j\cdot}$. The original PSRF, introduced by Gelman and Rubin (1992), roughly corresponds to $\hat{R}_2$. As the usual summaries computed from the posterior are the mean and the vari-

ance, it is reasonable to ensure that one obtains similar values for those over a number of independent simulations.

3.3 A hierarchy of approximations

All of the methods reviewed in this chapter have their particular strengths and weaknesses. One way to organise them is in terms of their accuracy versus their computational complexity. Such an attempt has been made in Figure 3.1.

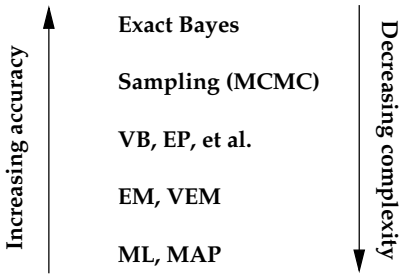


Figure 3.1: A hierarchy of approximations

The exact Bayesian treatment is the most accurate by its very definition. It is also the hardest to implement, requiring the evaluation of integrals over high dimensional spaces. If these integrals were tractable, there would be no need to discuss approximations.

In the hierarchy, sampling-based methods come next. Actually, they are not always considered approximate methods at all, as there is the guarantee that with persistent enough sampling, the samples will eventually come from the true posterior. But in the real world we are confined to a finite sample, often even to a rather small one, which makes it questionable to consider sampling exact.

Although otherwise a preferable approach, sampling is computationally intensive. The deterministic approximations, variational Bayes, expectation propagation and the related methods, are usually several orders of magnitude faster and yet "accurate" enough. The price to pay for the speed-up is added analytical complexity. Whereas MCMC methods can often be implemented by simply writing down the probabilistic model, the deterministic approximations require considerable pen-and-paper work in the derivation of the update rules. If the studied problem allows it, one can take one step further down in the hierarchy and neglect some of the posterior modelling by replacing part of the distributional estimates with point estimates. This can simplify the problem drastically and still avoid the many difficulties as-

sociated with the bottom of the hierarchy populated by the point estimation methods.

Above, the word accurate is inside quotes with good reasons. At first thought it would seem that the closer the approximation is to the true posterior the better it should be deemed. The matter, however, is not that simple at all. What is useful and what is not largely depends on the kind of analysis one wishes to conduct. Consider, for example, the model for noisy independent component analysis (for the definition of the model, see e.g. Section 5.2.1 in this thesis). It is well known and intuitively clear that any solution can be turned into another equally good solution by permuting the sources or changing their signs. Accordingly the posterior distribution has a plethora of modes that essentially represent the same solution. Computing a Bayes estimate of the sources by averaging over the posterior, then yields nothing sensible. In this light, an approximation that captures only one of the prominent modes, such as those one often finds with VB, might sometimes be more desirable than even the exact posterior.

Chapter 4

Framework for variational Bayesian learning

Variational Bayesian learning has proven to be a powerful approximate method for Bayesian inference. One drawback is its analytical complexity, meaning that substantial pen-and-paper work is required to derive the cost function and the update rules. This chapter focuses on the Bayes Blocks framework for variational Bayesian learning. The framework automates the necessary mathematical derivations leaving to the user only the burden of specifying his model. The model family is constrained, but not to the conjugate-exponential family for which a general variational inference procedure has been shown to be tractable by other authors (Winn and Bishop, 2005).

4.1 Bayes Blocks

Finding a suitable model for a problem is most often an iterative process. We start with some initial guess for a model and, based on experimenting, adjust the model incrementally to better describe the data. Bayesian inference can be a good guide in this endeavour as it provides sound quantitative advice on whether our model is getting better or worse in explaining the data. But it can be a hindrance as well, since the approximation methods (that are practically always needed) are laborious to derive and implement.

Bayes Blocks, the inference framework discussed in Publications I and II, is intended to alleviate the problem mentioned above by automating variational Bayesian inference for a certain class of models. Prototyping of various model candidates is thus made fast and effortless. The attempt of

this chapter is not to give a detailed summary of the related publications of the thesis, but rather to shed light on the operation of the framework with a few illustrative examples.

4.1.1 The building blocks

Bayes Blocks provides a set of blocks which can be combined, according to certain rules, to construct a wide variety of probabilistic models. The blocks can be divided into two categories: variable nodes and computational nodes. The variable nodes correspond to the observed and unobserved quantities in the model and the computational nodes provide means of combining these quantities in nontrivial ways. The nodes in both categories are listed in Table 4.1 which also describes the allowed connectivity. The Type column enumerates the node classes that can serve as the particular parent for the node. The node classes carry no deeper meaning; they simply identify the nodes that can appear at similar configurations in the model. Examples of how to read the table: the mean parent of a Gaussian variable can be anything with the exception of multinomial and Dirichlet variables whereas the parent of a nonlinearity can only be a Gaussian variable.

Class	Node	Parent	Type
	Variable nodes		
1	Gaussian	Mean	1, 2, 3
		Variance	1, 2
3	Rectified Gaussian	Scale	1, 2
3	Mixture of Gaussians	Mean	1, 2, 3
		Variance	1, 2
		Selector	4
4	Multinomial	Probability	5
5	Dirichlet	n/a	
	Computational nodes		
2	Sum		1, 2, 3
3	Product		1, 2, 3
3	Nonlinearity		1

Table 4.1: The blocks and the allowed connectivity. The Type column refers to the Class column, enumerating the Classes that can serve as the particular parent for the node.

The rules of connectivity follow from what kind of expectations can be computed in the forward direction in the network, and from what kind of potentials can be propagated backwards to the parent nodes. This is explained in detail in Publication I.

The usual convention for Gaussian variables is to use an inverse parametrisation for the variance so that the conditional distribution is of the form $p(x|\mu,\tau) = \mathcal{N}\left(x|\mu,\tau^{-1}\right)$. This ubiquitous practice is followed because then a Gamma prior for τ is conditionally conjugate to the likelihood. In Bayes Blocks this convention is abandoned and the variance is parametrised on the log-scale leading to a conditional distribution of the form $p(x|\mu,u) = \mathcal{N}\left(x|\mu,e^{-u}\right)$. There are a number of consequences. On the negative side, the requirements for conditional conjugacy are not satisfied and so the update rules are more complicated. On the positive side, the log-parametrisation opens many possibilities for modelling the variance that are not in one's reach if one is restricted to conditional conjugacy. The first part of Chapter 5 discusses several models that exploit this property.

4.1.2 Example: nonstationary ICA

Converting the mathematical description of a model to its Bayes Blocks implementation is a straightforward process. We will demonstrate this with a block implementation of nonstationary ICA. The observations $x_i(t)$ are the outcome of linearly mixing a number of independent sources $s_j(t)$. The nonstationarity is modelled by putting an AR(1) process prior on the log-variances of the sources. The probabilistic model is then

$$\begin{aligned} x_i(t) &\sim \mathcal{N}\Big(\sum_{j=1}^{M} a_{ij} s_j(t), e^{-v_i}\Big) \\ s_j(t) &\sim \mathcal{N}\left(0, e^{-u_j(t)}\right) \\ u_j(t) &\sim \mathcal{N}\left(u_j(t-1), e^{-w_j}\right) \\ a_{ij} &\sim \mathcal{N}\left(0,1\right) . \end{aligned}$$

The parameters v_i and w_j, controlling the variance of the observation noise and the variance of the innovation process of $u_j(t)$, would in reality have priors as well, but are here assumed constant to not complicate the example.

In the Bayes Blocks formalism a model is expressed in terms of the variable and computational nodes. One possible representation of the above model

is

$$
\begin{aligned}
z^{-1}u_j(t) &= u_j(t-1) \\
u_j(t) &\sim \mathcal{N}\left(z^{-1}u_j(t), e^{-w_j}\right) \\
s_j(t) &\sim \mathcal{N}\left(0, e^{-u_j(t)}\right) \\
a_{ij} &\sim \mathcal{N}\left(0, e^{-0}\right) \\
\mathrm{prod}_{ij}(t) &= a_{ij} \cdot s_j(t) \\
\mathrm{sum}_i(t) &= \sum_j \mathrm{prod}_{ij}(t) \\
x_i(t) &\sim \mathcal{N}\left(\mathrm{sum}_i(t), e^{-v_i}\right) .
\end{aligned}
$$

The above representation mostly corresponds to the original model specification with the exception that the delay, summation, and product are explicitly shown as entities in their own right. This more elaborate description, in turn, maps almost directly to the Python implementation shown in Listing 1.

4.1.3 The message passing scheme

The inference algorithm in Bayes Blocks is based on message passing. Each node sends and receives messages to and from its immediate neighbours, the objective being to find an approximate distribution that minimises $\mathcal{C}_{\mathrm{VB}}$ for the particular model. Since the algorithm is equivalent to VBEM, convergence to a stable point of $\mathcal{C}_{\mathrm{VB}}$ is guaranteed.

The following example illustrates the computations taking place in Bayes Blocks. We will go through the calculations for one variable in a simple submodel of a larger hierarchical model. The submodel consists of five unobserved variables with the dependency structure

$$
\begin{aligned}
p(x|s,v) &= \mathcal{N}\left(x|s, e^{-v}\right) \\
p(s|m,w) &= \mathcal{N}\left(s|m, e^{-w}\right) .
\end{aligned}
$$

The posterior approximation in Bayes Blocks is fully factorial so the approximation for the variables in the submodel factors as

$$
q(x,v,s,m,w) = q(x)q(v)q(s)q(m)q(w) .
$$

The variable of interest here is s. We will find the optimal $q(s)$ given the other approximations.

Listing 1 Bayes Blocks implementation of nonstationary ICA

```
net = PyNet(tdim)                                  # Create the net
f = PyNodeFactory(net)                             # Create a nodefactory

c0 = f.GetConstant("const+0", 0.0)                 # Create some constants
v, w = ...

zu = [f.GetDelayV(Label("zu", j), c0,              # z^{-1}u_j(t) = u_j(t-1)
        f.GetProxy(Label("pu", j),
                   Label("u", j)))
      for j in range(sdim)]

u = [f.GetGaussianV(Label("u", j),                 # u_j(t) ~ N(z^{-1}u_j(t), e^{-w_j})
                    zu[j], w[j])
     for j in range(sdim)]

s = [f.GetGaussianV(Label("s", j), c0, u[j])       # s_j(t) ~ N(0, e^{-u_j(t)})
     for j in range(sdim)]

a = [[f.GetGaussian(Label("a", i, j), c0, c0)      # a_ij ~ N(0, e^{-0})
      for j in range(sdim)]
     for i in range(xdim)]

prods = [[f.GetProdV(Label("prod", i, j),          # prod_ij(t) = a_ij * s_j(t)
                     a[i][j], s[j])
          for j in range(sdim)]
         for i in range(xdim)]

sums = []                                          # sum_i(t) = sum_j prod_ij(t)
for i in range(xdim):
    sums.append(f.GetSumNV(Label("sum", i)))
    for j in range(sdim):
        sums[i].AddParent(prods[i][j])

x = [f.GetGaussianV(Label("x", i),                 # x_i(t) ~ N(sum_i(t), e^{-v_i})
                    sums[i], v[i])
     for i in range(xdim)]
```

The part of the cost function affected by $q(s)$ is

$$\begin{aligned}\mathcal{C}_{\mathrm{VB}} &= \left\langle \log \frac{q(s)}{p(x|s,v)p(s|m,w)} \right\rangle_{q(x,v,s,m,w)} \\ &= \left\langle \log q(s) - \langle \log p(x|s,v) \rangle_{q(x,v)} - \langle \log p(s|m,w) \rangle_{q(m,w)} \right\rangle_{q(s)} . \quad (4.1)\end{aligned}$$

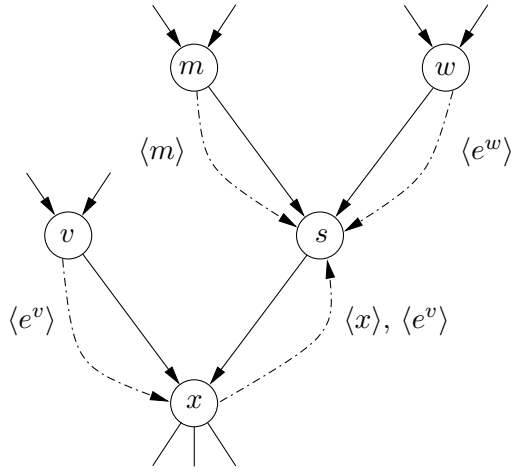


Figure 4.1: Updating $q(s)$. The solid lines show the logical dependencies and the dashed lines show the propagation of the expected values needed to update $q(s)$.

