Approximating methods for intractable probabilistic models: Applications in Neuroscience

Pedro A. d. F. R Højen-Sørensen

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

April 1, 2001

Department for Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Kongens Lyngby, Denmark

Abstract

This thesis investigates various methods for carrying out approximate inference in intractable probabilistic models. By capturing the relationships between random variables, the framework of graphical models hints at which sets of random variables pose a problem to the inferential step. The approximating techniques used in this thesis originate from the field of statistical physics which for decades has been facing the same type of intractable computations when analyzing large systems of interacting variables e.g. magnetic spin systems. In general, these approximating techniques are known as *mean field methods*.

The thesis provides a brief introduction to the basic methodology of learning and inference in graphical models as well as a short review of the various types of mean field approximations which recently have been shown to be efficient for carrying out approximate inference in intractable probabilistic models.

Starting from the naive mean field approximation we derive for the independent component analysis (ICA) model with instantaneous mixing general expressions for the posterior quantities needed to perform learning by Expectation-Maximization (EM). Furthermore, we explore the feasibility of going beyond the naive mean field approximation for this model. In fact, it turns out that the overcomplete ICA problem can be solved using a simple linear response correction to the mean sufficient statistics obtained by naive mean field approximation. In addition, we apply to the ICA problem an adaptive version of the [Thouless, Anderson and Palmer 1977] (TAP) mean field approach which is due to [Opper and Winther 2000c].

To illustrate the methodology on a real world problem, an explorative analysis of a functional magnetic resonance imaging (fMRI) dataset from a visual activation study is carried out using ICA with binary sources. It is shown this approach, which is computationally efficient, infers reasonable brain activation functions.

Finally, we outline various ways of carrying out approximate message passing in probabilistic models for which marginalization over some of the clique variables is intractable.

Resumé (Abstract in Danish)

I nærværende afhandling undersøgers forskellige teknikker til at udføre approximativ inferens i beregningsmæssigt tunge probabilistiske modeller. Med udgangspunkt i teorien om grafiske modeller er det muligt direkte at få et indblik i, hvornår approximerende metoder er påkrævede. De teknikker, som er benyttet i denne afhandling, har alle deres oprindelse i statistisk fysik. I løbet af de sidste årtier har folk indenfor dette område udviklet approximerende metoder for at kunne analysere systemer med mange vekselvirkende enheder som f.eks. magnetiske spin systemer. Disse metoder går samlet under betegnelsen *middelfeltsmetoder*.

Denne afhandling giver foruden en kort introduktion til inferens og parameterestimation i grafiske modeller også en oversigt over de forskellige metoder, som i tidens løb har vist sig egnede for approximativ inferens i probabilistiske modeller.

Med udgangspunkt i den naive middelfeltsapproximation udleder vi generelle udtryk for kilde-posteriorsandsynligheden i en model for "independent component analysis" (ICA) hvor kilderne blandes instantant. For denne model undersøger vi fordelene ved at benytte mere advancerede approximationer. Det viser sig, at det underbestemte tilfælde, hvor der er flere kilder end mikrofoner, kan løses ved en lineær responskorrektion af de sufficiente statistikker, som fås fra den naive middelfeltsapproximation. Endeligt anvender vi til dette ICA problem en adaptiv version af [Thouless, Anderson and Palmer 1977] (TAP) middelfeltsmetoden, som blev foreslået af [Opper and Winther 2000c].

Vi illustrerer metoden i en explorativ analyse af en sekvens af dynamiske hjerneskanbilleder optaget under et visuelt aktiveringsstudie. Det er vist, at denne metode, som er beregningsmæssig effektiv, rent faktisk er i stand til at finde plausible hjerneaktiveringsmønstre.

Endelig gives der forslag til, hvordan det er muligt at lave approximativ sekventiel inferens i probabilistiske modeller, hvor eksakt marginalisering over enkelte klikkepotentialer er umuligt.

Preface

The present thesis has been submitted in partial fulfillment for the Ph.D. degree in electrical engineering. The work documented in this thesis has been carried out at the Department for Informatics and Mathematical Modelling, Section for Digital Signal Processing at the Technical University of Denmark. The project was supervised by professor Lars Kai Hansen, associate professor Jan Larsen and Dr. Carl Edward Rasmussen. The work was commenced in February 1998 and completed in April 2001 with a six months stay at UC Berkeley in the fall of 1999.

The reader is expected to be well-versed in the most common machine learning techniques and terminology. These subjects are not introduced in this thesis since excellent treatments on these matters have existed for decades which means that the most common terminology and techniques of this field at the present time is regarded as well established. Furthermore, basic knowledge of the principles behind using magnetic resonance imaging techniques for in vivo imaging of the human brain is appreciated but not a prerequisite.

I have tried to set out the text in such a way that the main features stand out clearly; it may sometimes seem that I go to great lengths to explain the obvious, but that is how I am.

During the Ph.D. study the following papers have been written,

- de Freitas J. F. G., **Højen-Sørensen P. A. d. F. R.**, Jordan M. I. and Russell S.: Variational MCMC. Submitted to the *17th Conference on Uncertainty in Artificial Intelligence.* (2001).
- **Højen-Sørensen P. A. d. F. R.**, Winther O. and Hansen L. K.: Analysis of Functional Neuroimages using ICA with Adaptive Binary Sources. Submitted to *Journal of Neurocomputing* (2001).
- de Freitas J. F. G., Andrieu. C., **Højen-Sørensen P. A. d. F. R.**, Niranjan M. and Gee A. H.: Sequential Monte Carlo Methods for Neural Networks. In *Sequential Monte Carlo Methods in Practice.* Doucet A., de Freitas J. F. G. and Gordon N. (editors). Springer-Verlag. 2001.
- **Højen-Sørensen P. A. d. F. R.**, Winther O. and Hansen L. K.: Mean Field Approaches to Independent Component Analysis. Submitted to *Journal of Neural Computation* (2000).

- **Højen-Sørensen P. A. d. F. R.**, Winther O. and Hansen L. K.: Ensemble Learning and Linear Response Theory for ICA. In *Advances in Neural Information Processing Systems*, (NIPS 13), NIPS*00. (2000).
- **Højen-Sørensen P. A. d. F. R.**, de Freitas J. F. G. and Fog T.: On-line Probabilistic Classification with Particle Filters. In *Proceeding of IEEE International Workshop on Neural Networks for Signal Processing.* NNSP*00, (2000).
- **Højen-Sørensen P. A. d. F. R.**, Hansen L. K. and Rasmussen C. E.: Bayesian modelling of fMRI time series. In *Advances in Neural Information Processing Systems*, (NIPS 12), NIPS*99. (1999).
- **Højen-Sørensen P. A. d. F. R.**, Hansen L. K. and Rostrup E.: A Bayesian approach for estimating activation in fMRI time series. In *Proceedings of the 5th Int. Conf. on Functional Mapping of the Human Brain.* NeuroImage (1999).

Nomenclature

An attempt has been made to use standard symbols and operators consistently throughout the presentation. Although most symbols and operators are introduced along the way, the reader should with no effort be able to infer the meaning of any non-defined symbol from its context. In some passages of the text I have chosen to spontaneously suppress variables and indices to make the main features stand out more clearly. Again, the reader should be able to infer the exact details in these cases.

In a hopeful attempt to make the present exposition more readable I have chosen to make overloaded use of the expectation operator $\langle\cdot\rangle$. To be specific, given a function $f(\Omega)$ of a set of variables Ω, we define

$$\langle f\rangle_{1|\omega} \equiv \int d(\Omega\backslash\omega) f(\Omega) \ , \tag{0.1}$$

where $\omega \subseteq \Omega$, i.e. we integrate out all variables except the ones belonging to ω. For instance, suppose $f(x, y, z)$ is the joint probability density of the random variables X, Y and Z. Then, the marginal density $f(z)$ can be written as $\langle f(x, y, z)\rangle_{1|z}$. Furthermore, we define $\langle\cdot\rangle_1 \equiv \langle\cdot\rangle_{1|\emptyset}$ such that e.g. the normalizing constant for an unnormalized probability density f is given by $\langle f\rangle_1$.

Acknowledgements

I would like to thank my supervisors professor Lars Kai Hansen, Jan Larsen and Carl Edward Rasmussen for carrying out a liberal yet vigilant supervision of my Ph.D. project which allowed me the academic freedom to satisfy my own curiosity, while ensuring that I was following a reasonable path. Their vast experience in each of their respective fields have been of immense value and I have benefitted greatly from each of them.

I sincerely wish to express my gratitude to professor Mike Jordan for making it possible to visit him and his students at the Computer Science Division, University of California at Berkeley in the fall of 1999. It turned out to be six truly rewarding months both academically and socially. Participating in his highly stimulating reading groups gave me all new insights into many subjects of Machine Learning; insights that would otherwise have been much more painful to obtain. A warm thank goes to Kevin Murphy whom from the very first week made me feel welcomed and helped me with all sorts of practical stuff in Berkeley. I also thank Andrew Ng, Nando de Freitas and the rest of the RUGS members for inspiring discussions; in this context I would like to thank professor Stuart Russell for allowing me to attend the RUGS meetings.

The Technical University of Denmark is acknowledged for allowing me the opportunity of doing this work. The Otto Mønsted foundation and NIPS foundation is acknowledged for financial support to travel activities.

Thanks also goes to my fellow Ph.D. students for many hours of fun and good discussions. Furthermore, a special thanks goes to Thomas Fabricius, Ole Winther and Manfred Opper with whom enlightening and inspiring discussions motivated to get some work done.

I also thank my family and friends, in particular the members of the (well...almost) biweekly dinner-club ("madklub"), for helping me through the numerous times where all aspects of research seemed pointless.

Finally but most importantly, I thank Mette, my fiancee, for her endless love, support and encouragement.

Copenhagen, April 1, 2001

Pedro A. d. F. R. Højen-Sørensen

Contents

1. Introduction

The Road goes ever on and on
Down from the door where it began.
Now far ahead the Road has gone,
And I must follow, if I can,
Pursuing it with eager feet,
Until it joins some larger way
Where many paths and errands meet.
And wither then? I cannot say.

J.R.R. Tolkien

The ability to make inference is vital for any learning device to adapt and make decisions in changing environments. This is true for both human beings as well as machine learning algorithms. While human beings are able to carry out the inferential step quite efficiently e.g. to make predictions into the immediate future, this task is from a probabilistic point of view computationally intractable since it may involve calculations of high dimensional sums or integrals. Needless to say, it would be of great scientific and technological importance to find efficient approaches to solve this problem since it would constitute a significant step towards learning in structures as complex as the one found in the human brain.

1.1 The Brain, Graphical Models and Statistical Physics

The human brain is made up of approximately 10^{11} neurons which are organized into 10^7 elementary networks, each of which consists of 10^4 densely interconnected neurons. Through functional activation studies many of these elementary networks have been shown to be highly specialized to a specific task. The collective behavior of networks of densely interconnected neurons has provided us with a highly adaptive learning architecture which enables us to adapt rapidly to changing environments. Whereas many of these networks have been specialized and refined through biological evolution others are being modified constantly in our daily life. Another advantage of distributed architectures is that of robustness. Due to the dense connectivity we would expect the performance to these architectures to degrade gracefully.

From a practical point of view it would be nice if there existed a framework we could use in a principled way to come up with (and solve) new and interesting machine learning algorithms and which, in principle, could incorporate knowledge of how the brain is organized. Indeed there is such a framework, namely that of probabilistic networks and expert systems also known as graphical models. This framework provides a principled way of merging different *experts* or *specialized models* into one system which is probabilistically consistent. Like the human brain, the resulting model is a highly structured stochastic system. One powerful property of the framework of graphical models is that it makes use of the principle of modularity by exploiting the structure of the system to make inference by local and (hopefully) efficient computations.

For many systems, however, it turns out to be computationally intractable to carry out these local computations. This happens in particular when the elements (say the neurons in the brain or the random variables in a statistical model) involved in the local computation are densely interconnected. For several decades statistical physicists have been facing the same type of intractable computations when analyzing large systems of interacting variables e.g. magnetic spin systems. In the pursue to tackle this problem they have developed a rich class of approximating methods, collectively known as *mean field methods*, which in the last decade have been successfully applied to a large selection of intractable probabilistic models.

1.2 Functional neuroimaging

As mentioned already, it is of great scientific and technological interest to be able to make inference in highly structured stochastic system of similar scale as the human brain. Provided the ultimate goal in this scientific endeavor is to simulate the human brain we would need to have knowledge about e.g. the connectivity in the networks of neurons and furthermore know the functional significance of the different networks. Obviously, this can only be achieved through an enormous amount of interdisciplinary efforts. We will here take on a more pragmatic position and use small scale machine learning architectures or probabilistic models to analyze data from functional activation studies. Hence, the goal of this, in comparison, modest task is to gain insights into the distribution of the functional areas in the human brain.

Although it is well known that the brain undergoes physical changes when exposed to sensory input e.g. by modulating the strength of the synaptic junction, I still fell that the following disclaimer is imperative:

If you choose to proceed reading this text, you agree on getting your brain physically modified. Oops — you already did — sorry.

Thesis overview

This thesis is organized into seven chapters and eight appendices. The first chapter serves as an general introduction to the present exposition, while the remaining chapters form the main part of the thesis concerning methods for approximate inference in intractable probabilistic models and their applications to the analysis of functional neuroimages. In more detail, the contents of the individual chapters and appendices are:

Chapter 1 gives a general introduction to the thesis.

Chapter 2 provides a short introduction to learning and inference in graphical models. This chapter introduces two canonical models which are useful to have in mind throughout this thesis.

Chapter 3 reviews some of the various mean field approximations which have been successfully used to carry out approximate inference in intractable probabilistic models.

Chapter 4 makes use of advanced mean field methods to solve the intractable inference problem encountered in the generative model for probabilistic independent component analysis (ICA) with instantaneously mixed sources.

Chapter 5 presents a explorative analysis of functional magnetic resonance imaging (fMRI) data using probabilistic ICA with adaptive binary sources. This chapter also considers the problem of determining the number of latent sources.

Chapter 6 outlines various strategies for carrying out approximate message passing. However, this chapter, which presents work in progress, somehow lacks experimental support. The material has been included since it binds together most of the topics considered in this thesis.

Chapter 7 summarizes the work presented and outlines possible conclusions. Suggestions for some possible directions to carry on this work are also provided.

Appendix A summarizes the useful results for marginalizing and conditioning on variables in the Gaussian probability density.

Appendix B-H contains reprints of selected papers which have been authored and co-authored during the Ph.D. study.

2. Learning and Inference in Graphical Models

This chapter briefly introduces the basic notion of conditional independence which together with graph theory provide the theoretical foundation of localized computations for inference in probabilistic models. The material considered in this chapter is kept at a rather operational level, hence we will be omitting a large part of the proofs supporting the underlying theory. Readers interested in such details are referred to the excellent books of [Jensen 1996; Jordan 1998; Cowell et al. 1999; Jordan and Bishop 2001] which this chapter strongly relies on. Besides of providing a brief introduction to learning and inference in graphical models, this chapter mainly serves the purpose of introducing the terms and notation which will be used throughout this thesis.

2.1 Directed Acyclic Graphs (DAGs)

The idea of introducing graph theory in probabilistic modeling might at first sight seem unnecessary. However, not only is the graph topology a useful tool to assist in the actual modeling process of a specific problem, but it also bare the solution to how computationally efficient algorithms for doing inference can be devised. Using probability theory as our main starting point we will motivate the introduction of the graph theoretical concepts as they come in handy.