The expectation of $\log p(s|m,w)$ yields

$$\begin{aligned}\langle \log p(s|m,w)\rangle &= \left\langle \log \mathcal{N}\left(s|m,e^{-w}\right)\right\rangle \\ &= \log \mathcal{N}\left(s|\,\langle m\rangle\,,\langle e^w\rangle^{-1}\right) + \text{const.}\end{aligned} \tag{4.2}$$

The latter equality follows by expanding the quadratic form in the normal distribution and using the linearity of the expectation operation (remember that m and w are independent under q). Similarly

$$\langle \log p(x|s,v)\rangle = \log \mathcal{N}\left(\langle x\rangle\,|s,\langle e^v\rangle^{-1}\right) + \text{const.} \tag{4.3}$$

Substituting (4.2) and (4.3) back to (4.1) we get

$$\mathcal{C}_{\text{VB}} = \left\langle \log \frac{q(s)}{\mathcal{N}\left(\langle x\rangle\,|s,\langle e^v\rangle^{-1}\right)\mathcal{N}\left(s|\,\langle m\rangle\,,\langle e^w\rangle^{-1}\right)}\right\rangle_{q(s)} + \text{const.}$$

The two normal distributions combine to form a single (unnormalised) normal distribution $\mathcal{N}(s|\bar{s},\tilde{s})$ with parameters

$$\begin{aligned}\tilde{s} &= (\langle e^v\rangle + \langle e^w\rangle)^{-1} \qquad \text{and}\\ \bar{s} &= \tilde{s}(\langle e^v\rangle\,\langle x\rangle + \langle e^w\rangle\,\langle m\rangle).\end{aligned}$$

The variational problem is then solved by an invocation of Gibbs's inequality from which it follows that the optimal approximation $q(s) = \mathcal{N}(s|\bar{s},\tilde{s})$.

Figure 4.1 shows the flow of information in the model. We note that the update of $q(s)$ can be done by propagating certain expected values from

the nodes in s's Markov blanket. Of course the node must also know the types of the potentials these expected values are encoding. For example, a Gaussian potential coming through a rectification nonlinearity is obviously no longer Gaussian. But the node knows its neighbours and so can infer the types of the potentials to update its posterior appropriately.

Only Gaussian variables were considered in the example, but the same principles apply to the rest of the nodes in any allowed configuration. In the inference algorithm, information from only the immediate neighbours of the node are needed to update its state. From this it follows that the computational complexity of one update iteration is linear w.r.t. the number of connections in the model. The price to pay for the efficiency is that a fully factorial posterior approximation has to be used. This means that all posterior dependencies are neglected, which can sometimes cause unfortunate side effects in the model estimation (see e.g. Ilin and Valpola, 2005).

4.2 Other frameworks

Tools for automating Bayesian inference have been considered by many other authors as well. One such popular framework is BUGS (Spiegelhalter et al., 1995). The acronym stands for Bayesian inference Using Gibbs Sampling and, as the name suggests, the framework is intended to be a flexible tool for Bayesian analysis using Markov chain Monte Carlo methods. The use of MCMC makes BUGS a widely applicable piece of software for Bayesian inference, but also limits the size of the models that can be studied as MCMC methods involve intense computations.

More relevant to the discussion of Bayes Blocks is the framework by Winn and Bishop (2005) called VIBES (Variational Inference for BayEsian networkS). Similarly to Bayes Blocks, its inference algorithm is based on variational Bayesian learning. The supported model family is different, though, as their framework is confined to models in the conjugate-exponential family. On the one hand, VIBES is a more general framework than Bayes Blocks in that any distribution from the conjugate-exponential family can relatively easily be incorporated to the system. On the other hand, constructing nonlinear and variance models in the way that it is possible in Bayes Blocks cannot be done in VIBES as these kind of models do not meet the criteria of conditional conjugacy.

Chapter 5

Heteroscedastic modelling

Heteroscedasticity means nonstationarity of variance. For computational convenience, the opposite assumption, i.e. of homoscedasticity, is made in most standard probabilistic models. In this chapter, heteroscedastic modelling is discussed in two contexts. In the first part of the chapter, the noisy-ICA model is extended to several directions to include the modelling of the nonstationary variance. In the second part, heteroscedasticity in nonlinear regression is discussed where it leads to predictive uncertainty, i.e., to models that can predict not only the mean outcome of but also the uncertainty in the phenomenon of interest.

5.1 The trouble with heteroscedastic modelling

Even though heteroscedasticity is commonplace in many applications, its modelling is often neglected to avoid facing the associated computational complications. This section gives a simple but representative example of these troubles.

The problem we try to solve—first with ML and MAP, and later with variational methods—is estimating the mean and variance of a normal distribution from one observation. The model is

$$p(x|m,u) = \mathcal{N}\left(x|m, e^{-u}\right) \tag{5.1}$$

$$p(m) = \mathcal{N}\left(m|0, \tau_m^{-1}\right) \tag{5.2}$$

$$p(u) = \mathcal{N}\left(u|0, \tau_u^{-1}\right) . \tag{5.3}$$

Above, x is the observation and m and u are the parameters of the normal distribution who have Gaussian priors. In what follows the constants are set to $\tau_m = 1$ and $\tau_u = 1/25$. It is meaningful to study the above simple

model, because such a construction will frequently appear in the realistic models encountered later on in this chapter.

Before proceeding it is worth noting that estimating the variance from a single observation is only possible if there is some information about the mean. If we had an uninformative prior for m, that is $p(m) \propto 1$, the posterior of u would equal its prior. This is what common sense suggests. In the absence of a reference point, nothing can be said about the variability of a distribution by looking a single sample drawn from it.

It is then not surprising that attempts towards obtaining a maximum likelihood estimate, which implies that we ignore all the prior information, will yield no estimate whatsoever for the problem. This becomes immediate when the likelihood is written out

$$\begin{aligned} p(x|m,u) &= \frac{1}{\sqrt{2\pi e^{-u}}} \exp\left\{-\frac{1}{2e^{-u}}(x-m)^2\right\} \\ &\propto \exp\left\{-\tfrac{1}{2}(x-m)^2 e^u + \tfrac{1}{2}u\right\}. \end{aligned}$$

When we set $m := x$ this simplifies to $\exp(\frac{1}{2}u)$ which approaches infinity as we let $u \to \infty$. Since the likelihood is not bounded from above, there cannot be a ML estimate.

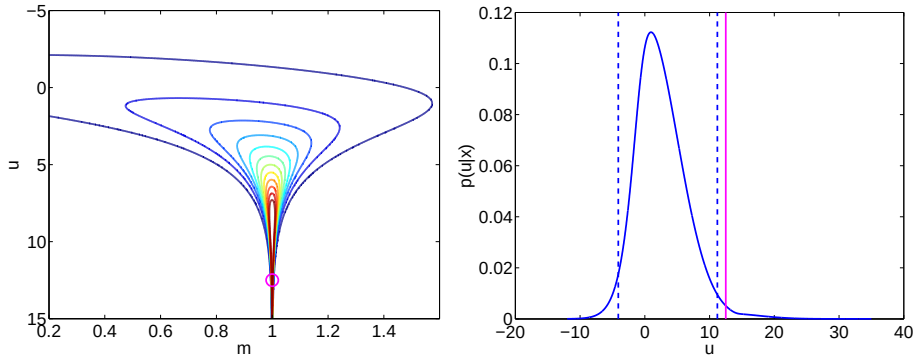
The situation is better with maximum a posteriori in that the MAP estimate does at least exist. The joint unnormalised posterior of m and u is readily obtained by an application of Bayes theorem

$$\begin{aligned} p(m,u|x) &\propto p(x|m,u)p(m)p(u) \\ &= \mathcal{N}\left(x|m,e^{-u}\right)\mathcal{N}\left(m|0,\tau_m^{-1}\right)\mathcal{N}\left(u|0,\tau_u^{-1}\right) \\ &\propto \exp\left\{-\tfrac{1}{2}(x-m)^2 e^u + \tfrac{1}{2}u - \tfrac{1}{2}\tau_m m^2 - \tfrac{1}{2}\tau_u u^2\right\}. \end{aligned} \tag{5.4}$$

Assuming that we have observed $x = 1$, the posterior has the shape shown in Figure 5.1(a). We will find the MAP estimate by visual inspection. It looks as though the optimal m equals one. By substituting $m \leftarrow 1$ in Eq. (5.4), we are left with the expression $\exp\{\frac{1}{2}u - \frac{1}{2}\tau_u u^2\}$, which is optimised with $u = 1/2\tau_u = 25/2$.[1] Intuitively this appears to be an extreme estimate given that there was considerable uncertainty in m. That this estimate is poor in representing the posterior probability mass is obvious in Figure 5.1(b) where the marginal posterior distribution $p(u|x)$ is shown.

So far we have not obtained a reasonable solution to our estimation problem, apart from the exact Bayesian treatment of course. Point estimates are too simple an approximation to the problem—they go awry in that

[1] Numerical analysis yields an optimum that equals this less rigorous estimate to several decimal places.



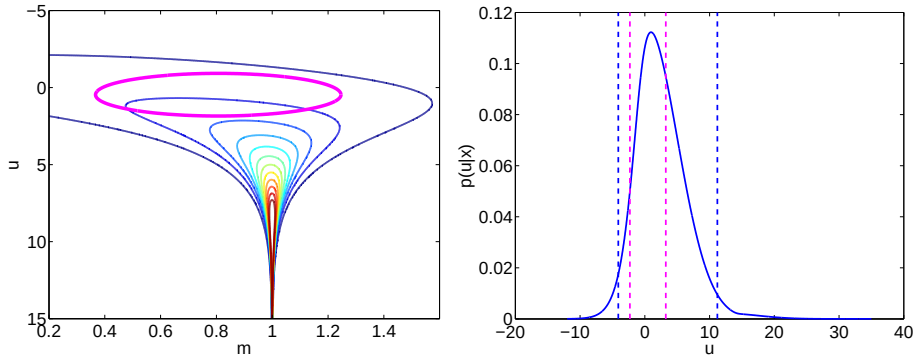
(a) (b)

Figure 5.1: MAP in the variance problem. (a) The joint posterior probability distribution $p(m, u|x)$. The magenta circle marks the MAP estimate. (b) The marginal distribution $p(u|x)$. The dashed lines show the 95% credible interval, and the magenta line shows the MAP estimate.

they attempt to find the region of high probability density when it is the probability mass that matters.

The whole point of this section is of course to demonstrate that approximation techniques that seek to summarise the true distribution more wholeheartedly than point estimates do are better suited for variance modelling. So as the final attempt, we shall find a VB approximation for our problem. We fix the approximation to be a product of two Gaussians $q(m, u) = \mathcal{N}\left(m|\mu_m, \sigma_m^2\right)\mathcal{N}\left(u|\mu_u, \sigma_u^2\right)$. Finding the optimal approximation is then a matter of minimising $\mathcal{C}_{\text{VB}}$ w.r.t. the variational parameters μ_m, σ_m^2, μ_u, and σ_u^2. Skipping the details of the calculations, it suffices to state that the global optimum is at $\mu_m = 0.806$, $\sigma_m^2 = 0.194$, $\mu_u = 0.468$, $\sigma_u^2 = 1.92$.

Figure 5.2(a) shows the true and the approximate posterior. The approximate distribution lies far from the region of high probability density, which indicates that the bulk of the probability mass is spread across the low density region. In Figure 5.2(b) the approximation $q(u)$ is contrasted to the marginal distribution $p(u|x)$. The exclusiveness property of the VB approximation, discussed in Section 3.1.3, is here clearly visible. The 95% credible interval of $q(u)$ is less than half of that of $p(u|x)$. Although the variances of the VB approximation do not reflect the true variability of the posterior, the approximation is nevertheless much more sensible than the earlier estimates we obtained.



(a) (b)

Figure 5.2: VB in the variance problem. (a) The joint posterior probability distribution $p(m, u|x)$. The magenta ellipse marks the 1-std contour of the VB approximation. (b) The marginal distribution $p(u|x)$. The 95% credible intervals are shown by the dashed lines. Blue stands for the true distribution and magenta for the approximate distribution.

5.2 Hierarchical modelling of variance

Independent component analysis (Hyvärinen et al., 2001), by its definition, seeks to find components from data that are statistically independent. But often in practice, the found components are only uncorrelated and not independent. Take, as an example, the two components shown in Figure 5.3, which were found by FastICA (Hyvärinen, 1999) from a set of MEG recordings.[2] Although the sources are uncorrelated, it is obvious that some dependencies exist between them. In fact, the physical explanation of the simultaneous burst of activity is that the monitored patient is biting his teeth. The phenomenon depicted in Figure 5.3 is not rare. Nonstationarity of variance is a common characteristic of many natural datasets such as image sequences and recordings of audio (Parra et al., 2001).

The subject of this section are hierarchical models for variance which capture the kind of dependencies illustrated above. The model for noisy ICA is extended in two ways. First, it is modified to find dependencies between the variances of the sources by using a set of higher-level latent variables termed variance sources. Second, the temporal correlations between the variances are taken into account by incorporating dynamics to the model. The models in this section can be implemented using Bayes Blocks so the inference procedure is not discussed.

[2]See (Vigário et al., 1998) for the description of the data.

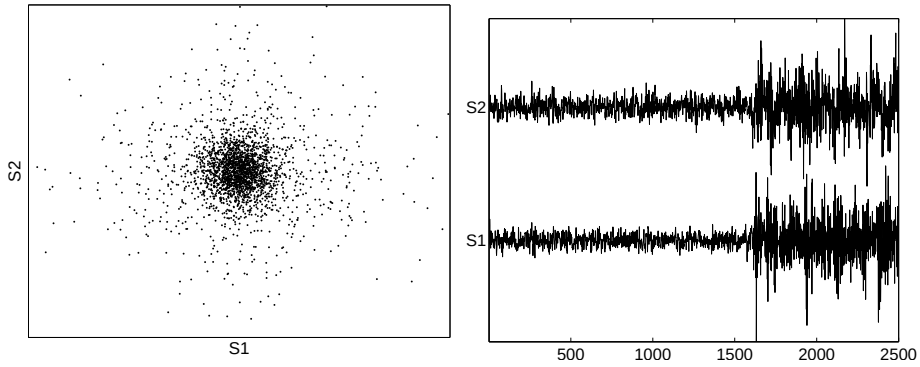


Figure 5.3: Two sources estimated from a set of MEG measurements. The scatter plot (on the left) does not reveal the dependency between the sources although it is obvious in the time-series plot (on the right).

5.2.1 Noisy ICA

The basis for the variance models in the subsequent sections is the noisy ICA model. The N dimensional noisy observations $\mathbf{x}_t$ are modelled as a linear combination of M independent sources $\mathbf{s}_t$:

$$\begin{aligned} \mathbf{x}_t &\sim \mathcal{N}\left(\mathbf{A}\mathbf{s}_t, \mathrm{diag}[\exp(-\mathbf{v}_x)]\right) \\ \mathbf{s}_t &\sim \mathcal{N}\left(\mathbf{0}, \mathrm{diag}[\exp(-\mathbf{u}_t)]\right) . \end{aligned} \tag{5.5}$$

The model is depicted in Figure 5.4(a). Conditional on $\mathbf{u}_t$, the sources have a Gaussian prior, but as the log-variance is allowed to take different values for different samples t, the marginal distribution becomes super-Gaussian. Hence the model can perform ICA (given that the assumption of supergaussianity of the sources holds). The variables $\mathbf{u}_t$ are termed variance nodes, and the subsequent models make abundant use of them.

There are known problems with estimating the noisy ICA model using VB. Ilin and Valpola (2005) have shown that the choice of the posterior approximation affects the obtained solution. A fully factorial approximation, such as that used in Bayes Blocks, favours a solution that is closer to PCA than ICA. In the variance modelling context this is a smaller concern as the objective is to find dependent sources and characterise the dependencies in a meaningful manner.

5.2.2 Variance sources

The example in the beginning of this section suggests an extension to the noisy ICA model. If there are dependencies between the variances of the

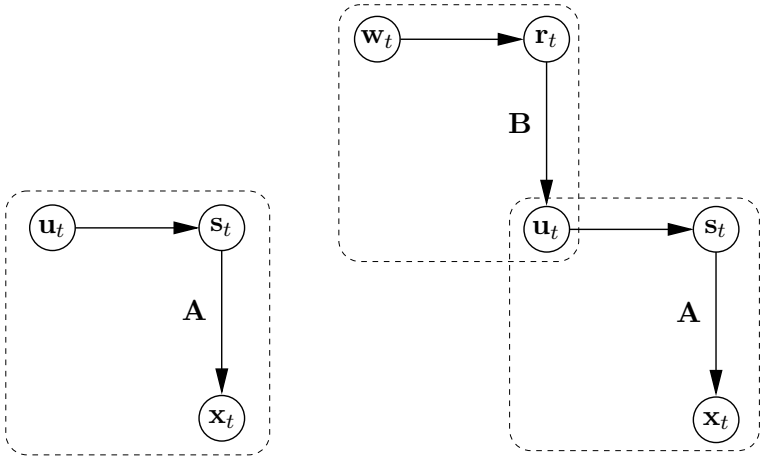


Figure 5.4: (a) The noisy ICA model and (b) the hierarchical variance model.

sources, it might be useful to model them. Several possibilities for capturing the dependencies could be considered, but the approach in Publication III is to use a similar model for the variances as is used for the observations. Figure 5.4(b) illustrates the resulting model structure. The bottom-right block is the noisy ICA model as it reads in (5.5) and the top-left block is a slightly modified replicate:

$$\begin{aligned} \mathbf{u}_t &\sim \mathcal{N}\left(\mathbf{B}\mathbf{r}_t, \mathrm{diag}[\exp(-\mathbf{v}_u)]\right) \\ \mathbf{r}_t &\sim \mathcal{N}\left(\mathbf{r}_{t-1}, \mathrm{diag}[\exp(-\mathbf{w}_t)]\right) . \end{aligned} \tag{5.6}$$

Having the dynamics in the variance sources (5.6) is of immense importance. There is a lot of uncertainty in the variance nodes $\mathbf{u}_t$ which renders expensive modelling of them impossible. If an iid model is used for $\mathbf{r}_t$ instead, the model needs significant evidence of dependency between the variance nodes before it is willing to use a variance source to capture that dependency. When, however, the dynamics are used, the cost of introducing a variance source to the model is considerably smaller which in turn makes it easier for the model to find the dependencies among the variance nodes.

In Publication III the modelling of the MEG data was studied in some detail using the hierarchical variance model. The model found several slowly changing variance sources of which the most prominent ones were related to the biting artifact.

Variance modelling in the context of ICA has been considered by some other authors as well. As a means to source separation, the nonstation-

arity of variance has been studied for example in the works of Pham and Cardoso (2001) and Hyvärinen and Hurri (2004). As a phenomenon in its own right, heteroscedasticity has been of interest especially to people doing research in computational neuroscience. To give some background: ICA applied to patches of images or image sequences yields a bank of filters that resemble the simple cells in the visual cortex of mammals (van Hateren and Ruderman, 1998). To achieve a model behaving like the complex cells, researchers have attempted to model the energies of the simple cells. Hyvärinen and Hoyer (2000) divide the cells into groups of equal sizes and maximise the sparseness within each group to obtain independent subspaces of simple cells. In (Hyvärinen et al., 2001) they use the correlations between the energies of the simple cells to find a topographic ordering for them. An extension to the standard (noise-free) ICA model, bearing similarity to the model presented in this section, has been suggested by Karklin and Lewicki (2005). Their model follows the iid assumption throughout—temporal correlations are not modelled at all. This might be beneficial in some applications, but can make it difficult to find weak, slowly changing variance sources, as discussed above.

5.2.3 Dynamic model for variance

The model in the previous section captured instantaneous variance dependencies between the sources. The dynamics used there was such that each variance source predicted only itself. It seems plausible in some applications, that the variance of one source could be indicative not only of its own future variance but also of future variances of other sources. Consider, for example, a not so hypothetical situation in the stock market. At the onset of a crisis in a certain industry, the variance of the returns of stock X starts to rise. Later the stocks Y and Z, belonging with X to the crisis stricken industry, start to show similar behaviour as X. So the stock X could have been used to predict the behaviour of the stocks Y and Z with an appropriate model.

In this section the noisy ICA model is extended to take temporal variance dependencies into account. Rather than directly modelling the variances of the sources, we will model their innovation processes. This changes the source prior to

$$\mathbf{s}_t \sim \mathcal{N}\left(\mathbf{s}_{t-1}, \mathrm{diag}[\exp(-\mathbf{u}_t)]\right) . \tag{5.7}$$

The sources are assumed to follow a first order AR process with identity dynamics, and the variability of the innovations is controlled by the variance nodes. They too follow a first order AR process, with unconstrained linear dynamics

$$\mathbf{u}_t \sim \mathcal{N}\left(\mathbf{B}\mathbf{u}_{t-1}, \mathrm{diag}[\exp(-\mathbf{v}_u)]\right) . \tag{5.8}$$

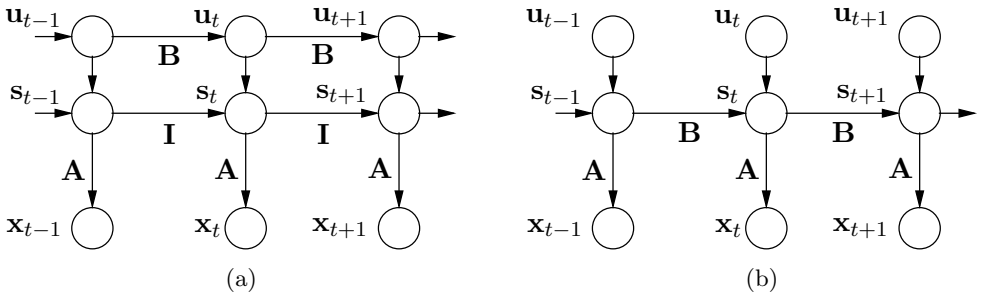


Figure 5.5: (a) An extension of noisy ICA where the temporal relations between the variance nodes are modelled. (b) A model otherwise similar to (a) except that the linear dynamics is moved from the variance nodes to the sources.