In this section we will consider graphs which are directed and acyclic. A graph is said to be *directed* if all its edges are directed and *acyclic* if it does not possess any cycles (along any directed path). A *directed path* can never cross itself and movement along a path never goes against the directions of the edges. A directed graph which is acyclic is called a *directed acyclic graph* (DAG). One important feature of DAGs is that they possess a, however not unique, *topological ordering*, i.e. it is always possible to find an ordering of the nodes such that for each node S_i all of its parents π_i precedes it in the ordering. The chain rule of probability theory states that any joint probability distribution $p(\boldsymbol{s})$ of N random variable $(S_1, S_2, \ldots, S_N)$ can be factorized as

$$p(s_1, s_2, \ldots, s_N) = \prod_{i=1}^{N} p(s_i | s_1, s_2, \ldots, s_{i-1}) \ , \tag{2.1}$$

where $p(s_1|s_0) \equiv p(s_1)$. Knowing this, we are now able to relate the factorization eq. (2.1) of an arbitrary joint distribution to the topology of a DAG. It is noted that the conditional probability associated to each random variable S_i in eq. (2.1) is conditioned on all the random variables $(S_1, S_2, \ldots, S_{i-1})$, i.e. the random variables are in fact a valid topological ordering with respect to conditioning. Using the fact that we can always relabel the nodes of a DAG into a topological ordering we see that the parents π_i of a node S_i are just the conditioning nodes in its conditional distribution, $p(s_i|s_1, s_2, \ldots, s_{i-1})$, in the factorization eq. (2.1). Figure 2.1(a) shows an example of a DAG with three nodes. All possible joint probability distributions of three random variables can be factorized according to this DAG. At this point it is not clear what insight we have achieved in relating the graph topology to the factorization of the joint distribution with respect to the DAG. This, however, becomes clear when edges starts to be removed from the graph. This naturally leads us to the introduction of Bayesian networks. A *Bayesian network* is a directed acyclic graph whose structure defines a set of conditional independence properties. To each node we associate a (local) conditional probability distribution, where the conditioning is on the parents of the node. The joint distribution is given by

$$p(\boldsymbol{s}) = \prod p(s_i|\pi_i) \ , \tag{2.2}$$

where $p(s_i|\pi_i)$ are the *local conditional probabilities* associated with the graph. We say that eq. (2.2) is a *recursive factorization* according to the DAG. As an example, consider the DAG in figure 2.1(b) which is a subgraph of the DAG shown in figure 2.1(a). Comparing the chain rule factorization eq. (2.1) with the recursive factorization eq. (2.2) for this Bayesian network it is seen that $p(s_3|s_1, s_2) = p(s_3|s_2)$, i.e. S_3 is *conditionally independent* of S_1 given S_2 which we usually write as $S_3 \perp S_1|S_2$. This shows that missing edges in the graph have a probabilistic interpretation in terms of conditional independence. Taking this example to the extreme by removing all edges we arrive at the DAG shown in figure 2.1(c) which implies that S_1, S_2 and S_3 are independent random variables. This is easily verified using the recursive factorization and Bayes rule. It is important to note that whereas missing edges in the graph necessarily do imply independence the edges that are present do not necessarily imply dependence.

We have now seen that a graphical model is associated with a family of probability distributions. In fact, as edges are being removed from the DAG the harder it gets to be a member of the associated family of probability distributions. Figure 2.1(a-c) illustrated this by considering successive subgraphs starting from the large family containing all joint distribution of three random variables and ending at the small family containing only fully factorized joint distributions. The conditional independence relation implied by these simple DAGs was readily obtained using Bayes rules but for large Bayesian networks such direct procedures of determining conditional independence be-

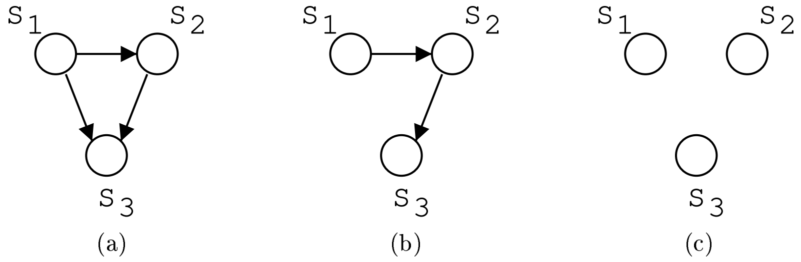


Fig. 2.1. Three DAGs consisting of 3 nodes implying different conditional independence assumption; (a) implies no conditional independence relations and thus able to capture all joint distribution consisting of three random variables; (b) implies the conditional independence relation $S_3 \perp S_1 | S_2$; (c) implies independent random variables.

tween random variables becomes quite tedious. It turns out, however, that the topology of the graph provides the inferential machinery for answering questions about probability distributions without one having to resort to direct calculations which, when done by hand, are prone to errors. This naturally leads to the introduction of the notion of *d-separation*.

2.1.1 d-separation

The notion of d-separation (shorthand for "directed separation") allows conditional independencies to be read directly from the graph. It is essentially the probabilistic counterpart to naive graph separation in the sense that it looks at the probabilistic connectivity instead of just the topological connectivity of the edges in the DAG. Using Bayes rule one can easily determine the probabilistic connectivity of the serial, diverging and converging connection shown in figure 2.2. The arrows in the figure shows the probabilistic connectivity of the DAG and the shaded nodes are the instantiated nodes. E.g., figure 2.2(c) shows that two nodes in a converging connection are marginal independent but *not* conditional independent given the intermediate node. This behavior is typically referred to as *explaining away*.

In addition to the serial, diverging and converging connections it is useful to consider the connection between a single parent and its single child. To determine the probabilistic connectivity of such *boundary connection* it is useful to note that the conditional independence statement $S_A \perp S_B | S_C$ is a property of the marginal distribution $p(s_A, s_B, s_C)$. Since marginalizing over a childless node is equivalent to simply removing the node (and all its edges to its parents) we can basically just add an extra node while leaving the joint distribution of the original parent and child invariant. We see that valid insertions of an extra node leads to either a serial or diverging connection and we can now read of the probabilistic connectivity of the boundary connection simply by using figure 2.2.

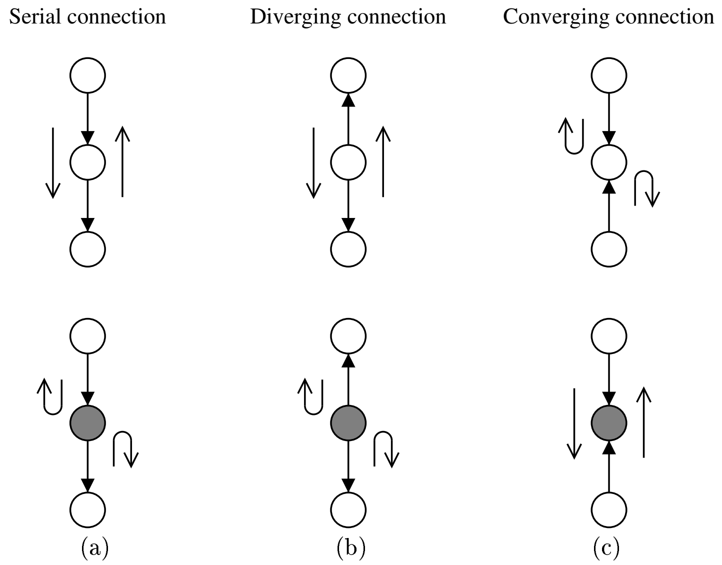


Fig. 2.2. Summarizes the concept of d-separation. The figure illustrates how two nodes communicates (shown as arrows) when their intermediate node yields an (a) serial (b) diverging and (c) converging connection. The top row shows the case there the intermediate node is not instantiated and the lower row shows the case when conditioning on the intermediate node (shown as shading). It is seen that the serial connection and the diverging connection is d-separated when conditioning on the intermediate node whereas the converging connection is d-separated when the intermediate node is not instantiated.

By repeatedly using the probabilistic connectivity of the serial, diverging and converging connections we are able to answer questions about conditional independence in complicated DAGs. This procedure is the *Bayes ball algorithm* of [Shachter 1998]. We have now motivated the following definition of d-separation.

Definition 2.1.1 (d-separation). [Jensen 1996] *Two variables A and B in a directed acyclic graph are d-separated if for all paths between A and B there is an intermediate variable C such that either*

1. *the connection is serial or diverging and the state of C is known*
2. *the connection is converging and neither C nor any of C's descendants have received evidence.*

An extension to d-separation is, however, needed to determine conditional independence relationships in Bayesian networks in which some of the nodes are deterministically determined given their parents. This leads to the notion of *D-separation* [Geiger et al. 1990]. The extension consists of regarding the nodes which are deterministically related to observed nodes as instantiated.

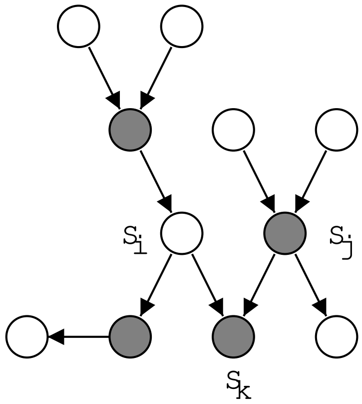


Fig. 2.3. Shows the Markov blanket $\boldsymbol{M}_i$ (illustrated by the shaded nodes) of node i. Instantiating node j removes the possibility of "explaining away" dependencies due to node k.

The *Markov blanket* $\boldsymbol{M}_i$ of a variable S_i is the parents of i, the children of i and the variables sharing a child with i. An example of a Markov blanket is shown in figure 2.3. Clearly, a node will always be d-separated from the rest of the network when all variables in its Markov blanket are instantiated.

In section 2.4.2 we will consider a particular DAG which is able to capture a large class of interesting models used in statistics, in particular Hidden Markov models, the Kalman filter/smoother and independent component analysis.

2.2 Undirected Graphical Models

One appealing property of Bayesian networks is that the joint probability distribution can be expressed as a product of local functions on the DAG associated to the network. In fact, these local functions turned out to be the local conditional probabilities $p(s_i|\pi_i)$. The general procedure for performing efficient inference in graphical models does not, however, directly exploit the topological structure of DAGs. Instead it makes use of the structure of another class of models referred to as *undirected graphical models* or *Markov Random Fields* (MRFs). A graph is said to be *undirected* if *all* the edges in the graph are not directed. Undirected and directed acyclic graphs are both special cases of *chain graphs* which are graphs that have no directed cycles.

As for a Bayesian network, the joint probability distribution of an undirected graphical model can be expressed as a product of local functions. In particular, we say that a probability density $p(\boldsymbol{s})$ factorize with respect to a given undirected graph $\mathcal{G}$ if

$$p(\boldsymbol{s}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(s_c) \ , \tag{2.3}$$

where the product is over the set of all cliques, $\mathcal{C}$, in the graph and ϕ_c is the *potential function* associated with clique c, i.e. a non-negative function depending only on the nodes, S_c, in the clique. A *clique* is a complete graph which is maximal in the sense that the graph can not be extended to include additional nodes without losing the property of being complete. A graph is called *complete* if every pair of nodes are pairwise linked, i.e. the graph is fully connected . The normalization constant, Z, (also known as the *partition function*) have been introduced explicitly in eq. (2.3) since there is no guarantee that a product of arbitrary functions is normalized and hence defines a probability distribution.

It is easily seen that the factorization eq. (2.3) implies that S_A is independent of S_B given S_C if the set of nodes C separates the nodes A from the nodes B, where by separation we mean naive graph-theoretic separation. A specific case of this is shown in figure 2.4. Let V denote the set of nodes associated with the undirected graph. Let $\tilde{A}$ denote the union of A and the nodes in $V \backslash C$ which are reachable from A and let $\tilde{B} = V \backslash (\tilde{A} \cup C)$. Clearly, every clique is composed of nodes from either $\tilde{A} \cup C$ or $\tilde{B} \cup C$, hence

$$p(\boldsymbol{s}) = p(s_{\tilde{A}}, s_{\tilde{B}}, s_C) = \frac{1}{Z} \Big(\prod_{c \in (\tilde{A} \cup C)} \phi_c \Big) \Big(\prod_{c \in (\tilde{B} \cup C)} \phi_c \Big) \tag{2.4}$$

$$= \frac{1}{Z} f(s_{\tilde{A}}, s_C) f(s_{\tilde{B}}, s_C) \ , \tag{2.5}$$

which explicitly shows that the joint distribution is a product of two functions; one function depending on s_A and the other depending on s_B. The conditional probability is then

$$p(s_{\tilde{A}}, s_{\tilde{B}} | s_C) = \frac{f(s_{\tilde{A}}, s_C) f(s_{\tilde{B}}, s_C)}{\langle f(s_{\tilde{A}}, s_C) \rangle_{1|s_C} \langle f(s_{\tilde{B}}, s_C) \rangle_{1|s_C}} \tag{2.6}$$

$$= \frac{p(s_{\tilde{A}}, s_C) p(s_{\tilde{B}}, s_C)}{\langle p(s_{\tilde{A}}, s_C) \rangle_{1|s_C} \langle p(s_{\tilde{B}}, s_C) \rangle_{1|s_C}} \tag{2.7}$$

$$= p(s_{\tilde{A}} | s_C) p(s_{\tilde{B}} | s_C) \ , \tag{2.8}$$

which shows that $S_{\tilde{A}} \perp S_{\tilde{B}} | S_C$, hence $S_A \perp S_B | S_C$. The property that S_A is independent of S_B given S_C if the set of nodes C separates the nodes A from the nodes B is often referred to as the *global Markov property*. That the factorization eq. (2.3) implies the global Markov property suggests that a product of potential functions is indeed the natural factorization of the joint distribution when considering undirected graphical models. Indeed, for strictly positive probability distributions the reversed implication, i.e. that the global Markov property implies the factorization eq. (2.3), can be obtained using the Hammersley-Clifford theorem.

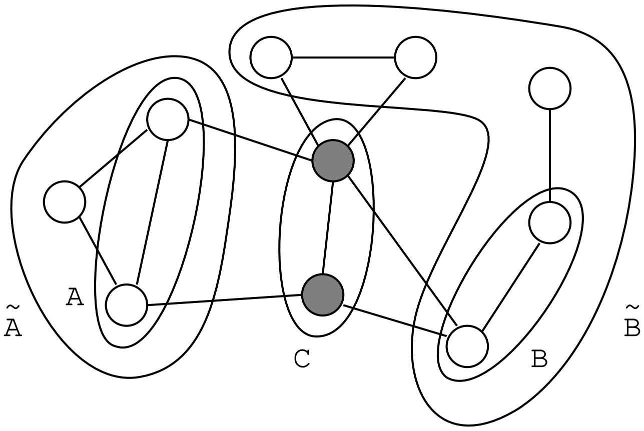


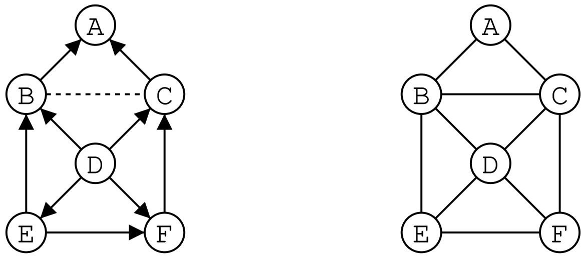
Fig. 2.4. The set C separates A from B since all paths from A to B have to pass through C. The same is seen to be true for the extended sets $\tilde{A}$ and $\tilde{B}$. Obviously is it sufficient to show that $\tilde{A}$ is conditional independent of $\tilde{B}$ given C.

The potential functions are commonly parameterized as $\phi_c = \exp(-E_c)$, where E_c is a unconstrained function. The joint probability distribution is then specified in term of the *Boltzmann distribution*

$$p(\boldsymbol{s}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{s})}, \tag{2.9}$$

where $E(\boldsymbol{s}) = \sum_{c \in \mathcal{C}} E_c$ is denoted the *energy function* and β is known as the inverse temperature. The inverse temperature is not directly part of the probabilistic model specification and it is clear that $\beta = 1$ in this case. Since the inverse temperature controls the smoothness of the joint distribution $p(\boldsymbol{s})$ it is, however, a useful quantity to have at hand when solving combinatorial optimization problems which have been casted onto undirected graphs, see e.g. simulated annealing [Kirkpatrick et al. 1983] and mean field annealing [Peterson and Söderberg 1989]).

In certain application domains it is more natural to state the generative model in terms of an undirected graphical model. Markov random fields have been widely used in image processing and computer vision ever since they were introduced in this context in the classic paper of [Geman and Geman 1984]. A nice treatment of various applications of MRFs to computer vision as well as an extensive list of references can be found in [Li 1995]. In section 2.4.1 we will as an example of a MRF consider the Boltzmann Machine which is an extension of the Hopfield network to include hidden units [Hertz et al. 1991]. In chapter 3, the Boltzmann Machine will due to its connection to statistical physics be our canonical example in the treatment of mean field methods for approximate inference. To see why and when such approx-



(a) (b)

Fig. 2.5. Shows the process of graph moralization. (a) Shows the original graph (in solid) as well as the edge added (in dashed) in the first step of "marrying" parents. (b) Shows the moralized graph.

imations are needed we have to consider the computational complexity in arbitrary graphical models. This is the subject of the next section.