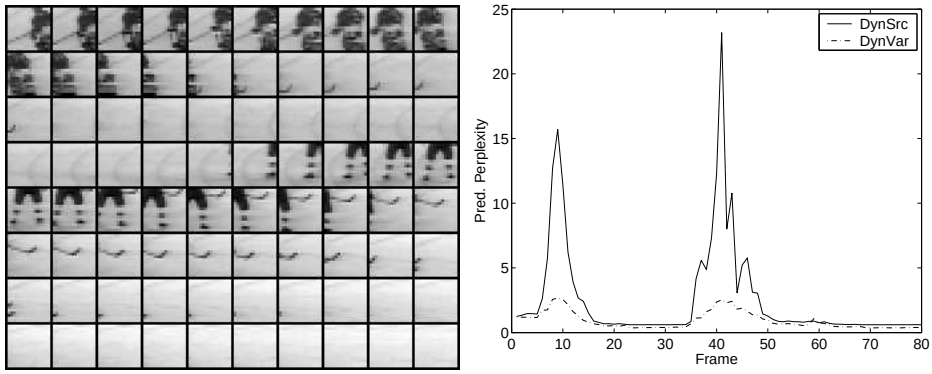
The model is illustrated in Figure 5.5(a).

To show that there exist prominent variance dependencies in realistic data, the model was used to analyse a video from an ice hockey game. To be able to make a quantitative assessment of the model's performance, it was compared to another similar model where the dynamical relations were directly sought in the sources rather than in the variance nodes (see Figure 5.5(b)). The task the two models were asked to perform, was to predict the next frame based on the previous frame. The training data consisted of 4000 frames of size 16x16. The performances of the models were measured as the predictive perplexity (PP)

$$\mathrm{PP} = \exp\left[-\frac{1}{256}\sum_{i=1}^{256}\log p(x_{i,t+1}|\mathbf{X}_{1:t})\right], \tag{5.9}$$

computed over a test set that was composed of the 80 frames shown in Figure 5.6(a). The PP values for these frames are plotted in Figure 5.6(b). Although the variance model does not provide better predictions for the means, it can quantify the uncertainty in its predictions which explains the much better PP values it obtains.

The stock market example in the beginning of this section is indeed not far fetched. Variance modelling in the time-series context is much studied in applied econometrics. There and in related fields variance is called volatility and it is intended to quantify the risk related to a financial instrument over a period of time. It is well known that volatility has temporal correlations and so there exist many methods based on that assumption. The most widely used models are ARCH (autoregressive conditional heteroscedasticity, Engle (1982)) and its generalisation GARCH (Bollerslev, 1986). The appeal of ARCH is its simplicity; the model estimation can be done with the ordinary least squares method. Other models have been suggested that—unlike



(a) (b)

Figure 5.6: (a) The test frames in the hockey image sequence and (b) the corresponding prediction accuracies for the two models measured as predictive perplexities. The shorthands DynSrc and DynVar refer to the models in Figure 5.5. DynVar corresponds to model (a) and DynSrc to model (b).

GARCH, wherein the conditional variance is a function of past squared errors and variances—postulate the volatility as a latent stochastic process. These kind of models are referred to as stochastic volatility models, and although they are more expressive compared to ARCH, their estimation is also considerably more difficult. Kim et al. (1998) present several MCMC based methods to this end which have a number of positive qualities. The use of MCMC makes the methods flexible in that the model can easily be extended to have more complicated dynamics. Also model selection can be done in a principled manner by calculating Bayes factors. A more recent method is the one by Zoeter et al. (2004). They present a fast, expectation propagation and Gaussian quadrature based approach to stochastic volatility. The first pass of their algorithm resembles unscented Kalman filtering (Julier and Uhlmann, 1997) but the initial posterior approximation can be improved by iterating the algorithm.

5.3 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.

Quantile regression (Koenker and Bassett, 1978; Yu et al., 2003) is the close equivalent of predictive uncertainty in statistics. The idea is to estimate the conditional quantiles of the response variable and thereby summarise the whole conditional density. By doing so, one can get as detailed description of the conditional density as one desires. The estimation of the quantiles, however, is not without its problems and often the methods suffer from the curse of dimensionality. In machine learning, quantile regression has not been so well known,[3] and the methods for predictive uncertainty have mostly relied on ideas already well established in the field such as Gaussian processes (Goldberg et al., 1998; Snelson and Ghahramani, 2006; Kersting et al., 2007) and neural networks (Weigend and Nix, 1994; Bishop, 1994; Williams, 1996; Cawley et al., 2006).

In Publication IV, 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 multilayer perceptron (MLP) networks to model both the location and the scale of the output distribution. The model can be summarised as

$$\begin{aligned} y_t &\sim \mathcal{N}\left(\mathrm{MLP}_y(\mathbf{x}_t), e^{-u_t}\right) \\ u_t &\sim \mathcal{N}\left(\mathrm{MLP}_u(\mathbf{x}_t), \tau^{-1}\right) . \end{aligned} \tag{5.10}$$

Above, y_t is the response variable and $\mathbf{x}_t$ is the vector of explanatory variables. 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 model is illustrated in Figure 5.7 for the case of four inputs, three hidden nodes in MLP_y, and two hidden nodes in MLP_u.

In the beginning of this chapter, it was demonstrated that the learning of heteroscedastic models can be difficult for simple methods. It was also shown that variational Bayes can largely avoid the associated problems. Unfortunately, VB for nonlinear models, such as that in Eq. (5.10), becomes

[3] The trend might be changing, though. See e.g. (Meinshausen, 2006).

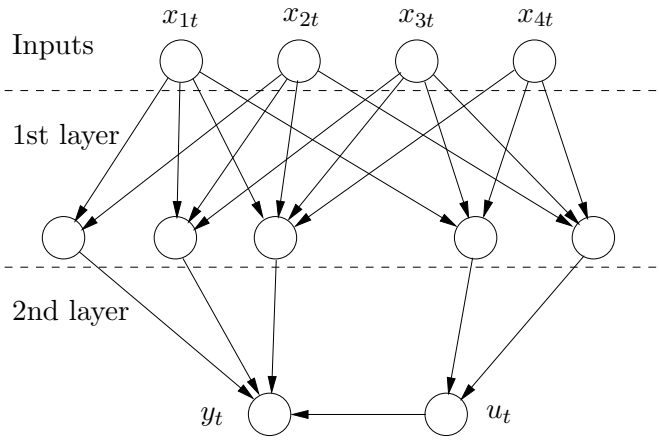


Figure 5.7: The model (5.10) schematically illustrated. In this example instance, there are four explanatory variables, three hidden nodes in MLP_y, and two hidden nodes in MLP_u.

involved both in analytic as well as in computational terms.[4] Hence the learning algorithm in Publication IV is based on the slightly weaker approximation technique, the variational EM algorithm, and only the the latent variables u_t and all the parameters in the second layer of the model have distributional estimates. The parameters in the first layer of the model, that is, the first-layer weights of the MLPs, have point estimates only.

Denoting the first-layer parameters as $\boldsymbol{\xi}$ and the second-layer parameters as $\boldsymbol{\psi}$, the outcome of the VEM algorithm is a distributional estimate $q(\boldsymbol{\psi})$ and a point estimate $\hat{\boldsymbol{\xi}}$. The predictive pdf, which is the object of interest in this modelling task, is then formally obtained from the integral

$$p(y_t|\mathbf{x}_t, \hat{\boldsymbol{\xi}}, \mathbf{X}, \mathbf{Y}) = \int p(y_t|u_t, \boldsymbol{\psi}, \mathbf{x}_t, \hat{\boldsymbol{\xi}}) p(u_t|\mathbf{x}_t, \boldsymbol{\psi}, \hat{\boldsymbol{\xi}}) q(\boldsymbol{\psi})\, d\boldsymbol{\psi}\, du_t\,.$$

The distribution $q(\boldsymbol{\psi})$ tends to be narrow and hence it can be approximated by a delta distribution $\delta(\boldsymbol{\psi} - \langle\boldsymbol{\psi}\rangle)$ without compromising the predictive density. Depending on the parameter τ, the distribution $p(u_t|\mathbf{x}_t, \boldsymbol{\psi}, \hat{\boldsymbol{\xi}})$ can have large variance in which case the predictive density would be poorly approximated if the integration over u_t were neglected. In Publication IV this integral was approximated with a suitably constructed finite mixture of Gaussians.

The importance of integrating over u_t, even if only approximately, is demonstrated in Figure 5.8. There, predictive densities for a one dimensional

[4] It is nevertheless not impossible to apply VB to such models. See e.g. the papers by Hinton and van Camp (1993), Barber and Bishop (1998), and Honkela and Valpola (2005) for examples where VB has been applied to nonlinear models by various means.

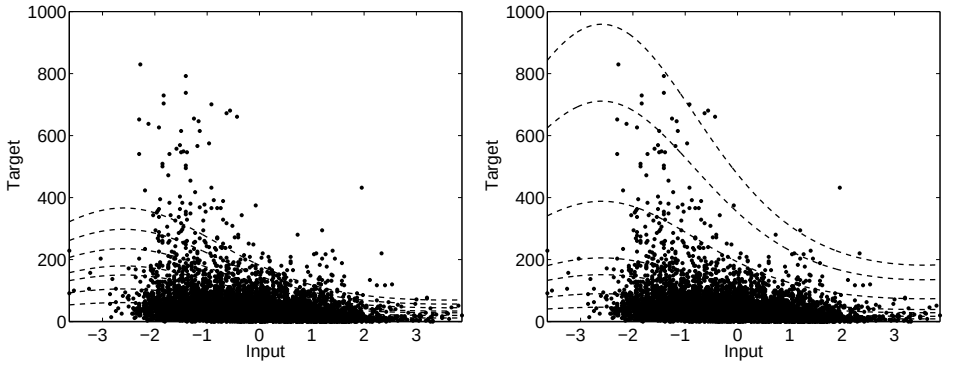


Figure 5.8: The effect of marginalising over u_t on the predictive pdf. The dashed lines denote the 0.5, 0.75, 0.9, 0.95, 0.99, 0.999 and 0.9999 credible intervals. On the left, the pdf of u_t has been collapsed to a delta distribution. On the right, marginalisation over the pdf of u_t has been approximately performed.

problem are shown. The first pdf is computed by collapsing $p(u_t|\mathbf{x}_t, \boldsymbol{\psi}, \hat{\boldsymbol{\xi}})$ to a delta distribution; the second pdf is computed by the approximate marginalisation discussed above. We can see that the first predictive pdf does not model the tail of the data distribution well by noting that far too many datapoints from the set of 10000 fall outside the 0.9999 credible interval (the topmost dashed line). The second predictive pdf does not suffer from this shortcoming.

The method summarised in this section was applied to all four datasets in the "Predictive uncertainty in environmental modelling" competition held at WCCI'06. The datasets varied in dimensionality from one input variable to 120 variables. The detailed results with the proposed approach can be found in Publication IV, and the summary of the competition, including results with other methods, is given in (Cawley et al., 2006). The proposed method performed well with all the datasets where heteroscedasticity was an important component being the overall winner of the competition.

Chapter 6

Astronomical applications

Two applications of approximate Bayesian inference to astronomical data analysis are discussed in this chapter. First, a problem of finding a set of prototypical star population spectra underlying a set of observed galaxy spectra is presented. Second, the estimation of time delays in gravitational lenses is discussed. The proposed solutions to these two problems are general machine learning algorithms—no detailed astrophysical modelling is involved. So although the methods are offered as solutions to the specific problems in astronomy, they are of wider applicability.

6.1 Analysis of galaxy spectra

6.1.1 Background

Consider a set of galaxy spectra such as that shown in Figure 6.1. Each spectrum is a collection of measured fluxes over a range of wavelengths. The overall shape of the spectrum as well as the wiggles (absorption lines) can be used to determine the age and the chemical properties of the galaxy. It is a fairly recent observation that some of the galaxy spectra can be composed of several stellar-subpopulation spectra (Nolan, 2002). In the earlier work, the subpopulations have been found by fitting a superposition of single stellar population models. The approach has relied on a brute-force search over a properly discretised parameter space of the model to find the decomposition. Needless to say, this has been computationally intensive. The data explosion in astronomy due to large sky survey projects has made it ever so important that analysis methods be applicable to large datasets. This has been one chief motivation for developing the methods presented in this section.

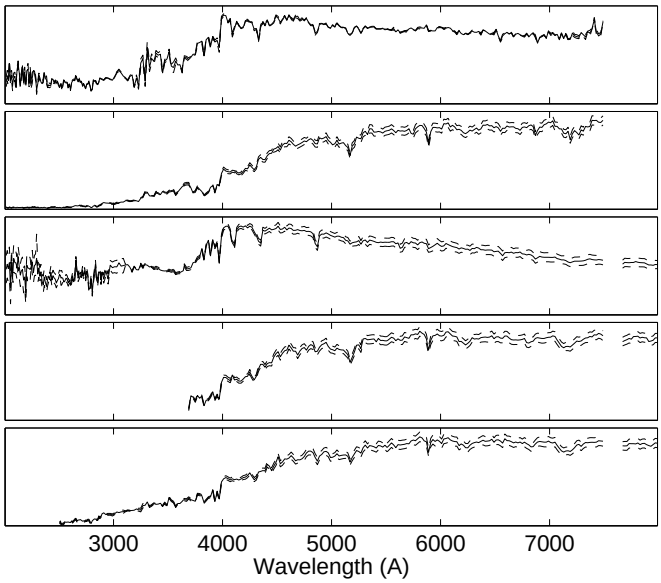


Figure 6.1: The spectra of five galaxies. The dashed lines show the measurement uncertainties in the data and the blank entries stand for missing values.

We denote the observed spectra as $\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_T) = (x_{it})$, where i indexes the galaxies and t the wavelength bins, and we denote the unobserved stellar subpopulation spectra as $\mathbf{S} = (\mathbf{s}_1 \dots \mathbf{s}_T) = (s_{jt})$, where j indexes the subpopulation and t the wavelength bins with one-to-one correspondence to those of the observed spectra. The problem is to find $\mathbf{S}$ given $\mathbf{X}$. The astrophysics related to the problem enters the modelling in two simple statements: (1) energies are positive and (2) they add up linearly. Two further assumption are made: (3) the spectral prototypes are independent in their distribution and (4) there is additive Gaussian noise in the observations. These specifications translate into the following probabilistic model

$$\begin{aligned} \mathbf{x}_t &\sim \mathcal{N}\left(\mathbf{A}\mathbf{s}_t, \mathrm{diag}(\boldsymbol{\tau})^{-1}\right) \\ \mathbf{s}_t &\sim \prod_j p(s_{jt})\,, \end{aligned} \tag{6.1}$$

with the constraints $p(s_{jt} < 0) = 0$ and $p(a_{ij} < 0) = 0$. The actual form of the distributions $p(s_{jt})$ is so far left unspecified. A priori, any form would as long as the distribution has nonnegative support.

6.1.2 Rectified factor analysis

The generative model in (6.1) looks much like the one for noisy ICA discussed in Section 5.2.1. Indeed, the term nonnegative noisy ICA could be

used for its estimation. For nonnegative ICA there exists several algorithms whose derivations are based on other principles than Bayesian inverse modelling; among them are the methods by Oja and Plumbley (2004) and Zheng et al. (2006). The much used nonnegative matrix factorisation (Lee and Seung, 1999) is not strictly ICA as there is no assumption of independence in the method, but it is nevertheless often used in source separation settings. The extension of nonnegative matrix factorisation by Hoyer (2004) adds sparseness constraints to the method, which brings it closer to ICA.

The Bayesian framework has some clear benefits though. For one, it is straightforward to handle missing values[1] and uncertainties in the measurements—both being features that are present in our application. Model comparison can also be done rigorously, which makes it possible to compare different modelling assumptions and to infer the model order. In the current application, this means that several hypotheses about the number of underlying stellar subpopulations can be tested.

The target in the modelling is to obtain estimates of the sources $\mathbf{s}_t$ and mixing proportions $\mathbf{A}$. How can this be accomplished? Even if the marginal posterior distributions could be computed, they would be useless due to symmetries in the model. Here, the variational Bayesian method is adopted both for the sake of computational efficiency as well as for breaking the symmetry in the model as VB approximates only one of the modes in the posterior pdf.

One of the first works on applying VB to nonnegative ICA is by Miskin (2000). As the prior distributions for the sources he uses rectified Gaussians:

$$p(s_{jt}) = \mathcal{N}^R(s_{jt}|m_j, v_j) = \frac{1}{Z(m_j, v_j)} u(s_{jt}) \mathcal{N}(s_{jt}|m_j, v_j) \ , \qquad (6.2)$$

where $Z(m_j, v_j) = \frac{1}{2}\operatorname{erfc}(-m_j/\sqrt{2v_j})$. This prior is convenient because then the model is in the conditional conjugate family.[2] The computations, however, are not tractable in the VB framework unless the location parameter m_j is set to zero so that the awkward normaliser vanishes. This has the unfortunate side effect that distributions biased toward zero are favoured. Examples of this phenomenon are presented later on in this section both in an artificial setting as well as in the galaxy spectra application.

Another way to formulate a nonnegatively supported prior is to specify it hierarchically: let s_{jt} be rectified version of a further latent variable r_{jt} i.e. $s_{jt} = \operatorname{cut}(r_{jt}) := \max(r_{jt}, 0)$. Now it does not make a difference what the

[1] Jaynes (2003) on missing data: "This is a problem that does not exist for us; Bayesian methods work by the same algorithm whatever data we have."

[2] Multiplying the Gaussian likelihood with the rectified Gaussian prior yields a rectified Gaussian posterior.

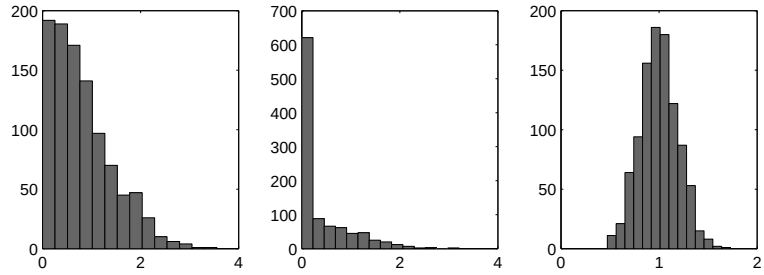


Figure 6.2: Histograms of the sources in the control experiment.

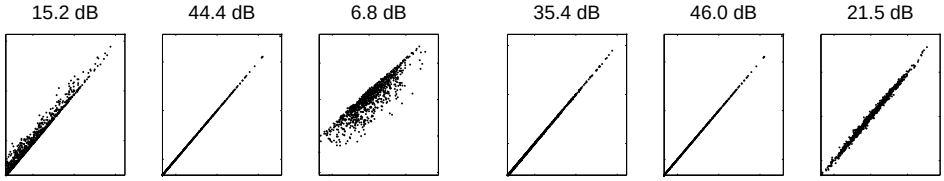
distribution of r_{jt} is—the nonnegativity constraint is automatically satisfied. Computationally it does matter, of course, and the prior is set to an ordinary Gaussian: $p(r_{jt}) = \mathcal{N}(r_{jt}|m_j, v_j)$. It might appear that an essentially same prior for s_{jt} is obtained as in Eq. (6.2). This is not the case though. If we marginalise r_{jt} out, the distribution of s_{jt} is a mixture of a rectified Gaussian and a Dirac delta at zero. The most important benefit of the chosen prior is the ability to have both of the hyperparameters m_j and v_j in the model. The prior can also be modified for modelling correlations between different s_{jt}, which makes it possible to consider autoregressive or other variants of the model.[3]

The variational inference procedure for the model sketched above is presented in detail in Publication V. Although the model is not in the conditional conjugate family, a variational Bayesian EM algorithm with free-form fully-factorial posterior approximation is tractable. The nonstandard part of the inference is the update rule for the factors r_{jt}. The free-form approximate posterior can be shown to be a mixture of a positive and a negative rectified Gaussian distribution:

$$q(r_{jt}) = \pi_{jt}^{+} \mathcal{N}^{R+}(r_{jt}|m_{jt}^{+}, v_{jt}^{+}) \; + \; \pi_{jt}^{-} \mathcal{N}^{R-}(r_{jt}|m_{jt}^{-}, v_{jt}^{-}) \,. \tag{6.3}$$

The effect the prior distribution of the sources has on the separation performance is demonstrated next. The model with zero-location rectified Gaussian priors is called positive factor analysis (PFA) and the model with rectification nonlinearities is called rectified factor analysis (RFA). The problems that the priors in PFA cause are well illustrated by the following control experiment. Three sources, whose histograms are shown in Figure 6.2, are mixed to obtain ten observations. Both of the models, PFA and RFA, are learnt, and the separation results are compared to the ground truth. In Figure 6.3 the estimated sources are plotted against the true sources. The performance is measured as the signal-to-noise ratio between the ground truth and the estimate, and these measures are shown above each plot.

[3]Publication V discusses the autoregressive variant.



(a) PFA (b) RFA

Figure 6.3: The separation results in the control experiment using both PFA and RFA. The estimated source is plotted against the ground truth, and the corresponding signal-to-noise ratio is shown above each plot.

RFA can reconstruct the original sources with high accuracy whereas PFA cannot. It is the third source that poses most difficulties to PFA; the SNR is as low as 6.8 dB. This source is also the one that disagrees the most with the prior used in PFA.

6.1.3 Results

In Publication VI, RFA and various other models were applied to the spectral dataset and the results were compared to the physical-model based approach. Here a few findings from that study are raised.

One of the questions that was initially asked was how many stellar subpopulations could be used to explain the spectra. There was a strong prior belief that no more than two subpopulations underlie the observations. The Bayesian evidence framework agreed that two sources are indeed enough. In Figure 6.4 the log-evidence[4] is plotted as the function of model order ranging from one source to four sources. The curves are shown for PFA and RFA.

The physical interpretability of the sources was another major concern in the study. The decompositions to two sources with PFA and RFA are shown in Figure 6.5. The first components of both models are almost identical. By comparison to the physical model, it was found to represent a typical old stellar subpopulation. The second component from RFA, on the other hand, was found to resemble a young subpopulation. In this instance, the two methods, PFA and RFA, did not agree with each other. With PFA, the second component is distorted toward zero and contains some spurious absorption lines. This is most likely due to the prior in PFA which pulls the posterior toward zero. This mismatch between the likelihood and the prior is clearly visible in the evidence plot: there is a considerable gap between

[4]Or more appropriately, the lower bound for the log-evidence.