2.3 Inference in Graphical Models

In this section we describe the general procedure for doing efficient inference in graphical models. As mentioned in section 2.2, this efficiency is achieved by exploiting the structure of the undirected graphical model. Often, however, the generative model has been specified in terms of a directed graphical model. Hence we need to consider how a directed graph can be recast into a undirected graph. This lead us to the concept of the graph moralization.

2.3.1 Graph Moralization

The joint distribution for the directed and undirected graphical model is given by respectively eq. (2.2) and eq. (2.3), i.e. in both cases a product of local functions on the graph. It is tempting just to drop the edge directions of the DAG and then identify the potential function $\phi_c(s_c)$ with the local conditional probability $p(s_i|\pi_i)$ for which the set of nodes $\{s_i\} \cup \pi_i$ is contained in c. This is, however, not in general a valid potential function since there is no guarantee that the set of nodes $\{s_i\} \cup \pi_i$ is contained in any clique, the reason being that some of the parents π_i might not be interconnected. The way to deal with this problem is to construct the *moral graph*, i.e. add undirected edges to all co-parents which are not currently joined and finally drop the directions of all remaining directed edges (see figure 2.5). Clearly, the probability distribution associated with the original DAG will still be a member of the family of distributions associated with the moral graph. The potential functions associated with clique c in the moral graph is then the product of all the local conditional probabilities $p(s_i|\pi_i)$ for which $\{s_i \cup \pi_i\}$ is contained in c. Obviously, we can now equivalently restate the Markov blanket $\boldsymbol{M}_i$ of node S_i as being the set of its neighbors in the moral graph.

2.3.2 The generalized potential representation

Since an arbitrary joint probability distribution cannot in general be expressed as a product of marginals, the factorization eq. (2.3) leaves us with little hope of equating the potential ϕ_c with the marginal $p(s_c)$. There is, however, an alternative way of expressing the joint distribution in which the potentials indeed *can* be identified with the marginals. As in section 2.2, consider a triple (A, B, C) of disjoint subsets of the vertex set V of an undirected graph $\mathcal{G}$, such that C separates A from B; this time, however, with the additional constraint that $\mathcal{G}_C$ is a complete subgraph of $\mathcal{G}_V$ and $V = (A \cup B \cup C)$. Such a triple is said to form a *decomposition* of the graph. Furthermore, we say that an undirected graph $\mathcal{G}$ is *decomposable* if either it is complete, or it possesses a proper decomposition (A, B, C) such that both subgraphs $\mathcal{G}_{A\cup C}$ and $\mathcal{G}_{B\cup C}$ are decomposable. A decomposition (A, B, C) is *proper* if both A and B are non-empty. Consider the graph decomposition $(\tilde{A}, \tilde{B}, C)$ shown in figure 2.4. Clearly, the joint distribution with respect to the graph can be expressed as in eq. (2.5), i.e.

$$p(\boldsymbol{s}) = p(s_{\tilde{A}}, s_{\tilde{B}}, s_C) = \frac{1}{Z} f(s_{\tilde{A}}, s_C) f(s_{\tilde{B}}, s_C) \ , \tag{2.10}$$

and by direct integration we find

$$p(s_{\tilde{A}}, s_C) = \frac{1}{Z} f(s_{\tilde{A}}, s_C) \tilde{f}(s_{\tilde{B}}, s_C) \ , \tag{2.11}$$

where $\tilde{f}(s_{\tilde{B}}, s_C) = \langle f(s_{\tilde{B}}, s_C) \rangle_{1|s_{\tilde{A}}, s_C}$. Note that the constrain that $\mathcal{G}_C$ is a complete subset of V makes sure that $\tilde{f}(s_{\tilde{B}}, s_C)$ can be represented in terms of complete subgraphs of $\mathcal{G}_C$. Similar results are obtained for the marginals, $p(s_{\tilde{B}}, s_C)$ and $p(s_C)$, which implies that $p(s_{\tilde{A}}, s_C)$ and $p(s_{\tilde{B}}, s_C)$ factorize with respect to the subgraphs $\mathcal{G}_{\tilde{A}\cup C}$ and $\mathcal{G}_{\tilde{B}\cup C}$ respectively, and that the joint distribution is given by

$$p(s_{\tilde{A}}, s_{\tilde{B}}, s_C) = \frac{p(s_{\tilde{A}}, s_C) p(s_{\tilde{B}}, s_C)}{p(s_C)} \ . \tag{2.12}$$

Conversely, let $p(s_{\tilde{A}}, s_C)$ and $p(s_{\tilde{B}}, s_C)$ factorize with respect to the subgraphs $\mathcal{G}_{\tilde{A}\cup C}$ and $\mathcal{G}_{\tilde{B}\cup C}$, respectively. Clearly, if $p(\boldsymbol{s})$ satisfies eq. (2.12) then it also factorize with respect to $\mathcal{G}$. If the graph $\mathcal{G}$ happens to be decomposable we can apply eq. (2.12) recursively. This shows that the joint distribution $p(\boldsymbol{s})$ associated to a decomposable graph can be expressed as

$$p(\boldsymbol{s}) = \frac{\prod_{c\in\mathcal{C}} \phi_c(s_c)}{\prod_{c'\in\mathcal{S}} \phi_{c'}(s_{c'})} \ , \tag{2.13}$$

where $\mathcal{C}$ is the set of cliques and $\mathcal{S}$ is the set of separators, i.e. the intersection of the adjacent cliques, in the decomposable graph $\mathcal{G}$. Equation (2.13) is

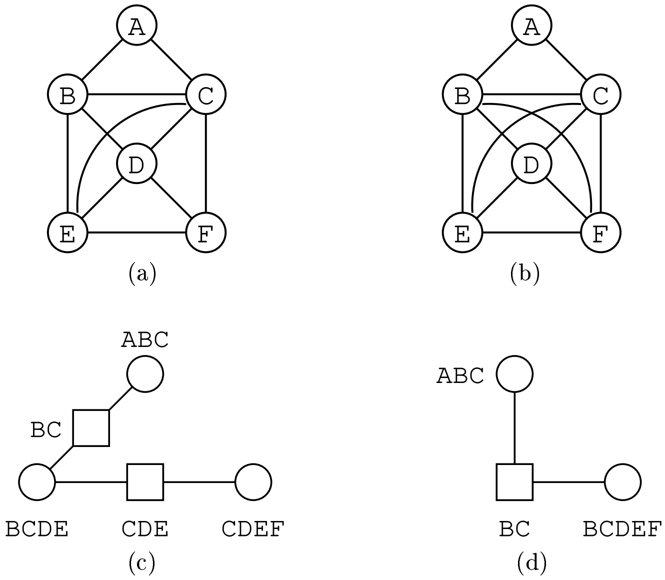


Fig. 2.6. Two triangulated graphs obtained from the DAG shown in figure 2.5 using node elimination ordering (a) $\{A, B, \cdot\}$ and (b) $\{A, D, \cdot\}$; (c,d) shows the junction tree associated with the triangulated graph (a) and (b), respectively.

known as the *generalized potential representation*. It can be shown that an undirected graph $\mathcal{G}$ is decomposable if and only if it is triangulated [Cowell et al. 1999]. An undirected graph is said to be *triangulated* (or chordal) if there are no cycles of length 4 or more distinct nodes without a short-cut. Obviously, a graph which is not triangulated can always be made so by adding extra edges in a suitable way. One such way is using a simple elimination procedure in which nodes are being successive removed from the graph; prior to removing a node S_i any non-connected neighbors of S_i are connected by adding fill-in edges. The union of S_i and its neighbors constitute the elimination clique corresponding to S_i. The edges of the triangulated graph are obtained as the union of the set of edges in the original (undirected) graph and the set of fill-in edges. The cliques in the triangulated graph can then be inferred from the elimination cliques. It is important to note that triangulation of a graph is not unique. Figure 2.6(a,b) shows two valid triangulations of the moral graph in figure 2.5(b), obtained using two different elimination orderings, namely $\{A, B, \cdot\}$ and $\{A, D, \cdot\}$. Figure 2.6(c,d) shows the junction tree corresponding to the two triangulated graphs. A tree of cliques is a *junction tree* if for each pair of cliques A, B, all cliques on the (unique) path between A and B contain the intersection $A \cap B$. Indeed, it can be shown that there exists a junction tree of cliques for the graph $\mathcal{G}$ if and only if $\mathcal{G}$ is decomposable [Cowell et al. 1999].

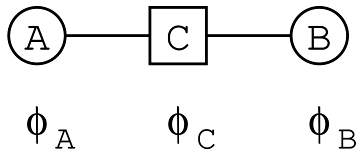


Fig. 2.7. Two cliques A and B and the separator $C = A \cap B$ between them. During the message passing scheme the clique and separator potentials are modified. In each message passing update the separator potential stores a copy of the message which is absorbed by the receiving clique. After a complete set of updates the two potentials have become consistent.

2.3.3 Message Passing on Junction Trees

Suppose there are only two cliques A and B in the triangulated graph; this e.g. turned out to be the case in figure 2.6(b,d). Now, let $C = A \cap B$ be the separator of A and B. The scenario is illustrated in figure 2.7. Since the graph is triangulated and hence decomposable then accordingly to eq. (2.13) the joint distribution can be written as

$$p(\boldsymbol{s}) = \frac{\phi_A(s_A)\phi_B(s_B)}{\phi_C(s_C)} \ . \tag{2.14}$$

However, from the specification of a probabilistic model we known the clique potentials ϕ_A and ϕ_B and furthermore we know that the joint distribution factorize as $p(\boldsymbol{s}) \propto \phi_A(s_A)\phi_B(s_B)$. Thus, a reasonable guess would be to initialize the separator potential to the partition function, $\phi_C = Z$. We can, however, without any loss of generality ignore the partition function and initialize $\phi_c = 1$. Since our objective is to interpret the potential functions as marginals we need a principled way of modifying the clique and separator potentials. Consider the following set of *marginal-propagation* updates [Jensen 1996].

$$\begin{aligned} \phi_C^* &= \langle \phi_A \rangle_{1|s_C} \ , \ \phi_B^* = (\phi_C^*/\phi_C)\phi_B \ , \ \phi_A^* = \phi_A \\ \phi_C^{**} &= \langle \phi_B^* \rangle_{1|s_C} \ , \ \phi_A^{**} = (\phi_C^{**}/\phi_C^*)\phi_A^* \ , \ \phi_B^{**} = \phi_B^* \end{aligned} \tag{2.15}$$

We can think of ϕ_C^* as a message passed from A to B and think of ϕ_C^{**} as a message passed from B to A. It is easily shown that each update (message passing) leaves the joint distribution invariant. After a complete set of updates the link between the two potentials have become consistent, that is $\langle \phi_A^{**} \rangle_{1|s_C} = \langle \phi_B^{**} \rangle_{1|s_C}$. This follows directly from eq. (2.15),

$$\langle \phi_A^{**} \rangle_{1|s_C} = \frac{\phi_C^{**}}{\phi_C^*} \langle \phi_A^* \rangle_{1|s_C} = \phi_C^{**} = \langle \phi_B^{**} \rangle_{1|s_C} \ . \tag{2.16}$$

Using the invariance of the joint distribution and eq. (2.16) we can now calculate the marginal

$$p(s_A) = \langle p(s_A, s_B, s_C) \rangle_{1|s_A} = \frac{\phi_A^{**}(s_A)}{\phi_C^{**}(s_C)} \langle \phi_B^{**}(s_B) \rangle_{1|s_C} = \phi_A^{**}(s_A) \ , \quad (2.17)$$

and similarly $p(s_B) = \phi_B$ and $p(s_C) = \phi_C$. This shows that inference can be seen as the process of achieving local consistency between the potential function of neighboring cliques. For arbitrary junction trees it is obvious that the junction tree property is a sufficient condition for local consistency to imply global consistency, hence the inference problem can be solved by a set of linked local computations in which links between neighboring cliques are being made consistent. These local computations can be carried out effectively using the following protocol.

Protocol 1 (Message passing scheme) [Jensen 1996] *A node A can send exactly one message to a neighboring clique B, and it may only be sent when A has received a message from each of its other neighboring cliques.*

The update, eq. (2.15), shows that the key to efficient inference algorithms is to form triangulated graphs which have small cliques, in terms of their state space. Thus, considering the example shown in figure 2.5 and 2.6 we could obtain the most efficient inference procedure by performing message passing on the junction tree shown in figure 2.6(c). This example shows that we need to find a good elimination ordering in order to obtain the triangulated cover which minimizes the sum of the state space sizes of the cliques. Unfortunately, it turns out that this is in general a NP-hard problem. Hence, various heuristics have been developed for triangulating non-chordal graphs. One such heuristic which is due to [Kjærulff 1990] is restated in the excellent procedural guide to inference in graphical models by [Huang and Darwiche 1996]. During an elimination procedure the heuristic suggests choosing the node that causes the least number of fill-in edges to be added, breaking ties by choosing the node that induces the cliques with the smallest state space. It turns out that this greedy heuristic produces reasonable triangulations in real world settings [Huang and Darwiche 1996].

In many applications we are interested in calculating the most probable configuration of all the dynamical variables instead of the entire marginals distributions. This calculation can be carried out computationally efficiently, merely by substituting the marginalizations by max-operators in the message passing updates eq. (2.15). This time, after a complete set of *max-propagation* updates the cliques have become max-consistent, i.e. for every pair of neighboring cliques A and B separated by C we have $\max_{A\setminus C} \phi_A^{**} = \max_{B\setminus C} \phi_B^{**}$ [Jensen 1996]. One highly celebrated special case of max-propagation is the *Viterbi algorithm* [Rabiner 1989] which is widely used in speech signal processing to find the MAP hidden sequence in Hidden Markov Models.

2.4 Two canonical models

Before proceeding to the problem of learning in graphical models let us consider two canonical graphical models; the first being undirected and the second being directed.

2.4.1 The Boltzmann machine

The Boltzmann machine (BM) [Hertz et al. 1991] is an undirected graphical model defined on a set $\boldsymbol{S} = \{\boldsymbol{V} \cup \boldsymbol{H}\}$ of N binary random variables $S_i = \{-1, 1\}$ where $\boldsymbol{V}$ and $\boldsymbol{H}$ is the set of visible and hidden units, respectively. The (Hopfield) energy function of a BM is at most quadratic, i.e. in general of the form

$$E(\boldsymbol{s}) = -\frac{1}{2}\sum_{i,j=1}^{N} J_{ij} s_i s_j - \sum_{i=1}^{N} \theta_i s_i = -\frac{1}{2}\sum_{i,j=0}^{N} J_{ij} s_i s_j \ , \tag{2.18}$$

where the interactions-weights are chosen symmetric such that $J_{ij} = J_{ji}$. The last equality is obtained by introducing an extra node S_0, which is clamped to state $+1$ and define J_{i0} to be equal to the threshold θ_i. Since the self-interaction weights, $J_{ii}(i > 0)$, leave the probability distribution invariant we can without loss of generality let $J_{ii} = 0$ for $i > 0$. When referring to the BM we will therefore, unless otherwise mentioned, be using the following parameterization

$$E(\boldsymbol{s}) = -\frac{1}{2}\sum_{i,j}^{N} J_{ij} s_i s_j - \sum_{i=1}^{N} \theta_i s_i \ , \tag{2.19}$$

where $J_{ii} = 0$. This undirected graphical model has a physical analog in the *Ising model* which is a model of magnetic systems. In the Ising model node, S_i represents the orientation of the spin at lattice site i; the spin is oriented "up" if $S_i = 1$ and "down" if $S_i = -1$.

2.4.2 A hidden state space model

In spite of their simple factorization, the graphical models considered in this section are able to account for a large class unsupervised models which have been proposed through time for modeling multidimensional data. All the generative models factorize according to the DAG shown in figure 2.8 which also shows one of the possible junction trees associated to this DAG. Hence, with this factorization we assume that the i'th observable Y_i is generated from the hidden state X_i which evolves according to a simple first-order Markov dynamics. The particular case where both the hidden state and observation are obtained as a linear mapping of the conditioning state corrupted with Gaussian noise is considered in [Roweis and Ghahramani 1999]. Starting from this

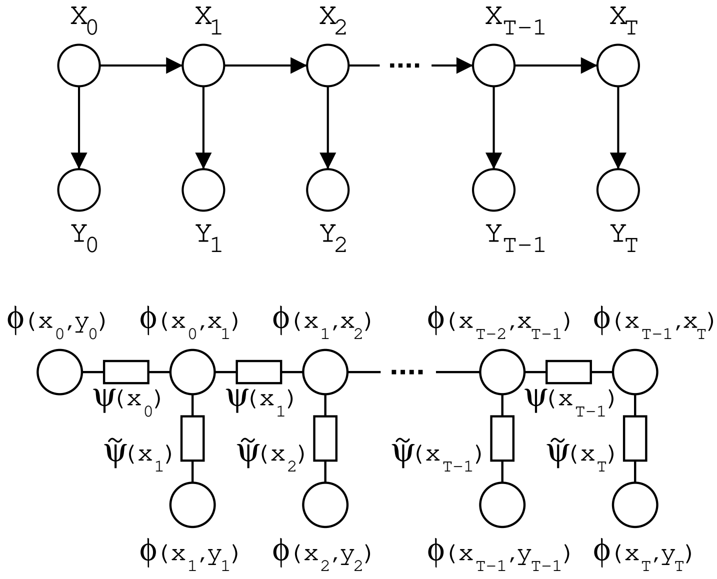


Fig. 2.8. Shows the DAG for the hidden state space model and one of the possible junction trees associated to it. The observation Y_i is conditional independent given the hidden state X_i which evolves according to a first order hidden Markov dynamic.

generative model they recover various well known statistical models as e.g. *factor analysis*, *principal components analysis* (PCA) and *Kalman filter models*. Furthermore, by applying additional non-linear mappings it is possible to obtain *vector quantization*, *hidden Markov models* (HMMs) and the generative model for *independent component analysis* (ICA). In the next section we follow [Jordan and Bishop 2001] and [Murphy 1998] and illustrate the message passing scheme on the Gaussian linear state space model or linear dynamical system. In chapter 6 we consider the same DAG in the context of a on-line classification model where approximate message passing is needed.

2.4.3 The Gaussian linear state space model

In this section we consider the Gaussian linear model which gives rise to the classical Kalman filter and Rauch-Tung-Striebel (RTS) recursions [Roweis and Ghahramani 1999]. This model which have been extensively investigated by the engineering and control communities for decades takes the form

$$\begin{aligned}
&\boldsymbol{X}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) \\
&\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\eta}_t \quad , \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta) \\
&\boldsymbol{Y}_t = \boldsymbol{B}\boldsymbol{X}_t + \varepsilon_t \quad , \qquad \varepsilon_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon) \ ,
\end{aligned} \tag{2.20}$$

where $\boldsymbol{A}$ is the state transition matrix, $\boldsymbol{B}$ is the observation matrix and $\{\boldsymbol{\eta}_t\}$ and $\{\boldsymbol{\varepsilon}_t\}$ are mutually independent white Gaussian noise sequences with covariance $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{\Sigma}_\varepsilon$, respectively. In the following derivation of the Kalman filter and RTS recursions we will be using the canonical parameterization of the Gaussian density, see appendix A. Hence, for this model the local conditional probabilities associated to the DAG shown in figure 2.8 are given by

$$
\begin{aligned}
p(\boldsymbol{x}_0) &\propto e^{-\frac{1}{2}(\boldsymbol{x}_0-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_0)} \propto e^{-\frac{1}{2}\boldsymbol{x}_0^T\boldsymbol{\Lambda}_0\boldsymbol{x}_0+(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0)^T\boldsymbol{x}_0} \\
p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) &\propto e^{-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{A}\boldsymbol{x}_{t-1})^T\boldsymbol{\Sigma}_\eta^{-1}(\boldsymbol{x}_t-\boldsymbol{A}\boldsymbol{x}_{t-1})} \\
&\propto e^{-\frac{1}{2}\begin{bmatrix}\boldsymbol{x}_{t-1}\\ \boldsymbol{x}_t\end{bmatrix}^T\begin{bmatrix}\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\\ -\boldsymbol{\Lambda}_\eta\boldsymbol{A} & \boldsymbol{\Lambda}_\eta\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_{t-1}\\ \boldsymbol{x}_t\end{bmatrix}} \\
p(\boldsymbol{y}_t|\boldsymbol{x}_t) &\propto e^{-\frac{1}{2}(\boldsymbol{y}_t-\boldsymbol{B}\boldsymbol{x}_t)^T\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}_t-\boldsymbol{B}\boldsymbol{x}_t)} \\
&\propto e^{-\frac{1}{2}\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{y}_t\end{bmatrix}^T\begin{bmatrix}\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} & -\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\\ -\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} & \boldsymbol{\Lambda}_\varepsilon\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{y}_t\end{bmatrix}} .
\end{aligned}
\tag{2.21}
$$

To solve the inference problem using the message passing scheme, every clique has to exchange exactly one bidirectional message with each of its neighbors in such a way that the message passing protocol is upheld. We achieve this by assigning one of the cliques in the junction tree as root and then carry out two sweeps each following the message passing protocol. In the first (forward) sweep every clique except the root sends a message towards the root node and in the second (backward) sweep messages are distributed from the root clique to all other cliques. Following [Murphy 1998; Jordan and Bishop 2001] we assign the clique $\{X_{T-1}, X_T\}$ to be the root clique. The separator potentials are initialized, i.e. $\psi = \tilde{\psi} = 1$, and the clique potentials ϕ are easily identified as

$$
\begin{aligned}
\phi(\boldsymbol{x}_0, \boldsymbol{y}_0) = p(\boldsymbol{x}_0)p(\boldsymbol{y}_0|\boldsymbol{x}_0) &= \mathcal{N}_*\left(\begin{bmatrix}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}_0+\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} & -\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\\ -\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} & \boldsymbol{\Lambda}_\varepsilon\end{bmatrix}\right) \\
\phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) &= \mathcal{N}_*\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\\ -\boldsymbol{\Lambda}_\eta\boldsymbol{A} & \boldsymbol{\Lambda}_\eta\end{bmatrix}\right) \\
\phi(\boldsymbol{x}_t, \boldsymbol{y}_t) = p(\boldsymbol{y}_t|\boldsymbol{x}_t) &= \mathcal{N}_*\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} & -\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\\ -\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} & \boldsymbol{\Lambda}_\varepsilon\end{bmatrix}\right)
\end{aligned}
\tag{2.22}
$$

Since the result of unconditional inference follows directly from the theory, we will now consider the case there evidence is present at $\{\boldsymbol{Y}_i\} = \{\dot{\boldsymbol{y}}_i\}$. Here the dot is used to emphasize instantiated random variables. We start by having a look at the forward sweep in which messages are being send towards the root. According to the message passing protocol the potential $\phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ has to wait sending its message toward the root until it has absorbed the messages stored in the updated separator potentials $\psi^*(\boldsymbol{x}_t)$ and $\tilde{\psi}^*(\boldsymbol{x}_{t+1})$.

Provided that the separator potential $\psi(\boldsymbol{x}_t)$ has already been updated, i.e. we have $\boldsymbol{\gamma}_{t|t}$ and $\boldsymbol{\Lambda}_{t|t}$ such that

$$\psi^*(\boldsymbol{x}_t) \propto e^{-\frac{1}{2}\boldsymbol{x}_t^T \boldsymbol{\Lambda}_{t|t}\boldsymbol{x}_t + \boldsymbol{\gamma}_{t|t}^T \boldsymbol{x}_t} \ , \tag{2.23}$$

we can easily let the potential $\phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ absorb the message stored in $\psi^*(\boldsymbol{x}_t)$,

$$\begin{aligned} \phi^\circ(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) &= \psi^*(\boldsymbol{x}_t)\phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) & (2.24)\\ &\propto \mathcal{N}_*\left(\begin{bmatrix}\boldsymbol{\gamma}_{t|t}\\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{K}_{t|t} & -\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\\ -\boldsymbol{\Lambda}_\eta \boldsymbol{A} & \boldsymbol{\Lambda}_\eta\end{bmatrix}\right) \ , & (2.25)\end{aligned}$$

where

$$\boldsymbol{K}_{t|t} = \boldsymbol{A}^T\boldsymbol{\Lambda}_\eta \boldsymbol{A} + \boldsymbol{\Lambda}_{t|t} \ . \tag{2.26}$$

The superscript $\circ$ is used to emphasize that the clique potential has only been partially updated. Since $\phi^\circ(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ is proportional to $p(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}|\dot{\boldsymbol{y}}_1^t)$ we can at this point find the one-step-ahead prediction or time update by marginalizing $\boldsymbol{x}_t$,

$$p(\boldsymbol{x}_{t+1}|\dot{\boldsymbol{y}}_1^t) = \mathcal{N}_*(\boldsymbol{\gamma}_{t+1|t}, \boldsymbol{\Lambda}_{t+1|t}) \ , \tag{2.27}$$

where

$$\begin{aligned} \boldsymbol{\gamma}_{t+1|t} &= \boldsymbol{\Lambda}_\eta \boldsymbol{A}\boldsymbol{K}_{t|t}^{-1}\boldsymbol{\gamma}_{t|t} & (2.28)\\ \boldsymbol{\Lambda}_{t+1|t} &= \boldsymbol{\Lambda}_\eta - \boldsymbol{\Lambda}_\eta \boldsymbol{A}\boldsymbol{K}_{t|t}^{-1}\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta \ . & (2.29)\end{aligned}$$

In order to complete the update of the clique potentials $\phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ we first need to update the evidence dependent separator potentials which is easily done

$$\begin{aligned} \tilde{\psi}^*(\boldsymbol{x}_t) &= \langle \delta(\boldsymbol{y}_t - \dot{\boldsymbol{y}}_t)\phi(\boldsymbol{x}_t, \boldsymbol{y}_t)\rangle_{1|\boldsymbol{x}_t} & (2.30)\\ &\propto e^{-\frac{1}{2}\boldsymbol{x}_t^T\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon \boldsymbol{B}\boldsymbol{x}_t + (\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon \dot{\boldsymbol{y}}_t)^T\boldsymbol{x}_t} \ , & (2.31)\end{aligned}$$

which in turn yields the updated clique potential

$$\begin{aligned} \phi^*(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) &= \tilde{\psi}^*(\boldsymbol{x}_{t+1})\phi^\circ(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) & (2.32)\\ &\propto \mathcal{N}_*\left(\begin{bmatrix}\boldsymbol{\gamma}_{t|t}\\ \boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon \dot{\boldsymbol{y}}_{t+1}\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}_{t|t} & -\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\\ -\boldsymbol{\Lambda}_\eta \boldsymbol{A} & \boldsymbol{\Lambda}_\eta + \boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon \boldsymbol{B}\end{bmatrix}\right) \ . & (2.33)\end{aligned}$$

Finally, we obtain $\psi^*(\boldsymbol{x}_{t+1})$ by marginalizing the updated clique potential with respect to $\boldsymbol{x}_t$ which yields the following datum or measurement updates

$$\begin{aligned} \boldsymbol{\gamma}_{t+1|t+1} &= \boldsymbol{\gamma}_{t+1|t} + \boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon \dot{\boldsymbol{y}}_{t+1} & (2.34)\\ \boldsymbol{\Lambda}_{t+1|t+1} &= \boldsymbol{\Lambda}_{t+1|t} + \boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon \boldsymbol{B} \ . & (2.35)\end{aligned}$$

To initialize the forward sweep in which messages are being send toward the root we need to know the boundary condition for the recursions. This is readily obtained

$$\psi^*(\boldsymbol{x}_0) = \langle \delta(\boldsymbol{y}_0 - \dot{\boldsymbol{y}}_0)\phi(\boldsymbol{x}_0, \boldsymbol{y}_0)\rangle_{1|\boldsymbol{x}_0} \tag{2.36}$$

$$\propto e^{-\frac{1}{2}\begin{bmatrix}\boldsymbol{x}_0\\ \dot{\boldsymbol{y}}_0\end{bmatrix}^T \begin{bmatrix}\boldsymbol{\Lambda}_0 + \boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} & -\boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\\ -\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} & \boldsymbol{\Lambda}_\varepsilon\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_0\\ \dot{\boldsymbol{y}}_0\end{bmatrix} + (\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0)^T\boldsymbol{x}_0} \tag{2.37}$$

$$\propto \mathcal{N}_*\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \boldsymbol{B}^T\boldsymbol{\Lambda}_\epsilon\dot{\boldsymbol{y}}_0, \boldsymbol{\Lambda}_0 + \boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B}\right) \ . \tag{2.38}$$

Equations (2.26), (2.28), (2.29), (2.34) and (2.35) together with the boundary condition eq. (2.38) constitute the information filter equations which are algebraically equivalent to the usual Kalman filter recursions [Anderson and Moore 1979].

We now turn to the backward sweep in which messages are being distributed from the root clique to all other cliques. The derivations of the backward recursions follows along the same lines as the ones leading to the forward recursions, i.e. we start by assuming that the separator potential $\psi^*(\boldsymbol{x}_{t+1})$ has already been updated

$$\psi^{**}(\boldsymbol{x}_{t+1}) \propto e^{-\frac{1}{2}\boldsymbol{x}_{t+1}^T\boldsymbol{\Lambda}_{t+1|T}\boldsymbol{x}_{t+1} + \boldsymbol{\gamma}_{t+1|T}^T\boldsymbol{x}_{t+1}} \ , \tag{2.39}$$

which is then used to update the clique potential

$$\begin{aligned}\phi^{**}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) &= \frac{\psi^{**}(\boldsymbol{x}_{t+1})}{\psi^*(\boldsymbol{x}_{t+1})}\phi^*(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})\\ &\propto \mathcal{N}_*\left(\begin{bmatrix}\boldsymbol{\gamma}_{t|t}\\ \boldsymbol{\xi}_{t+1|T}\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}_{t|t} & -\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\\ -\boldsymbol{\Lambda}_\eta\boldsymbol{A} & \boldsymbol{L}_{t+1|T}\end{bmatrix}\right) \ ,\end{aligned} \tag{2.40}$$

where we have introduced

$$\boldsymbol{\xi}_{t+1|T} = \boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\dot{\boldsymbol{y}}_{t+1} + \boldsymbol{\gamma}_{t+1|T} - \boldsymbol{\gamma}_{t+1|t+1} = \boldsymbol{\gamma}_{t+1|T} - \boldsymbol{\gamma}_{t+1|t} \tag{2.41}$$

$$\boldsymbol{L}_{t+1|T} = \boldsymbol{\Lambda}_\eta + \boldsymbol{B}^T\boldsymbol{\Lambda}_\varepsilon\boldsymbol{B} + \boldsymbol{\Lambda}_{t+1|T} - \boldsymbol{\Lambda}_{t+1|t+1} \ . \tag{2.42}$$

Given the updated clique potential $\phi^{**}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ we may now update the next separator potential along the chain, i.e. calculate $\psi^{**}(\boldsymbol{x}_t) = \langle\phi^{**}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})\rangle_{1|\boldsymbol{x}_t}$ which yields the following recursions

$$\boldsymbol{\gamma}_{t|T} = \boldsymbol{\gamma}_{t|t} + \boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\boldsymbol{L}_{t+1|T}^{-1}(\boldsymbol{\gamma}_{t+1|T} - \boldsymbol{\gamma}_{t+1|t}) \tag{2.43}$$

$$\boldsymbol{\Lambda}_{t|T} = \boldsymbol{K}_{t|t} - \boldsymbol{A}^T\boldsymbol{\Lambda}_\eta\boldsymbol{L}_{t+1|T}^{-1}\boldsymbol{\Lambda}_\eta\boldsymbol{A} \ . \tag{2.44}$$

Similarly, we update the last set of separator potential, i.e.

$$\tilde{\psi}^{**}(\boldsymbol{x}_{t+1}) = \langle\phi^{**}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})\rangle_{1|\boldsymbol{x}_{t+1}} \propto e^{-\frac{1}{2}\boldsymbol{x}_{t+1}^T\tilde{\boldsymbol{\Lambda}}_{t+1|T}\boldsymbol{x}_{t+1} + \tilde{\boldsymbol{\gamma}}_{t+1|T}^T\boldsymbol{x}_{t+1}}, \tag{2.45}$$

where

$$\tilde{\boldsymbol{\gamma}}_{t+1|T} = \boldsymbol{\xi}_{t+1|T} + \boldsymbol{\Lambda}_\eta\boldsymbol{A}\boldsymbol{K}_{t|t}^{-1}\boldsymbol{\gamma}_{t|t} \tag{2.46}$$

$$\tilde{\boldsymbol{\Lambda}}_{t+1|T} = \boldsymbol{L}_{t+1|T} - \boldsymbol{\Lambda}_\eta\boldsymbol{A}\boldsymbol{K}_{t|t}^{-1}\boldsymbol{A}^T\boldsymbol{\Lambda}_\eta \ . \tag{2.47}$$

The base case $\psi^{**}(x_{T-1})$ for the backward recursion of ψ are found straightforwardly by the marginalizing the root potential $\phi^{**}(x_{T-1}, x_T)$ with respect to x_t.