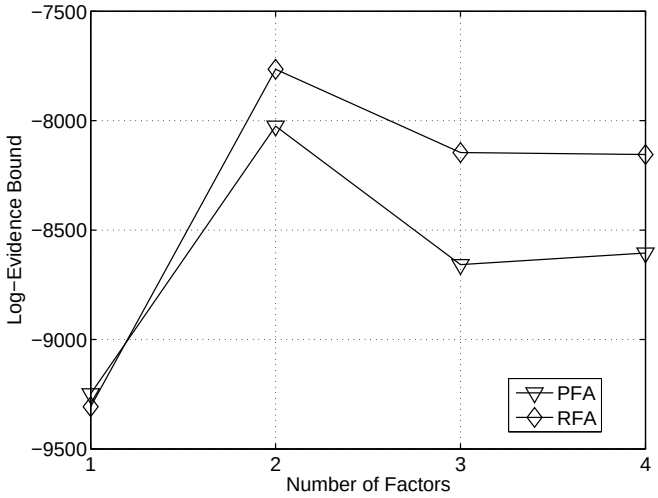
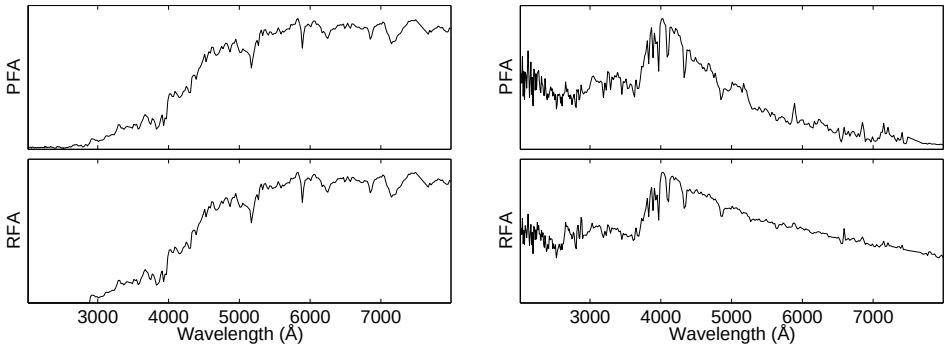


Figure 6.4: The log-evidence as the function of the number of factors.



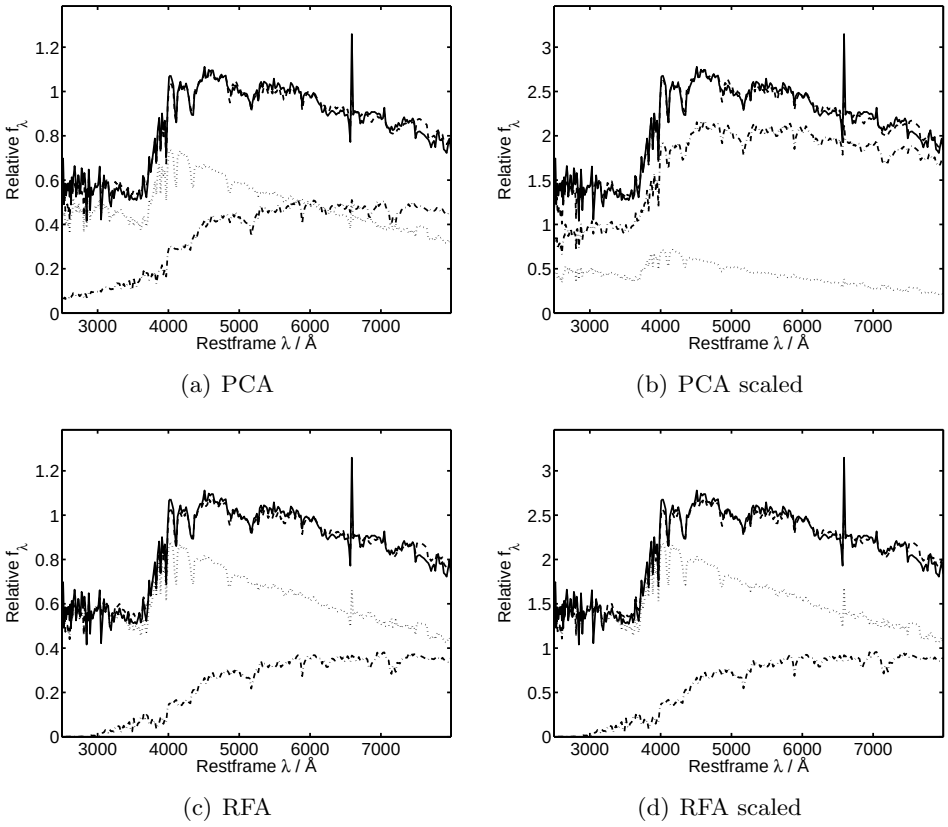
(a) First source (b) Second source

Figure 6.5: The first and the second source estimated from the galaxy spectra data using both PFA and RFA.

PFA and RFA in favour of the latter method.

Decompositions with model orders higher than two were studied too, but it was indeed found that the subsequent components did not have any physical interpretation—a finding that was coherent with the Bayesian evidence analysis.

In Publication VI, PCA produced fairly similar results compared to those of RFA. As discussed in the article, this is a matter of luck mainly, as PCA is not a method for source separation. When, however, the eigenvalues of the data covariance matrix differ substantially, PCA does distinguish between different rotations of the sources. This has clearly been the case with



(a) PCA

(b) PCA scaled

(c) RFA

(d) RFA scaled

Figure 6.6: The effect of rescaling the spectra on PCA and RFA. Since PCA is inherently not a source separation method, the scaling drastically changes the proportions of the two components used to explain the spectrum. The decomposition with RFA is identical with and without the scaling.

the particular dataset used in Publication VI. Although it will come as no surprise to anybody working in the field of source separation, it is here anyway demonstrated that a slight rescaling of some of the observations drastically changes the decomposition in the case of PCA. It is also shown that the decomposition found by RFA is not affected by the rescaling exercise. Figure 6.6 presents a spectrum and its decomposition to two sources as found by either PCA or RFA. With PCA there is a considerable change to be noted in the decomposition when the data is rescaled whereas the results with RFA stay almost identical.

After the initial study in Publication VI, Nolan et al. (2007) have applied RFA to a much larger dataset, obtained from the archives of Sloan Digital Sky survey, which consisted of over 13000 early-type galaxy spectra. The findings with this much richer dataset were similar to those in Publication VI, the conclusion from the methodological point of view being that

RFA is indeed capable of finding the young stellar subpopulation when present in a galaxy spectrum. The authors of that study see RFA as "a powerful tool for studying in detail, both globally and individually, the evolution of early-type galaxies" (Nolan et al., 2007).

6.2 Estimation of time delays in gravitational lensing

6.2.1 Background

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 Figure 6.7 for illustration). 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. The significance of estimating the delays in such systems stems from the early observation that they can be used in determining important cosmological quantities (Refsdal, 1964).

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.[5] 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. With all the gaps and the noise in the data, the interpolation can, however, introduce spurious features to the data which cause the cross-correlation analysis to fail (Cuevas-Tello et al., 2006).

6.2.2 Bayesian time-delay estimation with irregularly sampled signals

In Publication VII, a method for estimating the delay between irregularly sampled signals is presented. Since interpolation on the data that is noisy and contains gaps has its risks, that is avoided. Instead the two observed

[5] In the regular-sampling case, even the Bayesian formulation of the delay-estimation problem has the cross-correlation function as a sufficient statistic (Scargle, 2001).

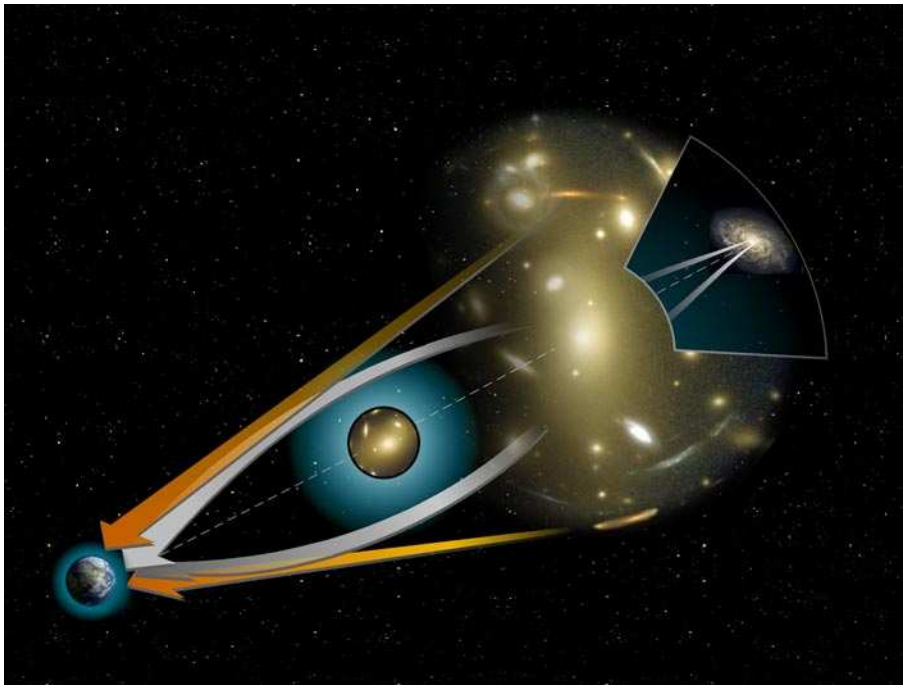


Figure 6.7: Schematic illustration of gravitational lensing. (Image courtesy of NASA.)

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. This is illustrated in Figure 6.8.

The subsequent instances of the source need of course to be related to each other in some manner to make the estimation of the delay possible. There are two requirements that seem sensible: (1) when separated by a small time gap ϵ, the source instances $s(t)$ and $s(t+\epsilon)$ should be close to each other, and (2) when separated by a large gap E, the source instances $s(t)$ and $s(t+E)$ should not be strongly dependent on each other. A prior that satisfies the above informal requirements is the Wiener process:

$$s(t_{i+1}) - s(t_i) \sim \mathcal{N}\left(0, [(t_{i+1} - t_i)\,\sigma]^2\right) . \tag{6.4}$$

This prior encodes the notion of "slow variability" into the model which is an assumption that is implicitly present in many of the other methods as well. Indeed, if the source would fluctuate a lot compared to the sampling frequency, it would render the delay estimation problem practically impossible.

The latent source can be marginalised out of the model analytically which leads to specific type of Kalman-filter equations. In addition to the de-

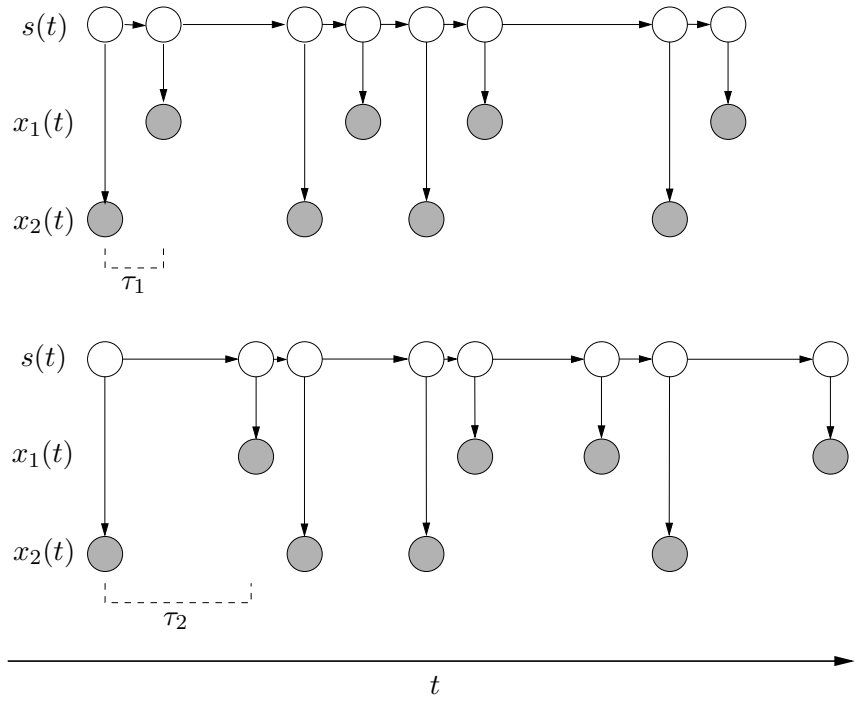


Figure 6.8: The delay τ affects the observation times and hence the structure of the model.

lay, there are several parameters in the model controlling the scaling and shifting of the source as well as the noise levels. These are all nuisance parameters and they are further marginalised out using a Metropolis-Hastings sampler. The dimensionality of the parameter space, after having got rid of the source, is counted in tens. For this low dimensionality sampling is computationally efficient. In what follows, the above sketched Bayesian delay estimation algorithm is called BEDBUSS (short for "Bayesian Estimation of Delays Between Unevenly Sampled Signals").