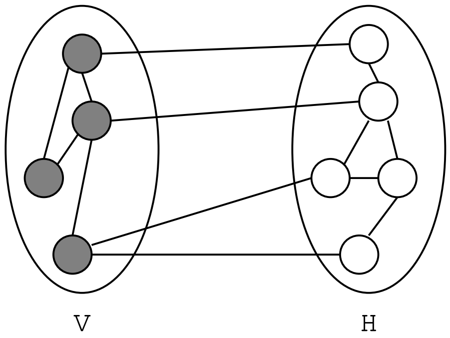


Fig. 2.9. Illustration of the generic learning problem for graphical models. Given a set of visible nodes $\boldsymbol{V}$, we want to estimate the model parameters (not shown in the figure) in the presence of a set of hidden nodes $\boldsymbol{H}$. Hence, to carry out ML learning we have to be able to calculate the posterior density of the hidden variables.

2.5 Learning in Graphical Models

This section describes two different ways for learning the parameters in graphical models in the maximum likelihood framework, i.e. to choose the model parameters maximizing the marginal probability density of a given set of visible (evidence) nodes. The general situation is depicted in figure 2.9 in which the node set $\boldsymbol{S}$ of the graphical model has been partitioned into the set of visible nodes $\boldsymbol{V}$ and the set of hidden nodes $\boldsymbol{H}$.

2.5.1 Boltzmann Learning

The Boltzmann learning rule [Hertz et al. 1991] was introduced by Hinton and Sejnowski as a procedure for doing supervised learning of Boltzmann machines. However, we will here consider a general undirected graphical model with node set $\boldsymbol{S} = \{\boldsymbol{V}, \boldsymbol{H}\}$, where $\boldsymbol{V}$ is the set of visible nodes and $\boldsymbol{H}$ is the set of hidden nodes. Let $E_{\boldsymbol{\theta}}$ denote the energy function associated to the undirected graphical model with parameters $\boldsymbol{\theta} = \{\theta_i\}$. The joint probability distribution then takes the form

$$p(\boldsymbol{s}|\boldsymbol{\theta}) = \frac{1}{Z} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{s})} , \tag{2.48}$$

where $Z = \left\langle e^{-E_{\boldsymbol{\theta}}(\boldsymbol{s})} \right\rangle_1$ is the partition function associated to the joint distribution. To do maximum likelihood parameter estimation we need to calculate the (log) probability of the visible nodes given the model parameters, i.e. marginalize out the hidden nodes

$$\log p(\boldsymbol{v}|\boldsymbol{\theta}) = \log \left\langle \frac{1}{Z} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})} \right\rangle_{1|\boldsymbol{v}} = \log \frac{Z_{\boldsymbol{v}}}{Z} , \tag{2.49}$$

where we have introduced the partition function $Z_{\boldsymbol{v}} = \left\langle e^{-E_{\boldsymbol{\theta}}(\boldsymbol{s})} \right\rangle_{1|\boldsymbol{v}}$ associated to the posterior

$$p(\boldsymbol{h}|\boldsymbol{\theta}, \boldsymbol{v}) = \frac{1}{Z_{\boldsymbol{v}}} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{h},\boldsymbol{v})} \ . \tag{2.50}$$

This clearly shows that we have to be able to make inference (about hidden states) in order to carry out learning. Using the simple relation

$$\frac{\partial \log Z}{\partial \theta_i} = \frac{1}{Z} \frac{\partial}{\partial \theta_i} \left\langle e^{-E_{\boldsymbol{\theta}}(\boldsymbol{s})} \right\rangle_1 = - \left\langle \frac{\partial E_{\boldsymbol{\theta}}(\boldsymbol{s})}{\partial \theta_i} \right\rangle_p \ , \tag{2.51}$$

the Boltzmann learning rule is readily obtained as the gradient ascent on the likelihood, i.e.

$$\Delta\theta_i = \eta \left\{ \left\langle -\frac{\partial E_{\boldsymbol{\theta}}}{\partial \theta_i} \right\rangle_{p|\boldsymbol{v}} - \left\langle -\frac{\partial E_{\boldsymbol{\theta}}}{\partial \theta_i} \right\rangle_p \right\} \ , \tag{2.52}$$

where η is the learning rate. Equation (2.52) shows that the learning process consists of an unlearning (or student) component, $\langle \cdot \rangle_p$, in which the parameters are free to move and a learning (or teacher) component, $\langle \cdot \rangle_{p|\boldsymbol{v}}$, in which the visible nodes are clamped. As an example consider applying the Boltzmann learning rule (2.52) on the Boltzmann machine eq. (2.19). It is seen that this requires the calculation of correlations between nodes. This is, however, computational infeasible if the system contains large cliques. For such systems we need to consider approximating methods for calculating correlations which will be the subject of chapter 3.

2.5.2 Learning by Expectation Maximization (EM)

The EM algorithm was proposed by [Dempster et al. 1977] as an iterative approach for doing maximum likelihood estimation in the presence of hidden variables. Again we consider a graphical model with node set $\boldsymbol{S} = \{\boldsymbol{V}, \boldsymbol{H}\}$, where $\boldsymbol{V}$ is the set of visible nodes and $\boldsymbol{H}$ is the set of hidden nodes. The joint probability density associated with the model is $p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\theta_i\}$ is the collection of model parameters. Like for Boltzmann learning the objective of the EM algorithm is to maximize the likelihood, i.e.

$$L(\boldsymbol{\theta}) = \log p(\boldsymbol{v}|\boldsymbol{\theta}) = \log \langle p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta}) \rangle_{1|\boldsymbol{v}} \ . \tag{2.53}$$

A lower bound on the likelihood can be obtained using *Jensens inequality*[1],

$$L(\boldsymbol{\theta}) = \log \left\langle \frac{p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{q(\boldsymbol{h})} \right\rangle_q \geq \left\langle \log \frac{p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{q(\boldsymbol{h})} \right\rangle_q = L^*(q, \boldsymbol{\theta}) \ , \tag{2.54}$$

[1] Jensens inequality states $\langle \phi(X) \rangle \geq \phi(\langle X \rangle)$, where ϕ is a convex function. [Cover and Thomas 1991]

where $q(\boldsymbol{h})$ is an arbitrary probability density over the hidden variables. It can now be shown that the EM algorithm essentially performs gradient ascent on the lower bound, L^* [Neal and Hinton 1998]. Noting that the likelihood is independent of the hidden variables $\boldsymbol{H}$, the slack Δ between the likelihood and the lower bound is readily obtained

$$\Delta = L(\boldsymbol{\theta}) - L^*(q, \boldsymbol{\theta}) = \left\langle \log \frac{q(\boldsymbol{h})p(\boldsymbol{v}|\boldsymbol{\theta})}{p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})} \right\rangle_q = KL(q \,\|\, p_{\boldsymbol{H}}) \ , \tag{2.55}$$

where $p_{\boldsymbol{H}} = p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{\theta})$ is the posterior probability of the hidden states $\boldsymbol{H}$ and $KL(q \,\|\, p) = \langle \log q/p \rangle_q$ is the *Kullback-Leibler (KL) distance* (or divergence) between q and p. The EM algorithm alternates between maximizing L^* with respect to the distribution, q (E-step), and the parameters $\boldsymbol{\theta}$ (M-step), respectively, keeping the other fixed. In the E-step the KL distance is minimized when q is equal to the posterior, $p_{\boldsymbol{H}}^{(k)} = p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{\theta}^{(k)})$ of the hidden nodes, eq. (2.55). Alternatively, this can be seen by performing free-form optimization of q on the lower bound, L^*. Since q does not depend on $\boldsymbol{\theta}$ the M-step amounts to maximizing the expected *complete likelihood*, $p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})$ with respect to $p_{\boldsymbol{H}}^{(k)}$,

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \langle \log p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta}) \rangle_{p_{\boldsymbol{H}}^{(k)}} \ . \tag{2.56}$$

Since the bound is tight at the beginning of each M-step and furthermore, the E-step does not change $\boldsymbol{\theta}$, the combined EM step is guaranteed to not decrease the likelihood after each combined EM step, i.e.

$$L(\boldsymbol{\theta}^{(k-1)}) = L^*(q^{(k)}, \boldsymbol{\theta}^{(k-1)}) \leq L^*(q^{(k)}, \boldsymbol{\theta}^{(k)}) \leq L(\boldsymbol{\theta}^{(k)}) \ , \tag{2.57}$$

where the first equality follows after having performed the E-step and the first inequality comes from the M-step and the last inequality follows from the lower bound eq. (2.54). When the EM algorithm has converged to a fixed-point $\boldsymbol{\theta}^*$ we know that $\boldsymbol{\theta}^*$ is a maximum for $L^*(q, \boldsymbol{\theta}^*)$ and that L and L^* are equal at $\boldsymbol{\theta}^*$. Assuming L and L^* are both differentiable this implies that $\boldsymbol{\theta}^*$ is a stationary point (not necessarily a local maximum) of L. In practice, however, convergence to saddle points or local minima in the likelihood is rarely seen. The big advantage of using EM is that we can make explicit use of the factorization of the complete log likelihood and hence decouple the estimation problem. All we need to calculate is the *expected sufficient statistics* (with respect to the posterior of the hidden variables) and solve a maximization problem. As mentioned already in section 2.3.3, it is in general not possible to solve the inference problem exactly. Likewise, it might be the case that the maximization problem has no analytically solution. However, since both the expectation step and maximization step in turn maximizes the same lower bound on the likelihood, we are allowed to make only partial E-steps and M-steps without loosing the monotonic increase in the likelihood, eq. (2.57).

In this case the EM algorithm is called a *generalized* EM (GEM) algorithm. E.g. instead of using the true posterior in the E-step we could consider a family of tractable distributions and pick the member minimizing Δ. Similarly, we are allowed to use our favorite numerical optimizer to either partially or completely solve the maximization step. It should be noted, however, that by using an approximation to the true posterior we are no longer guaranteed to get the same estimates as ML-estimation. This is due to the fact that it may be impossible to achieve $L^* = L$ within the chosen family of approximating distributions. Hopefully, we have chosen the family of approximating distributions large enough to be close (in the KL sense) to the true posterior. In such case the GEM algorithm would yield reasonable ML estimates.

3. Mean Field Approximations

The previous chapter showed that our success of making probabilistic learning and inference in dense grahical models depends on the size of the largest clique in the triangulated graph since we potentially need to sum over all posible configurations of the clique variables. This chapter reviews some of the various types of mean field (MF) methods that have been proposed in context of statistical physics for computing the partition function which is the most intractable sum we need to consider for probabilistic models. Common for the mean field approaches are that they only are strictly valid in the limit of infinite numbers of degrees of freedom. This makes MF approaches particularly well suited for analyzing systems behavior in the thermodynamic limit ($N \to \infty$) as well as solving combinatorial optimization problems ($\beta \to \infty$) which have been casted onto graphical models. However, mean field approximations may still be a valid approximation for finite system size since dense graphs can be probabilistically simple e.g. averaging phenomena can make nodes relatively insensitive to the particular configuration of its neighboring nodes. In this chapter we let without loss of generality $\beta = 1$ to ease the notation.

3.1 The saddle-point approximation

The saddle-point approximation[1] (also known as the method of steepest descent) was proposed by [Peterson and Anderson 1987] as a method to overcome the computational intensive task of learning of BMs and later proposed as a method for solving combinatorial optimization problems [Peterson and Söderberg 1989]. The main idea of the saddle-point approach is to replace a sum over discrete variables by a integral over a set of auxiliary variable. To be specific, given a system of N dynamical variables with energy function, $E(\boldsymbol{s})$, we are interested in an approximation to the partition function

[1] The term saddle-point approximation originates from the fact that the real and imaginary parts of an analytic function, $f(z) = u(x, y) + jv(x, y)$, where $z = x + jy$, must satisfy the Cauchy-Riemann equations, that is $\partial u/\partial x = \partial v/\partial y$ and $\partial u/\partial y = -\partial v/\partial x$.

$$Z = \int d\boldsymbol{s} e^{-E(\boldsymbol{s})} \; . \tag{3.1}$$

By inspired hindsight we define the N-dimensional Dirac delta function as

$$\delta(\boldsymbol{x}) = \left(\frac{1}{2\pi}\right)^N \int d\boldsymbol{h} e^{j\boldsymbol{h}^T\boldsymbol{x}} \; , \tag{3.2}$$

whereby the partition function can be calculated as

$$Z = \int d\boldsymbol{m} d\boldsymbol{s} \delta(\boldsymbol{s}-\boldsymbol{m}) e^{-E(\boldsymbol{m})} \propto \int d\boldsymbol{h} d\boldsymbol{m} e^{-E_e(\boldsymbol{m},\boldsymbol{h})} \; , \tag{3.3}$$

where we have defined the effective (complex) energy function

$$E_e(\boldsymbol{m},\boldsymbol{h}) = E(\boldsymbol{m}) + (j\boldsymbol{h}^T\boldsymbol{m} - \sum \log \int ds_i e^{jh_i s_i}) \; . \tag{3.4}$$

Since the introduction of the auxiliary variables have rendered the dynamical variables, $\boldsymbol{S}$, independent, the sum over the state space has now become tractable. Assuming that E_e is an analytic function the integral can be approximated using saddle-point integration [Marsden and Hoffman 1987],

$$Z \propto e^{-\beta E_e(\boldsymbol{m}^*,\boldsymbol{h}^*)} \tag{3.5}$$

where the saddle-points $(\boldsymbol{m}^*, \boldsymbol{h}^*)$ are given by the *mean field equations*

$$\frac{\partial E_e}{\partial h_i} = 0 \Rightarrow m_i^* = \frac{\int ds_i s_i e^{h_i s_i}}{\int ds_i e^{h_i s_i}} \quad , \quad \frac{\partial E_e}{\partial m_i} = 0 \Rightarrow h_i^* = -\frac{\partial E}{\partial m_i} \; . \tag{3.6}$$

For the moment being we have to think of the fixed-point solution, $\{m_i^*\}$, as a set of order parameters i.e. combinations of dynamical variables that do not average to zero for any value of control parameters, in the thermodynamical limit. To circumvent this problem, a modification to the described saddle-point approximation has been proposed by [Bhattacharyya and Keerthi 1999] which relies on the generalized steepest descent theorem [Marsden and Hoffman 1987]. The generalized steepest descent theorem yields the asymptotic expansion

$$\int_\gamma d\boldsymbol{\zeta} f(\boldsymbol{\zeta}) e^{\beta E_e(\boldsymbol{\zeta})} \propto f(\boldsymbol{\zeta}_0) e^{\beta E_e(\boldsymbol{\zeta}_0)} \; , \tag{3.7}$$

where $f(\boldsymbol{\zeta})$ is a bounded continuous function on the path of integration $\boldsymbol{\gamma}$ and $\boldsymbol{\zeta}_0$ is the fixed-point solution $E_e'(\boldsymbol{\zeta}_0) = 0$. Since the factor of proportionality is independent of f we can apply saddle-point integration to both the nominator and denominator of $\langle f(\boldsymbol{S})\rangle$, that is

$$\langle f(\boldsymbol{S})\rangle = \frac{\int d\boldsymbol{s} f(\boldsymbol{s}) e^{-\beta E(\boldsymbol{s})}}{\int d\boldsymbol{s} e^{-\beta E(\boldsymbol{s})}} = \frac{\int_\gamma d\boldsymbol{\zeta} f(\boldsymbol{\zeta}) e^{-\beta E_e(\boldsymbol{\zeta})}}{\int_\gamma d\boldsymbol{\zeta} e^{-\beta E_e(\boldsymbol{\zeta})}} = f(\boldsymbol{\zeta}_0) \; , \tag{3.8}$$

and hence $\langle S_i \rangle = m_i^*$ and $\langle S_i S_j \rangle = m_i^* m_j^*$. It is seen that this approximation directly addresses the problem of calculating averages of dynamical variables in a system with energy function E. However, in the the next section we will determine the nature of the fixed-point solution $\boldsymbol{m}^*$ for the original saddle-point approximation as stated in [Peterson and Anderson 1987].

3.2 Variational methods

Generally, variational methods seek computationally tractable bounds on the partition function by simplifying the intractable joint probability distribution e.g. by modifying the local conditional probability functions. Variational methods come in various flavors; however, in the present review we mainly consider the Kullback-Leibler variational bound which provides a lower bound on the partition function. The first application of the KL variational bound to BNs was done by [Saul et al. 1996] in the context of sigmoid belief networks and recently in a Bayesian setting for graphical models [Attias 2000]. An alternative, however closely related, approach to obtain tractable bounds is to make use of the concept of convex duality [Rockafellar 1970]. This was done in [Jaakkola and Jordan 1999] to obtain an upper bound on the probability of positive findings in the QMR database.

3.2.1 Kullback-Leibler variational bound

Due to its operational simplicity, the KL variational bound is presently the most common mean field method for approximate learning and inference in graphical models. Given an intractable energy function, E, we are interested in the partition function, Z, of the probability density

$$p(\boldsymbol{s}) = \frac{1}{Z} e^{-E(\boldsymbol{s})} \ . \tag{3.9}$$

Consider a family of tractable distributions, q, for which the partition function Z_0 is tractable. Now, our goal is to pick the member of q which approximates p the best. As our measure of closeness between the two distributions we will use the KL divergence

$$KL(q \| p) = \langle \log q/p \rangle_q = \langle \log q \rangle_q - \langle \log p \rangle_q \ . \tag{3.10}$$

Since $KL(q\|p) \geq 0$ we get the lower bound, Z_q^*, on the partition function

$$\log Z \geq \log Z_q^* = - \langle E \rangle_q - \langle \log q \rangle_q \ , \tag{3.11}$$

where the average, $\langle E \rangle_q$ of the energy function with respect to the tractable distribution is called the *variational energy*. The equality (3.11) can equivalently be expressed in terms of the *variational free energy* , F_q^*, which provides an upper bound on the *free energy*

$$F = -\log Z \leq F_q^* = \langle E \rangle_q + \langle \log q \rangle_q = \langle E \rangle_q - \mathcal{H}(q) \ , \tag{3.12}$$

where we have introduced the *differential entropy* of q [Cover and Thomas 1991],

$$\mathcal{H}(q) = -\langle \log q \rangle_q \ . \tag{3.13}$$

Inserting q into eq. (3.11) we recover what is known in statistical physics as the *Gibbs-Bogoliubov-Feynman inequality* [Zhang 1996],

$$\log Z \geq \log Z_0 - \langle E - E_0 \rangle_q \ , \tag{3.14}$$

where E_0 is the energy function of the tractable probability distribution. Equation (3.11) has translated the problem of minimizing the KL distance between the approximating distribution, q, and the intractable distribution, p, into a problem of maximizing a lower bound on the partition function. This shows that the feasibility of the KL variational bound depends in our ability to calculate the variational free energy F_q^*. A widely used family of approximating distributions is obtained by the *naive mean field* (NMF) *ansatz* [Parisi 1988],

$$q(\boldsymbol{s}) = \prod q_i(s_i) \ , \tag{3.15}$$

in which all the dynamical variables, $\{S_i\}$, are independent, i.e. the approximating distribution is the family of non-interacting systems. Calculating the variational (functional) derivative of the variational free energy F_q^* with respect to q_i we readily obtain the free-form optimized marginal distribution

$$q_i \propto e^{-\langle E(\boldsymbol{s}) \rangle_{q_{\setminus i}}} \ , \tag{3.16}$$

which we will denote the *naive mean field distribution* of S_i. The average $\langle \cdot \rangle_{q_{\setminus i}}$ is taken with respect to the marginal distribution $q_{\setminus i} = \prod_{j \neq i} q_j$. Note that the equality in eq. (3.11) is attained if and only if the target distribution is factorized. This follows directly from the identity

$$KL(q \| p) = 0 \Leftrightarrow q = p \ . \tag{3.17}$$

An important property of the NMF approach is that only local operations are needed for updating the mean field distribution eq. (3.16) [Haft et al. 1999]. This is easily seen by expressing the energy function in terms of the joint distribution

$$q_i \propto e^{\left\langle \log p(s_i|\boldsymbol{s}_{\setminus i}) p(\boldsymbol{s}_{\setminus i}) \right\rangle_{q_{\setminus i}}} \propto e^{\left\langle \log p(s_i|\boldsymbol{s}_{\setminus i}) \right\rangle_{q_{\setminus i}}} \tag{3.18}$$

$$\propto e^{\langle \log p(s_i|\boldsymbol{M}_i) \rangle_{q_{\boldsymbol{M}_i}}} \ , \tag{3.19}$$

where we in the last line have introduced the Markov blanket, $\boldsymbol{M}_i$, of the variable S_i. This is a very important property since it shows the optimization process itself leads to coupling of dynamical variables even though they are assumed independent.

Whereas the interpretation of the fixed-points of the saddle-point method is not very transparent, this is more straightforward in the variational approach. Assume that the order statistic m_i^* of the saddle-point approximation is to be interpreted as the mean of node S_i with respect to its naive mean field distribution q_i, i.e.

$$m_i^* = \langle S_i \rangle_{q_i} \ . \tag{3.20}$$

Comparing eq. (3.6) and eq. (3.16) it follows that the saddle-point method and the NMF approach are equivalent, i.e. they yield the same results, if

$$\langle E(\boldsymbol{S}) \rangle_{q_{\setminus i}} = S_i \frac{\partial E(\boldsymbol{m})}{\partial m_i} \ . \tag{3.21}$$

In the case of the BM we see that condition (3.21) holds and for both approaches we recover the classic mean field equations

$$\langle S_i \rangle_{q_i} = \tanh \left(\sum_{j \neq i} J_{ij} \langle S_j \rangle_{q_j} + \theta_i \right) \ . \tag{3.22}$$

Usually, the marginal distribution, q_i, is assumed to be of some specific form parameterized with a set of additional variational parameters, $\{\lambda_i\}$, from which the corresponding averages can be inferred. The bound optimization is then performed with respect to the variational parameters [Jordan et al. 1998]. To illustrate this, consider the node S_i of the Boltzmann machine shown in figure 3.1. We wish to approximate the target distribution in a family of approximating distributions where all the nodes are independent and each node distribution, q_i, is parameterized by λ_i. Equation (3.19) showed that S_i will be coupled to the nodes in its Markov blanket and hence coupled to the variational parameters in its Markov blanket.

Another advantage of the variational methods is that the approximating distribution can be aimed more directly at the factorization of the target distribution, hence obtaining a better approximation. Instead of using the naive mean field ansatz it is possible to exploit substructures of the original distribution which are computational tractable [Saul and Jordan 1996]. This feature of the KL variational bound was exploited in context of factorial hidden Markov model [Ghahramani and Jordan 1997]. However, for some architectures, such as the QMR database and the layered sigmoid belief network, it is not straightforward to identify tractable substructures. Alternatively we could just choose a specific family of approximating distributions, q, which we think is able to capture particular properties of the target distribution. For instance, in [Jaakkola and Jordan 1998] a mixture of mean field distributions were used to capture higher-order interactions of a multimodal target distribution.

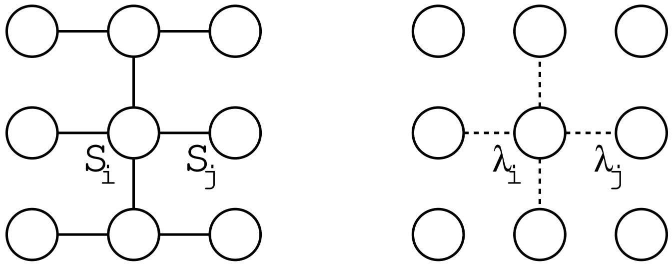


Fig. 3.1. (a) A node S_i in a Boltzmann Machine. (b) KL variational transformation using the NMF ansatz. The mean field equations of S_i yields a deterministic relationship which only depends on the mean fields of the nodes in the Markov blanket of S_i. The deterministic relationship is illustrated by the dashed lines.

3.2.2 Linear Response Correction

A limitation of variational mean field theory using factorized trial distributions is that it only treats "self-interactions" correctly, while producing trivial second moments, i.e. $\langle S_i S_i \rangle = \langle S_i \rangle \langle S_j \rangle$ for $i \neq j$. As pointed out by [Kappen and Rodríguez 1998b] this naive mean-field approximation may fail completely in some cases when applied to Boltzmann learning which was introduced in this context by [Peterson and Anderson 1987]. Instead, they went on to propose an efficient learning algorithm based on linear response (LR) theory. Linear response theory gives a recipe for computing an improved approximation to the covariances directly from the solution to the NMF equations [Parisi 1988]. Consider an intractable distribution p corresponding to a system with energy function E. The trick is to impose an external field (or bias) $\boldsymbol{h} = \{h_i\}$ to the intractable system, i.e. we consider a modified (and still intractable) system with energy function

$$E_{\boldsymbol{h}} = E - \sum h_i s_i \ . \tag{3.23}$$

The mean of the random variable S_i with respect to the probability distribution of the modified system can be obtained utilizing the external field

$$\langle S_i \rangle_{p_{\boldsymbol{h}}} = \frac{1}{Z_{\boldsymbol{h}}} \frac{\partial Z_{\boldsymbol{h}}}{\partial h_i} \tag{3.24}$$

where $Z_{\boldsymbol{h}} = \left\langle e^{-E_{\boldsymbol{h}}(\boldsymbol{s})} \right\rangle_1$ is the partition function of the modified system with probability density $p_{\boldsymbol{h}}$. The mean of the random variable S_i with respect to the probability distribution, p, of the original system is straightforwardly obtained by simply removing the external field, i.e.

$$\langle S_i \rangle = \langle S_i \rangle_{p_{\boldsymbol{h}}} \Big|_{\boldsymbol{h}=\boldsymbol{0}} \ . \tag{3.25}$$

Clearly, the second moment can be obtained following the same procedure

$$\langle S_i S_j \rangle_{p_{\boldsymbol{h}}} = \frac{1}{Z_{\boldsymbol{h}}} \frac{d^2 Z_{\boldsymbol{h}}}{dh_j dh_i} = \frac{1}{Z_{\boldsymbol{h}}} \frac{dZ_{\boldsymbol{h}} \langle S_i \rangle_{p_{\boldsymbol{h}}}}{dh_j} \tag{3.26}$$

$$= \langle S_i \rangle_{p_{\boldsymbol{h}}} \langle S_j \rangle_{p_{\boldsymbol{h}}} + \frac{d \langle S_i \rangle_{p_{\boldsymbol{h}}}}{dh_j} , \tag{3.27}$$

which follows directly using eq. (3.24) repeatedly. Equation (3.27) which is valid for expectations with respect to the density $p_{\boldsymbol{h}}$ of the modified intractable system is known as the *linear response theorem* [Parisi 1988]. However, provided $\langle S_i \rangle_{q_{\boldsymbol{h}}}$ obtained by the NMF approach is a reasonable approximation to $\langle S_i \rangle_{p_{\boldsymbol{h}}}$, this equation can indeed be used to get a nontrivial approximation to the covariance with respect to the modified system

$$\chi_{ij}^{\boldsymbol{h}} = \langle S_i S_j \rangle_{q_{\boldsymbol{h}}} - \langle S_i \rangle_{q_{\boldsymbol{h}}} \langle S_j \rangle_{q_{\boldsymbol{h}}} = \frac{d \langle S_i \rangle_{q_{\boldsymbol{h}}}}{dh_j} , \tag{3.28}$$

which in turn can be used to obtain a nontrivial approximation to the covariance with respect to the original intractable distribution

$$\chi_{ij} = \chi_{ij}^{\boldsymbol{h}} \big|_{\boldsymbol{h}=\boldsymbol{0}} . \tag{3.29}$$

In the case where $\boldsymbol{h}$ is in itself the linear terms in E the mean is given directly by

$$\langle S_i \rangle = \frac{1}{Z} \frac{\partial Z}{\partial h_i} , \tag{3.30}$$

where $Z_{\boldsymbol{h}} = \langle e^{-E(\boldsymbol{s})} \rangle_1$ in the partition function of the intractable distribution p. Similar, we get the following relation between the mean and covariance

$$\chi_{ij} = \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle = \frac{d \langle S_i \rangle}{dh_j} , \tag{3.31}$$

which is the basic expression for getting an improved estimate of the covariance within the mean field ansatz.

3.2.3 KL versus backward KL minimization

The main reason for using the KL divergence as the distance measure between the tractable approximating density, q, and the target density, p, is that it only requires computations of expectations with respect to the tractable distribution. However, as pointed out by [Jordan et al. 1998], another motivation for using the KL distance is its connection to the concept of convex duality [Rockafellar 1970]. It can be verified that the log partition function

$$\log Z = \log \left\langle e^{-E(\boldsymbol{s})} \right\rangle_1 , \tag{3.32}$$

is a convex function of $-E$. Hence, the log partition function can be expressed in terms of an integral[2] of $-E$ over a variational kernel $\boldsymbol{q}$ and the dual function of $\log Z$, more specifically

$$\log Z = \sup_q \{-\langle E\rangle_q - f^*(q)\} \ , \tag{3.33}$$

where the dual function f^* of $\log Z$ is given by

$$f^*(q) = \sup_E \{-\langle E\rangle_q - \log Z\} \ . \tag{3.34}$$

Hence, to calculate the dual function f^* we need to find the energy function E^* that maximizes $-\langle E\rangle_q - \log Z$. The maximizing energy function is obtained straightforwardly by free-form optimization which yields the result

$$E^* = -\log Zq \ . \tag{3.35}$$

Inserting the maximizing energy function eq. (3.35) into eq. (3.34) we get the dual function of $\log Z$,

$$f^*(q) = \langle \log q\rangle_q = -\mathcal{H}(q) \ , \tag{3.36}$$

which we recognize as minus the differential entropy of q. Finally, by inserting the dual function into eq. (3.33) we recover the KL variational lower bound eq. (3.11).