As the delays in real gravitational lensing systems are not known for certain, controlled comparisons to other methods must be made with artificial data where the ground truth is known. In Publication VII, BEDBUSS is compared against three popular methods. These are the discrete correlation function (Edelson and Krolik, 1988; Lehar et al., 1992), interpolation followed by cross-correlation analysis (e.g. Kundic et al., 1997), and the dispersion spectra (Pelt et al., 1994). Here, results with a recently developed kernel-based method (Cuevas-Tello, 2007) are shown also.

Three groups of datasets were generated, the SNR of the observations being different in each of them.[6] Examples of the datasets are shown in Figure 6.9.

[6]The SNRs were 20 dB, 14 dB, and 8 dB. This range was motivated by real datasets:

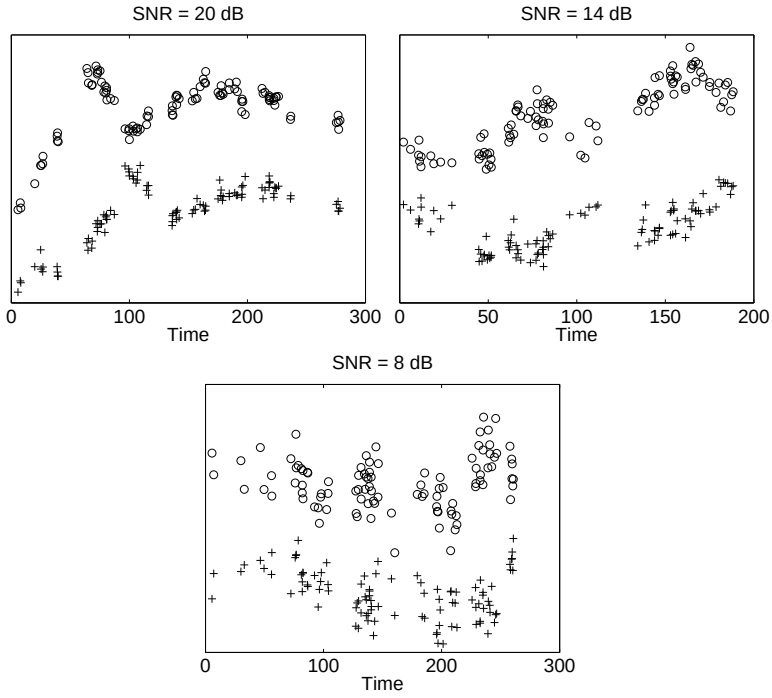


Figure 6.9: One example dataset from each of the three groups. The SNR of the group is shown above each plot.

In low noise, it is rather easy to visually identify the delay, which was 35 units in all of the cases. In high noise, the problem is already considerably more difficult. The five methods, including the one proposed here, were used to estimate the delay for each dataset, and the average absolute errors were computed.[7] These are shown in Figure 6.10. All the methods perform well in low noise but the accuracies of the first three methods start to deteriorate in medium and high noise. Between those three methods, the performance does not vary much. The Kernel method does slightly better in low noise than any other method, but loses somewhat to BEDBUSS in medium and high noise.

with the particular real datasets discussed in the next section, the estimated SNRs range from 10 dB to 25 dB.

[7]The simulations with the kernel method were performed by its author for his thesis work and the results for this method are quoted as they are presented in the thesis (Cuevas-Tello, 2007).

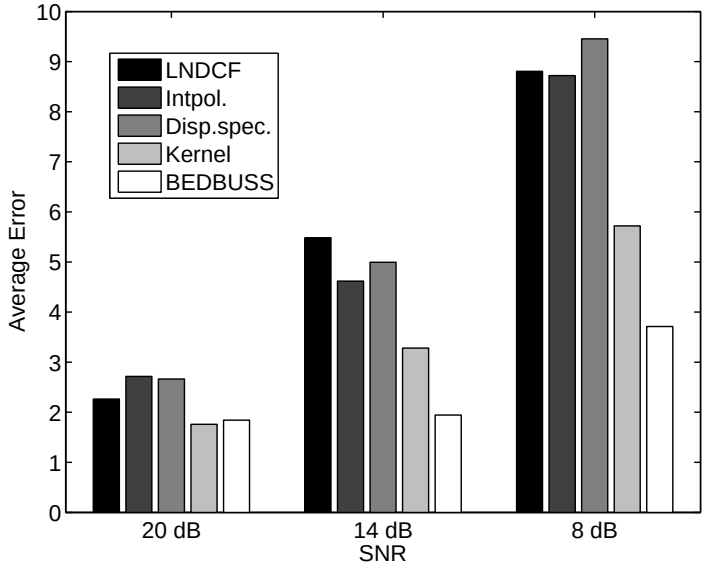


Figure 6.10: Average errors of the methods for the three groups of datasets. The shorthand LNDCF stands for locally normalised discrete correlation function and intpol. stands for interpolation followed by standard cross-correlation analysis.

6.2.3 Results

We have applied BEDBUSS to several gravitational lensing systems and have reported initial results from that study in an astronomy meeting (Harva and Raychaudhury, 2005). Here the method is illustrated with two lensing systems: B0218 and PG1115. Hubble telescope images of the lenses as well as the measured time-series are shown in Figure 6.11. B0218 serves as an example of a system where little controversy over the delay exists, whereas PG1115 is an example of the opposite: the estimates of its delays vary depending on who is doing the estimation.

With BEDBUSS we obtain the posterior distribution of the delay, or rather a fair amount of samples from it. To compare with previous estimates of the delays, we summarise the posterior by its mean and std. These values along the previous estimates found in the literature are shown in Table 6.1.

In the case of B0218, we get a very similar estimate of the delay compared to the previous attempt. With PG1115, however, the situation is somewhat different. Although our estimates are not in strident disagreement with the earlier measurements, they do not exactly equal them either. But as already said, PG1115 is a lensing system over which there has been controversy before, as can be noted in Table 6.1. The posterior distributions of the

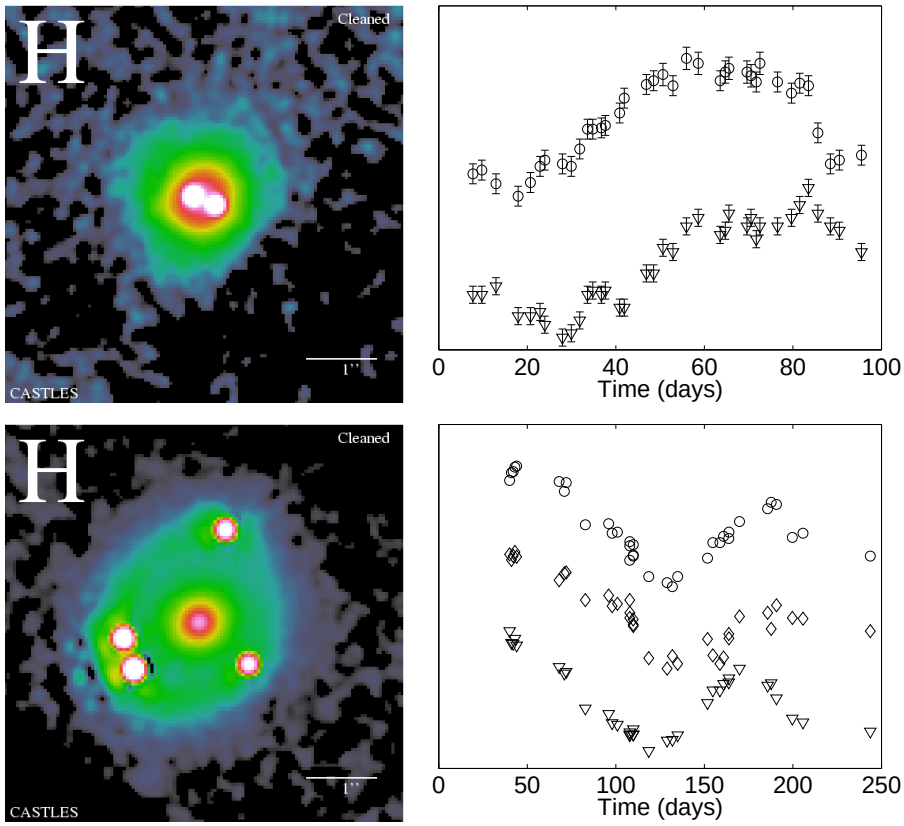


Figure 6.11: Top: the two images of B0218 and the corresponding intensity measurements. Bottom: Same for PG1115 (it has four images of which the two closest to each other are merged). The images were obtained from CASTLES.

System	Images	Our Delay	Previous Measures	Reference
B0218+357		10.9 ± 0.7	10.5 ± 0.4	Biggs et al. (1999)
PG1115+080	AC	-11.7 ± 1.7	-13 ± 1	Barkana (1997)
			-9.4 ± 3.4	Schechter (1997)
	BC	-22.7 ± 1.8	-25 ± 1	Barkana (1997)
			-23.7 ± 3.4	Schechter (1997)

Table 6.1: Our estimates of time delays compared to previous results.

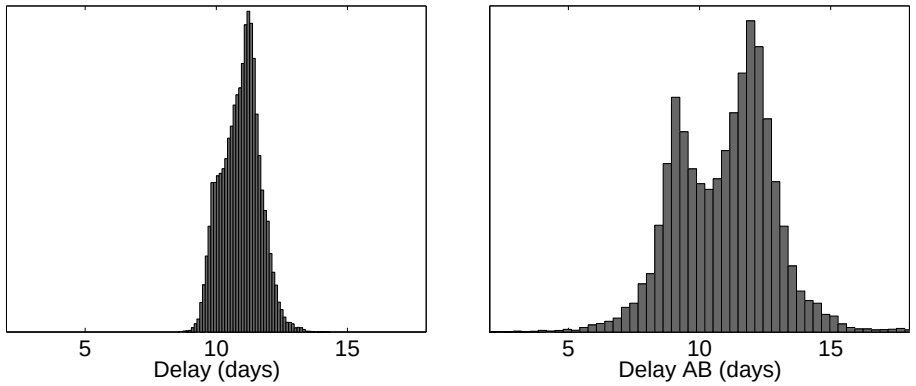


Figure 6.12: Left: the posterior over the delay in B0218. Right: same for the delay AB in PG1115.

delays for these two systems, shown in Figure 6.12, clearly point to the reason why it might be easy to get consistent results with the one system and inconsistent with the other. Whereas the distribution of B0218 delay is well concentrated around its mean value, the distribution of PG1115 delay spans a wide range of values and is strongly multimodal. This suggests that the data obtained from PG1115 so far is not sufficient for precise determination of the delays.

Chapter 7

Discussion

In a recent meta study (Poirier, 2006), the impact of Bayesian inference was investigated. Articles in economics and statistics journals were classified to be either Bayesian or non-Bayesian by a simple criterion of whether the word "Bayes" or "Bayesian" was contained in the text. The findings do not come as a surprise: in statistics there has been a steady upwards trend of Bayesianity since the 1970's, with an especially sharp rise in the mid 90's, most likely caused by the MCMC revolution. In the other fields investigated, the growth was not found to be that dramatic. Indeed, Poirier concludes his article writing: "The pessimistic Bayesian might say things have barely started in the other disciplines."