In section 3.2.1 we found the optimal trial distribution q_i within the mean field ansatz was given by the naive mean field distribution eq. (3.16) when using $KL(q\|p)$ as our measure of closeness. Due to the asymmetry of the KL divergence it is natural to ask what would be the optimal trial distribution if we instead use $KL(p\|q)$ as our distance measure between q and p. Let us denote this distance as the *backward KL divergence (BKL)* between q and p,

$$BKL(q\|p) = KL(p\|q) = \langle \log p/q\rangle_p \ . \tag{3.37}$$

Hence, to calculate the backward KL distance we have to take expectations with respect to the intractable density p. The optimal trial distribution q_i within the NMF ansatz is readily obtained by free-form optimizing BKL, i.e.

$$\frac{\delta}{\delta q_i} KL(p\|q) = -\frac{\delta}{\delta q_i}\langle \log q_i\rangle_p + \lambda_i \tag{3.38}$$

$$= -\int d\boldsymbol{s}' p(\boldsymbol{s}')\delta(s_i' - s_i)/q_i(s_i') + \lambda_i \tag{3.39}$$

$$= -p(s_i)/q_i(s_i) + \lambda_i = 0 \ , \tag{3.40}$$

[2] For convex $f(\boldsymbol{x})$ we have $f(\boldsymbol{x}) = \sup_{\boldsymbol{q}}\{\boldsymbol{q}^T\boldsymbol{x} - f^*(\boldsymbol{q})\}$, where the dual function $f^*(\boldsymbol{q}) = \sup_{\boldsymbol{x}}\{\boldsymbol{q}^T\boldsymbol{x} - f(\boldsymbol{x})\}$ and $\boldsymbol{q} = \{q_i\}$ is the set of variational parameters. In this case we take the linear combination to the infinite limit hence treating the sum as an integral and $\boldsymbol{q}$ as a function instead of a set of parameters.

where λ_i is the Lagrange multiplier insuring the normalization of q_i. This shows that the optimal q_i is given by

$$q_i = p(s_i) \ . \tag{3.41}$$

That is, q_i is obtained by matching the moments of the marginal $p(s_i)$. To gain additional insight into the difference between the two distance measures let us consider a specific example where the approximating probability density is taken from the exponential family

$$q(x|\boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}} e^{-\boldsymbol{\theta}^T \boldsymbol{f}(x)} = e^{-\boldsymbol{\theta}^T \boldsymbol{f}(x) - \log Z_{\boldsymbol{\theta}}} \ . \tag{3.42}$$

Start by considering the KL divergence as our measure of closeness between the family of approximating densities q and the target density p. To find the minimizing parameter $\boldsymbol{\theta}^*$ we take the derivative of the KL distance

$$\frac{\partial}{\partial \boldsymbol{\theta}} \langle \log q/p \rangle_q = -\frac{\partial}{\partial \boldsymbol{\theta}} \left(\boldsymbol{\theta}^T \langle \boldsymbol{f} \rangle_q + \log Z_{\boldsymbol{\theta}} + \langle \log p \rangle_q \right) \tag{3.43}$$

$$= -\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \langle \boldsymbol{f} \rangle_q^T \right) \boldsymbol{\theta} + \frac{\partial}{\partial \boldsymbol{\theta}} \langle \log p \rangle_q \right) \tag{3.44}$$

$$= \boldsymbol{\chi} \boldsymbol{\theta} - \frac{\partial}{\partial \boldsymbol{\theta}} \langle \log p \rangle_q \ , \tag{3.45}$$

where the second equality makes use of $\langle \boldsymbol{f} \rangle_q = -(\partial / \partial \boldsymbol{\theta}) \log Z_{\boldsymbol{\theta}}$. In the last equality we have made use of the linear response correction calculation from section 3.2.2 to show that

$$-\frac{\partial}{\partial \boldsymbol{\theta}} \langle \boldsymbol{f} \rangle_q^T = \left\langle \boldsymbol{f}^T \frac{\partial}{\partial \boldsymbol{\theta}} \left(\boldsymbol{\theta}^T \boldsymbol{f} + \log Z_{\boldsymbol{\theta}} \right) \right\rangle_q \tag{3.46}$$

$$= \left\langle \boldsymbol{f}^T \left(\boldsymbol{f} + \frac{\partial}{\partial \boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \right) \right\rangle_q \tag{3.47}$$

$$= \left\langle \boldsymbol{f}^T \left(\boldsymbol{f} - \langle \boldsymbol{f} \rangle_q \right) \right\rangle_q = \boldsymbol{\chi} \ , \tag{3.48}$$

where $\boldsymbol{\chi}$ is the covariance matrix given by

$$\boldsymbol{\chi} = \langle \boldsymbol{f} \boldsymbol{f}^T \rangle_q - \langle \boldsymbol{f} \rangle_q \langle \boldsymbol{f}^T \rangle_q \ . \tag{3.49}$$

Due to the changed sign of the bias term $\boldsymbol{\theta}$ the sign has changed in these expressions for the mean and covariance compared to the similar expressions obtained in section 3.2.2. The last term in eq. (3.45) can be calculated as

$$-\frac{\partial}{\partial \boldsymbol{\theta}} \langle \log p \rangle_q = \left\langle \left(\frac{\partial}{\partial \boldsymbol{\theta}} \left(\boldsymbol{\theta}^T \boldsymbol{f} + \log Z_{\boldsymbol{\theta}} \right) \right) \log p \right\rangle_q \tag{3.50}$$

$$= \left\langle \left(\boldsymbol{f} - \langle \boldsymbol{f} \rangle_q \right) \log p \right\rangle_q \tag{3.51}$$

$$= \langle \boldsymbol{f} \log p \rangle_q - \langle \boldsymbol{f} \rangle_q \langle \log p \rangle_q \tag{3.52}$$

This shows that the minimizing parameter $\boldsymbol{\theta}^*$ for the KL distance has to satisfy

$$\boldsymbol{\chi}\boldsymbol{\theta}^* = \langle \boldsymbol{f} \rangle_q \langle \log p \rangle_q - \langle \boldsymbol{f} \log p \rangle_q \ . \tag{3.53}$$

Now consider the backward KL distance as the measure of closeness. Taking the derivatives of BKL yields

$$\frac{\partial}{\partial \boldsymbol{\theta}} \langle \log p/q \rangle_p = -\frac{\partial}{\partial \boldsymbol{\theta}} \langle \log q \rangle_p = \left\langle \frac{\partial}{\partial \boldsymbol{\theta}} \left(\boldsymbol{\theta}^T \boldsymbol{f} + \log Z_{\boldsymbol{\theta}} \right) \right\rangle_p \tag{3.54}$$

$$= \langle \boldsymbol{f} \rangle_p - \langle \boldsymbol{f} \rangle_q \tag{3.55}$$

where we in the first equality use that p is independent of $\boldsymbol{\theta}$. This shows that the minimizing parameter $\boldsymbol{\theta}^*$ for the BKL distance has to satisfy

$$\langle \boldsymbol{f} \rangle_p = \langle \boldsymbol{f} \rangle_q \ . \tag{3.56}$$

Provided the target distribution p is also from the exponential family

$$p(x|\boldsymbol{\theta}_0) = \frac{1}{Z_{\boldsymbol{\theta}_0}} e^{-\boldsymbol{\theta}_0^T \boldsymbol{f}(x)} = e^{-\boldsymbol{\theta}_0^T \boldsymbol{f}(x) - \log Z_{\boldsymbol{\theta}_0}} \ , \tag{3.57}$$

the minimizing parameter $\boldsymbol{\theta}^*$ must satisfy

$$\boldsymbol{\chi}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) = \mathbf{0} \qquad \text{and} \qquad \langle \boldsymbol{f} \rangle_p = \langle \boldsymbol{f} \rangle_q \ , \tag{3.58}$$

for KL and BKL minimization, respectively.

Let us consider a specific example where the target and approximating distribution is given respectively by

$$p(x|\boldsymbol{\theta}_0) \propto e^{-\frac{1}{2}\kappa(x-\alpha)^2(x+\alpha)^2} \qquad \text{and} \qquad q(x|\boldsymbol{\theta}) \propto e^{-\frac{1}{2}\tau(x-\mu)^2} \ , \tag{3.59}$$

where μ and τ is the mean and precision of the Gaussian trial distribution. We start by minimizing the KL distance

$$\langle \log q/p \rangle_q = \frac{1}{2} \left\langle \log \tau + \kappa(x^4 - 2x^2\alpha^2) - \tau(x^2 + \mu^2 - 2x\mu) \right\rangle_q + K \tag{3.60}$$

$$= \frac{1}{2} \left(\log \tau + \kappa \left\langle x^4 \right\rangle_q - (2\kappa\alpha^2 + \tau) \left\langle x^2 \right\rangle_q + \tau\mu^2 \right) + K \tag{3.61}$$

$$= \frac{1}{2} \left(\log \tau + \kappa \left\langle x^4 \right\rangle_q - 2\kappa\alpha^2 \left(\tau^{-1} + \mu^2 \right) - 1 \right) + K \ , \tag{3.62}$$

where K is a constant independent of κ and τ. In this example, the KL distance only depends on the first, second and fourth order moment of the Gaussian trail density. Using tables of standard integrals (e.g. [Gradshteyn and Ryzhik 1980]), a general expression for calculating moments of the Gaussian eq. (3.59) can be obtained

$$\langle x^n \rangle_q = \mu^n n! \sum_{k=0}^{\lfloor n/2 \rfloor} \frac{(2\tau\mu^2)^{-k}}{(n-2k)!k!} \ , \tag{3.63}$$

which in turn is used to calculate the fourth order moment

$$\left\langle x^4 \right\rangle = \mu^4 + 6\tau^{-1}\mu^2 + 3\tau^{-2} \ . \tag{3.64}$$

The derivative of eq. (3.62) with respect to the mean of the approximating Gaussian is given by

$$\frac{\partial}{\partial \mu} \left\langle \log q/p \right\rangle_q = \mu\kappa \left(2\mu^2 + 6\tau^{-1} - 2\alpha^2\right) \ , \tag{3.65}$$

which shows that the stationary points of the KL divergence must satisfy

$$\mu = 0 \qquad \text{or} \qquad \mu^2 = \alpha^2 - 3\tau^{-1} \quad \text{where} \quad \alpha^2 \geq 3\tau^{-1} \ . \tag{3.66}$$

For $\mu = 0$ the stationary points are attained at

$$\tau = -\kappa\alpha^2 \pm \sqrt{\kappa(\kappa\alpha^4 + 6)} \ . \tag{3.67}$$

However, only the positive solution is a valid solution since we know as a fact that κ is positive. The stationary points corresponding to $\mu^2 = \alpha^2 - 3\tau^{-1}$ are attained at

$$\tau_{\pm} = 2\left(\kappa\alpha^2 \pm \sqrt{\kappa(\kappa\alpha^4 - 3)}\right) \quad \text{where} \quad \alpha^4 \geq 3/\kappa \ . \tag{3.68}$$

The valid stationary point can be found by taking a closer look at the two possible solutions $\tau_{\pm}$ in eq. (3.68) when α is large. We see that $\alpha^2\tau_{-} \to 0$ for $\alpha^4 \gg 3/\kappa$, hence violating the condition that $\alpha^2 \geq 3\tau^{-1}$ in eq. (3.66). However, for all valid choices of α and κ we have $\alpha^2\tau_{+} \geq 6$ which indeed satisfies $\alpha^2 \geq 3\tau^{-1}$. Hence, in the bifurcating α region the solution to the precision is given by τ_{+}. By computing the Hessian of the KL divergence it can be verified that these stationary points indeed correspond to the maxima of the KL divergence, which is summarized in

$$(\mu, \tau) = \begin{cases} (0, -\kappa\alpha^2 + \sqrt{\kappa(\kappa\alpha^4 + 6)}) & \text{for } \alpha^4 < 3/\kappa \\ (\pm\sqrt{\alpha^2 - 3\tau^{-1}}, 2\left(\kappa\alpha^2 + \sqrt{\kappa(\kappa\alpha^4 - 3)}\right)) & \text{for } \alpha^4 \geq 3/\kappa \end{cases} \tag{3.69}$$

Figure (3.2)(a-b) shows as a function of α, the optimal mean μ and the precision τ of the Gaussian trial distribution in the case where $\kappa = 3$, i.e. where the phase transition appears at $\alpha = 1$. Furthermore, the figure shows the corresponding MCMC estimates of the mean and precision obtained by averaging 200 parameter estimates obtained by Hybrid Monte Carlo simulations [Neal 1993; MacKay 1998]; each of which draws 10000 samples (discarding 1000 as burn-in samples) using 10 leapfrog iterations with step size 0.1 and unit variance Gaussian momentum updates. This simulation shows that sampling based methods (at least phenomenological) examine the same type symmetry breaking as the ones occurring when minimizing the KL divergence. By

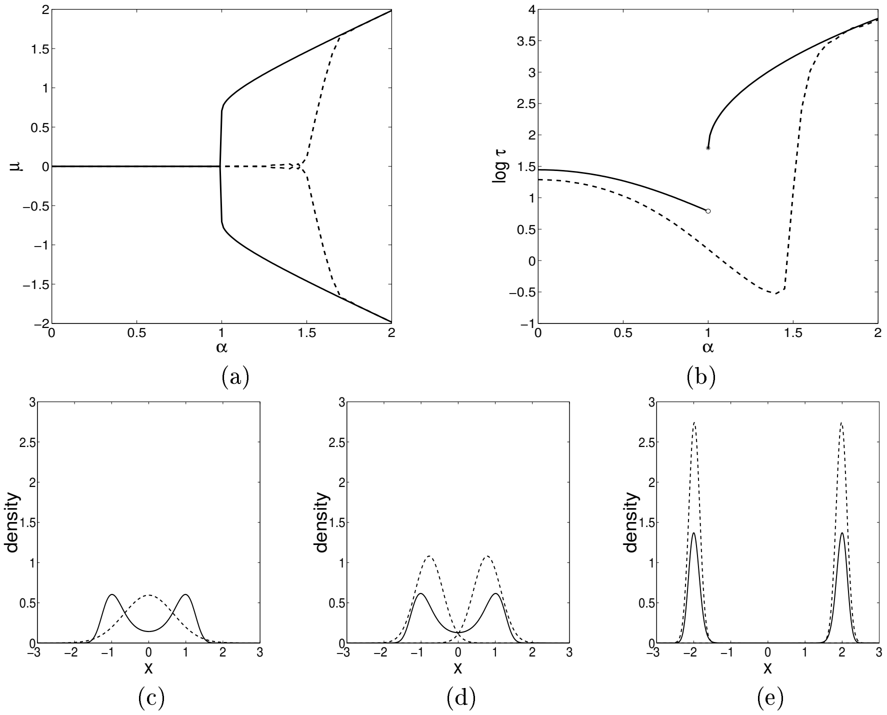


Fig. 3.2. Symmetry breaking in the KL divergence. Shows (a) the true and simulated mean μ and (b) the true and simulated precision τ in the case where $\kappa = 3$. The target density (solid) and the possible approximating densities (dashed) is shown for (c) $\alpha = 0.99$, i.e. just before the phase transition (d) for $\alpha = 1.01$ i.e. immediately after the phase transition and (e) for $\alpha = 2.0$ i.e. in the case of well separated modes in target density. Both bifurcated solutions are shown in (d) and (e).

varying the Hybrid Monte Carlo parameters and hence the diffusion length of the sampler the location of bifurcation point can be changed. However, the key point is that for any values for these control parameters the Monte Carlo methods will for a particular α undergo a phase transition which furthermore suggests that the mean field approximation in some sense possesses an intrinsic temperature. The observation that sample-based methods and the naive mean field approximation, at least quantitatively, yield the same results shows that mean field methods in fact are a reasonable approach for approximating intractable densities; give and take the usual pros and cons of both methods.

We know from eq. (3.41) that the optimal factorized trial distribution obtained by minimizing the BKL divergence simply matches the moments of marginal of the target distribution. Hence, for every choice of α the mean of the Gaussian trial density is constantly $\mu = 0$ which in turn implies that no phase transition will appear when using the BKL divergence as the measure of distance between the target and factorized trail density.

Besides the computational issues mentioned in the beginning of the section this example suggests another reason for favoring the KL divergence over the BKL divergence when approximating intractable densities. Clearly, the moment matching property of the BKL is undesirable when approximating multimodal densities, e.g. consider the case shown in figure 3.2(e). Obviously, we would not consider a Gaussian with mean 0 as being a useful approximation for this target density. Indeed, a much more reasonable approximation would be to pick up one of the modes like the approximation obtained by minimizing the KL distance. When considering graphical models with hidden variables such multi-modality is likely to appear in the marginal densities of the children of the hidden variables. This is especially true then the hidden variables are multinomial.

The main point of the section was to highlight some of the qualitatively differences between approximation obtained by minimizing the KL versus the BKL divergence. Loosely speaking, the KL divergence places emphasis on not inferring unlikely values at the cost of not inferring some of the likely values, whereas the BKL divergence places emphasis on inferring all likely values at the cost of inferring some of the unlikely values [Frey 1998]. Indeed, we see in figure 3.2 that the mean field approximation tends to underestimate the variance of the true target density.

Recently, some authors (e.g. [Ghahramani and Beal 2000]) have suggested using importance sampling from the KL variational approximation to obtain unbiased estimates of various quantities of interest. However, a good importance sampler should in general be more heavy tailed than the target density. Hence, since the KL variational approximation tends to underestimated the variance this endeavor to yield unbiased estimates should be carried out with caution. The situation is best illustrated by considering the variance of the importance weights, i.e.

$$\left\langle (p/q)^2 \right\rangle_q - \left(\langle p/q \rangle_q \right)^2 = \left\langle p^2/q \right\rangle_1 - 1 \tag{3.70}$$

$$= \frac{Z_q}{Z_p^2} \int dx e^{-2E_p(x)+E_q(x)} - 1 \ , \tag{3.71}$$

where we in the last equation have expressed the target density p and approximating density q in term of the Boltzmann distribution with energy function E_p and E_q, respectively. Hence, the variance converges if $E_q \leq 2E_p$ and diverges otherwise. E.g. assume that both the target and approximating density are Gaussians with variance σ_p^2 and σ_q^2 respectively. In that case the variance of the importance weights diverges if $\sigma_q^2 \leq \sigma_p^2/2$. When the variance of the importance weights diverges the variance of the importance estimate of say $\langle f \rangle$ will also tend to diverge provided the function f does not decay fast enough towards zero. Depending on the functional form of the target and approximating density this effect becomes more or less pronounced.