Vast contributions to the development of Bayesian probability theory having come from physicists, and astrophysicists in particular, one could easily conjecture that it is the mode of statistics most applied in astrophysics. That is not quite the case. Loredo (1990) mentions the irony of Laplace, one of the notable figures in astronomy, having been strongly Bayesian and yet the Bayesian approach being little used among astronomers. Since Loredo's paper, Bayesian methods have become more popular in the field, but as noted by Scargle (2001), the pace in which this happens is "agonizingly slow." He believes that the reason for this modest rate is the (perceived) complexity in implementing Bayesian procedures, and that the remedy will be easy-to-use tools for Bayesian analysis becoming available.

Bayesian methods, and machine learning algorithms in general, are usually tailored to solve one specific problem at a time. New problems then call for tailoring of new algorithms. It is clear that similar modelling patterns recur again and again in applications, and thus it would be highly useful if those common denominators could somehow be captured and reused without always having to start the modelling exercise from scratch. Bayes Blocks, the

variational Bayesian inference framework discussed in Chapter 4, is an effort to that direction. For a specific model family, it completely automates the inference procedure. It does suffer, of course, from the same problem that any framework does. When one's model falls outside the framework's domain, one either needs to extend the framework to suit one's wishes or work without its assistance. It is hardly likely that any inference engine could be a panacea to each and every conceivable problem.

Sometimes Bayesian methods are criticised for being overly complex, that simpler methods could be used to solve the same problems as accurately and more efficiently. And perhaps the criticism is on some occasions to the point. The Bayesianisation of an algorithm is sensible only if there are some benefits from doing so. The question is then, what are the benefits of applying Bayesian methods? With the problems studied in this thesis, the answer varies. In the case of the variance models discussed in Chapter 5, the answer is that simpler methods produce, if not numerical explosions as the precision parameters tend to infinity, then at least some less severe form of overfitting. To put it shortly, simple methods just do not work satisfactorily in those problems.

Consider as another example the delay estimation task of Section 6.2. Using the same model, an estimate of the delay could be found by a gradient search on the marginal likelihood, without the need to resort to Monte Carlo integration. But what about quantifying the uncertainty, which there seems to be plenty? Ad hoc devices for obtaining error bars can be found in the literature. Compute the estimate by leaving one of the observations out and repeat for each observation. The mean and standard deviation of this procedure then serve as the estimate and error bars. But why leave just one observation out at a time? Why not two, three, or four? By controlling this number one can obtain as wide or as narrow error bars as one desires. This is practical but not so rigorous.

Rigour is certainly one of the chief appeals of Bayesian inference. That it is derived from first principles which are easy to grasp and to accept, makes it a trustworthy method. Perhaps it will indeed be a matter of the computational techniques to get advanced enough, for Bayesian inference to become the standard tool for statistical data analysis.

Bibliography

Barber, D. and Bishop, C. (1998). Ensemble learning for multi-layer networks. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems 10*, pages 395–401. The MIT Press, Cambridge, MA, USA.

Barkana, R. (1997). Analysis of time delays in the gravitational lens PG 1115+080. *The Astrophysical Journal*, 489:21–28.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. J. Wiley.

Biggs, A. D., Browne, I. W. A., et al. (1999). Time delay for the gravitational lens system B0218+357. *Monthly Notices of the Royal Astronomical Society*, 304:349–358.

Bishop, C. M. (1994). Mixture density networks. Technical Report NCRG/4288, Neural computing research group, Aston University.

Bishop, C. M. (1999). Variational principal components. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'99)*, volume 1, pages 509–514, Edinburgh, UK.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Cawley, G. C., Haylock, M. R., and Dorling, S. R. (2006). Predictive uncertainty in environmental modelling. In *Proc. 2006 IEEE World Congress on Computational Intelligence (WCCI'06)*, pages 11096–11103, Vancouver, BC, Canada.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13.

Cuevas-Tello, J. C. (2007). *Estimating Time Delays between Irregularly Sampled Time Series*. PhD thesis, School of Computer Science, University of Birmingham, Birmingham, UK.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38.

Edelson, R. A. and Krolik, J. H. (1988). The discrete correlation function: a new method for analysing unevenly sampled variability data. *The Astrophysical Journal*, 333:646–659.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50:987–1007.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, Florida.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511.

Ghahramani, Z. and Beal, M. (2001). Propagation algorithms for variational Bayesian learning. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 507–513. The MIT Press, Cambridge, MA, USA.

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532.

Goldberg, P., Williams, C. K. I., and Bishop, C. M. (1998). Regression with input-dependent noise: A Gaussian process treatment. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*. MIT Press, Cambridge, MA.

Gregory, P. C. (2005). *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge University Press.

Harva, M. and Raychaudhury, S. (2005). A new Bayesian look at estimation of gravitational lens time delays. In *Abstracts RAS National Astronomy Meeting 2005*, Birmingham, UK.

Hastings, W. K. (1970). Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Heskes, T. and Zoeter, O. (2002). Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proc. 18th Conf. on Uncertainty in Artificial Intelligence (UAI'02)*, pages 216–233, San Francisco, CA.

Hinton, G. E. and van Camp, D. (1993). Keeping neural networks simple by minimizing the description length of the weights. In *Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, pages 5–13, Santa Cruz, CA, USA.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417.

Honkela, A. and Valpola, H. (2004). Variational learning and bits-back coding: an information-theoretic view to Bayesian learning. *IEEE Transactions on Neural Networks*, 15(4):800–810.

Honkela, A. and Valpola, H. (2005). Unsupervised variational Bayesian learning of nonlinear models. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 593–600. MIT Press, Cambridge, MA, USA.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12:1705–1720.

Hyvärinen, A., Hoyer, P., and Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, 13:1527–1558.

Hyvärinen, A. and Hurri, J. (2004). Blind separation of sources that have spatiotemporal dependencies. *Signal Processing*, 84(2):247–254.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. J. Wiley.

Ilin, A. and Valpola, H. (2005). On the effect of the form of the posterior approximation in variational learning of ICA models. *Neural Processing Letters*, 22(2):183–204.

Jaakkola, T. S. and Jordan, M. I. (1997). Bayesian logistic regression: a variational approach. In *Proc. 1997 Conf. on Artificial Intelligence and Statistics*, pages 283–294.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. In Jordan, M., editor, *Learning in Graphical Models*, pages 105–161. The MIT Press, Cambridge, MA, USA.

Julier, S. J. and Uhlmann, J. K. (1997). A new extension of the Kalman filter to nonlinear systems. In *Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defense Sensing, Simulation and Controls*.

Karklin, Y. and Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation*, 17:397–423.

Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W. (2007). Most likely heteroscedastic Gaussian process regression. In *Proc. 24th International Conference on Machine learning (ICML'07)*, pages 393–400, Corvalis, Oregon.

Kim, S., Shepard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Kundic, T., Colley, W. N., et al. (1997). A robust determination of the time delay in 0957+561A, B and a measurement of the global value of Hubble's constant. *The Astrophysical Journal*, 482:75–82.

Lappalainen, H. and Miskin, J. (2000). Ensemble learning. In Girolami, M., editor, *Advances in Independent Component Analysis*, pages 75–92. Springer-Verlag, Berlin.

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

Lehar, J., Hewitt, J. N., Burke, B. F., and Roberts, D. H. (1992). The radio time delay in the double quasar 0957+561. *The Astrophysical Journal*, 384:453–466.

Lehman, R. S. (1955). On confirmation and rational betting. *The Journal of Symbolic Logic*, 20(3):251–262.

Loredo, T. J. (1990). From Laplace to Supernova SN 1987A: Bayesian inference in astrophysics. In Fougere, P. F., editor, *Maximum Entropy and Bayesian Methods*, pages 81–142. Kluwer Academic Publishers.

MacKay, D. J. C. (1995). Developments in probabilistic modelling with neural networks—ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications. Proc. of the 3rd Annual Symposium on Neural Networks*, pages 191–198.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Minka, T. (2001). Expectation propagation for approximate Bayesian inference. In *Proc. 17th Conf. on Uncertainty in Artificial Intelligence (UAI'01)*, pages 362–369.

Miskin, J. (2000). *Ensemble Learning for Independent Component Analysis*. PhD thesis, University of Cambridge, UK.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Neal, R. M. (1995). Suppresing random walks in Markov chain Monte Carlo using ordered overrelaxation. Technical Report 9508, Dept. of Statistics, University of Toronto.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31(3):705–767.

Neal, R. M. and Hinton, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 355–368. The MIT Press, Cambridge, MA, USA.

Nolan, L. (2002). *The Star Formation History of Elliptical Galaxies*. PhD thesis, The University of Edinburgh, UK.

Nolan, L., Raychaudhury, S., and Kabán, A. (2007). Young stellar populations in early-type galaxies in the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 375(1):381–387.

Oja, E. and Plumbley, M. (2004). Blind separation of positive sources by globally convergent gradient search. *Neural Computation*, 16:1811–1825.

Opper, M. and Winther, O. (2005). Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2204.

Parisi, G. (1998). *Statistical Field Theory*. Westview Press.

Parra, L., Spence, C., and Sajda, P. (2001). Higher-order statistical properties arising from the non-stationarity of natural signals. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 786–792. The MIT Press, Cambridge, MA, USA.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pelt, J., Kayser, R., Refsdal, S., and Schramm, T. (1994). Time delay controversy on QSO 0957+561 not yet decided. *Astronomy & Astrophysics*, 286(1):775–785.

Pham, D.-T. and Cardoso, J.-F. (2001). Blind separation of instantaneous mixtures of nonstationary sources. *Signal Processing*, 49(9):1837–1848.

Poirier, D. J. (2006). The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, 1(4):969–980.

Refsdal, S. (1964). 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.

Scargle, J. D. (2001). Bayesian estimation of time series lags and structure. In *Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2001)*.

Schechter, P. L. (1997). The quadruple gravitational lens PG 1115+080: Time delays and models. *The Astrophysical Journal Letters*, 475:L85–L88.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Skilling, J. (2006). Nested sampling for general Bayesiam computations. *Bayesian Analysis*, 1(4):833–860.

Snelson, E. and Ghahramani, Z. (2006). Variable noise and dimensionality reduction for sparse Gaussian processes. In *Proc. 22nd Int. Conf. on Uncertainty in Artifical Intelligence*, Arlington, Virginia. AUAI Press.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1995). BUGS: Bayesian inference using Gibbs sampling. Available at `http://www.mrc-bsu.cam.ac.uk/bugs/`.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

van Hateren, J. H. and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1412):2315–2320.

Vigário, R., Jousmäki, V., Hämäläinen, M., Hari, R., and Oja, E. (1998). Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*, pages 229–235. MIT Press, Cambridge, MA.

Weigend, A. S. and Nix, D. A. (1994). Predictions with confidence intervals (local error bars). In *Proceedings of the Internatioal Conference on Neural Information Processing (ICONIP'94)*, pages 847–852, Seoul, Korea.

Williams, P. M. (1996). Using neural networks to model conditional multivariate densities. *Neural Computation*, 8(4):843–854.

Winn, J. and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, 6:661–694.

Winther, O. and Petersen, K. B. (2007). Flexible and efficient implementations of Bayesian independent component analysis. *Neurocomputing*, 71:221–233.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350.

Zheng, C.-H., Huang, D.-S., Sun, Z.-L., Lyu, M. R., and Lok, T.-M. (2006). Nonnegative independent component analysis based on minimizing mutual information technique. *Neurocomputing*, 69:878–883.

Zoeter, O., Ypma, A., and Heskes, T. (2004). Improved unscented Kalman smoothing for stock volatility estimation. In *Proc. 2004 IEEE International Workshop on Machine Learning for Signal Processing*, pages 143–152, São Luis, Brazil.