3.3 Perturbational methods

We have seen that the analysis of the partition function Z is central in mean field theory because once it is calculated all statistical information about the system can be deduced from it. This suggests that the quality of the inferred statistics depends strongly upon how well the approximating partition function matches the true partition function of the system. One inherent limitation of the KL variational bound methods is that they do not suggest ways to tighten the bound beyond the family of approximating distributions e.g. the family of factorized densities. However, by extending the idea of variational bounds, it is possible to derive methods in which more accurate approximations can be achieved in a systematic way. Common for these methods is that they perturb particular functions around a computationally tractable density.

3.3.1 The Plefka expansion

The naive mean field distribution eq. (3.16) was obtained by minimizing the variational free energy in the family of factorized distributions. However, as mentioned in section 3.2.1 the minimized variational free energy will be different from the true free energy when the target distribution is not factorized, i.e. it is impossible to get an arbitrary approximation to the free energy within this framework. Instead of minimizing the variational free energy F^*_q subject to the constraint that the trial density q is factorized we now minimize a slightly modified variational free energy $F^*_{q,\xi}$ subject to the constraint that q belongs to the family of distributions for which $\langle \boldsymbol{S} \rangle_q = \boldsymbol{m}$ [Opper and Winther 2000a], i.e.

$$G_\xi(\boldsymbol{m}) = \min_q \{F^*_{q,\xi}\} \quad \text{subject to} \quad \langle \boldsymbol{S} \rangle_q = \boldsymbol{m} \ , \tag{3.72}$$

where the modified variational free energy is given by

$$F^*_{q,\xi} = \langle \xi E \rangle_q + \langle \log q \rangle_q \ . \tag{3.73}$$

The modified variational free energy essentially redefines the energy of the system to be ξE. This minimization problem is easily solved by taking the functional derivative

$$\frac{\delta F^*_{q,\xi}}{\delta q} = \frac{\delta}{\delta q}\left(\xi \langle E \rangle_q + \langle \log q \rangle_q + \sum h_i (m_i - \langle s_i \rangle_q)\right) \tag{3.74}$$

$$= \xi E(s) + \log q + 1 - \sum h_i s_i \ , \tag{3.75}$$

where we in the first line have introduced a set of Lagrange multipliers $\{h_i\}$ enforcing the constraint $\langle \boldsymbol{S} \rangle_q = \boldsymbol{m}$. Equating eq. (3.75) to zeros we find the optimal distribution within the family is given by

$$q_{\xi,\boldsymbol{h}} \propto e^{-\xi E(\boldsymbol{s})+\sum h_i s_i} \ , \tag{3.76}$$

where the dependence of ξ and $\boldsymbol{h}$ is shown explicitly in the subscript of $q = q_{\xi,\boldsymbol{h}}$. Finally, by inserting the optimal probability density into eq. (3.72) we get

$$G_\xi(\boldsymbol{m}) = \sum h_i m_i - \log Z_{q_{\xi,\boldsymbol{h}}} = \sum h_i m_i + F_{q_{\xi,\boldsymbol{h}}} \ , \tag{3.77}$$

where $\boldsymbol{h}$ is given implicit by the constraint $m = \langle \boldsymbol{S} \rangle_{q_{\xi,\boldsymbol{h}}}$ and $Z_{q_{\xi,\boldsymbol{h}}}$ is the normalizing constant of $q_{\xi,\boldsymbol{h}}$. Since the mean vector $\boldsymbol{m}$ is physically more meaningful than $\boldsymbol{h}$, it is appropriate to consider it as free and treat $\boldsymbol{h}$ as being dependent on $\boldsymbol{m}$. The set of mean field equations are obtained by

$$\frac{\partial G_\xi}{\partial m_i} = \sum_{i'} \frac{\partial G_\xi}{\partial h_{i'}} \frac{\partial h_{i'}}{\partial m_i} \tag{3.78}$$

$$= \sum_{i'} \left(m_{i'} + \sum_k h_k \frac{\partial m_k}{\partial h_{i'}} - \frac{\partial}{\partial h_{i'}} \log Z_{q_{\xi,\boldsymbol{h}}} \right) \frac{\partial h_{i'}}{\partial m_i} \tag{3.79}$$

$$= \sum_{i'} \sum_k h_k \frac{\partial m_k}{\partial h_{i'}} \frac{\partial h_{i'}}{\partial m_i} = h_i \ , \tag{3.80}$$

where we have assumed that $\boldsymbol{m} = \langle \boldsymbol{S} \rangle_{q_{\xi,\boldsymbol{h}}}$ can be uniquely solved for any fixed $\boldsymbol{m}$ and ξ, such that $\partial \boldsymbol{m}/\partial \boldsymbol{h}$ is the inverse of $\partial \boldsymbol{h}/\partial \boldsymbol{m}$. The first step towards getting the original free energy is to set $\boldsymbol{h} = \boldsymbol{0}$ which translates the mean fields equation into

$$\frac{\partial G_\xi}{\partial m_i} = 0 \ , \tag{3.81}$$

Unfortunately, we are not in a position to evaluate G_ξ at $\xi = 1$ which would give us the Gibbs free energy G for the system we are actually interested in. Clearly, the normalization of $q_{\xi,\boldsymbol{0}}$ is as intractable as the normalization of the original density so it is not clear what we have gained by this approximation to p. However, the parameter ξ makes it possible to interpolate between the factorized distribution ($\xi = 0$) and the distribution of interest $q_{\boldsymbol{h}}$ ($\xi = 1$). As proposed by [Plefka 1982], the trick is to expand eq. (3.77) as around the tractable solution at $\xi = 0$. In other words, since $m = \langle S \rangle_{q_{\xi,\boldsymbol{h}}}$ is intractable for all $\xi \neq 0$ and tractable for $\xi = 0$ we approximate G_ξ around $\xi = 0$, i.e.

$$G_\xi(\boldsymbol{m}) = G_0(\boldsymbol{m}) + \sum_{k=1}^{\infty} \frac{1}{k!} \left. \frac{\partial^k G_\xi}{\partial \xi^k} \right|_{\xi=0} \xi^k \ . \tag{3.82}$$

In [Bhattacharyya and Keerthi 1999] the Plefka expansion is seen more directly from a KL variational perspective. Again the starting point is to consider a modified density

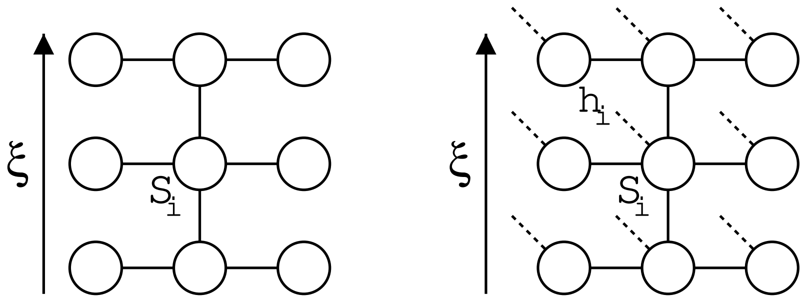


Fig. 3.3. The Plefka expansion from a KL variational bound perspective. (a) The modified target system is obtained by adding a global field ξ (change temperature, controlling the interaction) to the original system. (b) The approximating system is obtained by adding an additional external field (dashed lines) to the modified target system. The approximating density is now found minimizing the KL distance.

$$q_{\xi,\boldsymbol{h}}(\boldsymbol{s}) = \frac{1}{Z_{\xi,\boldsymbol{h}}} e^{-\xi E(\boldsymbol{s}) + \sum h_i s_i} \ , \tag{3.83}$$

where $Z_{\xi,\boldsymbol{h}} = \left\langle e^{-\xi E(\boldsymbol{s}) + \sum h_i s_i} \right\rangle_1$ is the associated partition function. Remember that $p = q_{1,\mathbf{0}}$ is the target distribution of interest. Let $q_\xi = q_{\xi,\mathbf{0}}$ be the approximating density. By using the Kullback-Leibler distance as the measure of closeness between the two distributions we obtain

$$KL(q_{\xi,\boldsymbol{h}} \,\|\, q_\xi) = \langle \log q_{\xi,\boldsymbol{h}} / q_\xi \rangle_{q_{\xi,\boldsymbol{h}}} \tag{3.84}$$

$$= \log Z_\xi - \log Z_{\xi,\boldsymbol{h}} + \sum h_i \langle S_i \rangle_{q_{\xi,\boldsymbol{h}}} \ , \tag{3.85}$$

where the last equation makes use of the fact that $\xi \langle E \rangle$ cancels out due to its appearance in both $\log q_{\xi,\boldsymbol{h}}$ and $\log q_\xi$. Note that the approximating family of densities have the target distribution as a member which obviously implies that it is possible to obtain the exact free energy of the target system. Using the fact that $KL(\cdot\|\cdot) \geq 0$ we recognize eq. (3.77) as the right hand side of

$$F_\xi \leq F_{\xi,\boldsymbol{h}} + \sum h_i \langle S_i \rangle_{q_{\xi,\boldsymbol{h}}} \ . \tag{3.86}$$

This view of the Plefka expansion is illustrated schematically in figure 3.3.

Let us have a closer look at the zeroth order term in the Plefka expansion. Noting that the partition function is independent of $\boldsymbol{S} = \{S_i\}$, we can write the zeroth order term as

$$G_0(\boldsymbol{m}) = \sum h_i m_i - \log Z_{q_{0,\boldsymbol{h}}} = \left\langle \sum h_i S_i - \log Z_{q_{0,\boldsymbol{h}}} \right\rangle_{q_{0,\boldsymbol{h}}} \tag{3.87}$$

$$= \left\langle \log \frac{e^{\sum h_i S_i}}{Z_{q_{0,\boldsymbol{h}}}} \right\rangle_{q_{0,\boldsymbol{h}}} = \langle \log q_{0,\boldsymbol{h}} \rangle_{q_{0,\boldsymbol{h}}} = -\mathcal{H}(q_{0,\boldsymbol{h}}) \ . \tag{3.88}$$

This shows that the zeroth order term is just the negative entropy of the factorized distribution. Before calculating any higher order derivatives of the

Gibbs free energy it is important to note that

$$\frac{1}{Z}\frac{\partial Z}{\partial \xi} = -\langle E\rangle_{q_{\xi,\boldsymbol{h}}} + \sum \frac{\partial h_i}{\partial \xi} m_i \ , \tag{3.89}$$

since $\boldsymbol{h}$ is in fact a function of $\boldsymbol{m}$ and ξ due to the constraint $\boldsymbol{m} = \langle \boldsymbol{S}\rangle_{q_{\xi,\boldsymbol{h}}}$. Let us for simplicity just consider the first and second derivatives of the Gibbs free energy. Using eq. (3.89) the first derivative readily becomes

$$\frac{\partial G_\xi}{\partial \xi} = \frac{\partial}{\partial \xi}\left(\sum h_i m_i - \log Z_{q_{\xi,\boldsymbol{h}}}\right) = \langle E\rangle_{q_{\xi,\boldsymbol{h}}} \ , \tag{3.90}$$

which in turn is used to calculate the second derivative

$$\frac{\partial^2 G_\xi}{\partial \xi^2} = \frac{\partial}{\partial \xi}\langle E\rangle_{q_{\xi,\boldsymbol{h}}} = \left\langle E(-E + \sum \frac{\partial h_i}{\partial \xi} S_i)\right\rangle_{q_{\xi,\boldsymbol{h}}} - \langle E\rangle_{q_{\xi,\boldsymbol{h}}} \frac{1}{Z}\frac{\partial Z}{\partial \xi} \tag{3.91}$$

$$= \langle E\rangle^2_{q_{\xi,\boldsymbol{h}}} - \left\langle E^2\right\rangle_{q_{\xi,\boldsymbol{h}}} + \left\langle E \sum \frac{\partial h_i}{\partial \xi}(S_i - m_i)\right\rangle_{q_{\xi,\boldsymbol{h}}} \ . \tag{3.92}$$

To calculate the second order derivative we need to evaluate

$$\left.\frac{\partial h_i}{\partial \xi}\right|_{\xi=0} = \left.\frac{\partial^2 G_\xi}{\partial \xi \partial m_i}\right|_{\xi=0} = \frac{\partial}{\partial m_i}\langle E\rangle_{q_{0,\boldsymbol{h}}} \ , \tag{3.93}$$

which is obtained using eq. (3.80) and (3.90). We see that the KL variational bound eq. (3.11) and (3.12) is obtained as the first order expansion of G. In [Kappen and Rodríguez 1998a], the linear response correction to the correlations was given by the Hessian of the Gibbs free energy. To illustrate the basic methodology of the Plefka expansion let us as an example consider our canonical example of the BM. Direct use of eq. (3.76) shows that for $\xi = 0$ any system of binary variables $S_i \in \{-1, 1\}$ has an optimal distribution given by

$$q_{0,\boldsymbol{h}}(\boldsymbol{s}) = \prod \frac{e^{h_i s_i}}{e^{h_i} + e^{-h_i}} \ , \tag{3.94}$$

and hence the mean of $\langle S_i\rangle$ is given by

$$m_i = \frac{e^{h_i} - e^{-h_i}}{e^{h_i} + e^{-h_i}} = \tanh(h_i) \ , \tag{3.95}$$

which in turn is used to solve for the Lagrange multipliers

$$h_i = \frac{1}{2}\log\frac{1 + m_i}{1 - m_i} \ . \tag{3.96}$$

Inserting eq. (3.96) into eq. (3.94) shows that we, not surprisingly, could have parametrized the factorized distribution directly in terms of the means, i.e.

$$q_{0,\boldsymbol{h}}(\boldsymbol{s}) = \prod \left(\frac{1+m_i}{2}\right)^{\delta_{s_i,1}} \left(\frac{1-m_i}{2}\right)^{\delta_{s_i,-1}} . \tag{3.97}$$

The zeroth order term, i.e. the entropy, of this binary probability distribution with $S_i \in \{-1, 1\}$ is then given by

$$G_0 = \sum \left(\frac{1+m_i}{2}\right) \log\left(\frac{1+m_i}{2}\right) + \left(\frac{1-m_i}{2}\right) \log\left(\frac{1-m_i}{2}\right) . \tag{3.98}$$

In order to calculate higher order terms we need to introduce the energy function of the specific system under consideration. The first order derivative needed in a Plefka expansion of the Boltzmann machine eq. (2.19) is easily obtained using eq. (3.90),

$$\left.\frac{\partial G_\xi}{\partial \xi}\right|_{\xi=0} = \left.\langle E \rangle_{q_{\xi,\boldsymbol{h}}}\right|_{\xi=0} = -\frac{1}{2}\sum_{i,j} J_{ij} m_i m_j - \sum_i \theta_i m_i . \tag{3.99}$$

Using the fact that

$$\left.\frac{\partial h_i}{\partial \xi}\right|_{\xi=0} = -\sum_{j\neq i} J_{ij} m_j - \theta_i , \tag{3.100}$$

the second order derivative can be obtained by tedious calculations

$$\left.\frac{\partial^2 G_\xi}{\partial \xi^2}\right|_{\xi=0} = -\frac{1}{2}\sum J_{ij}^2 (1-m_i^2)(1-m_j^2) . \tag{3.101}$$

The second order mean field equations are readily found by solving eq. (3.81) using the second over Plefka expansion

$$m_i = \tanh\left(\theta_i + \sum_{j\neq i} J_{ij} m_j + \sum_{j\neq i} J_{ij}^2 m_i (1-m_j^2)\right) . \tag{3.102}$$

This constitutes the classic TAP equations derived in [Thouless et al. 1977] for the SK model of disordered magnetic materials [Sherrington and Kirkpatrick 1975]. However, contrary to the original derivation of the TAP equations the perturbational derivation does not assume any knowledge about the distribution of the couplings. This is a useful property of the perturbational methods since the distribution of the couplings is usually not part of the specification of a probabilistic model. Finally, one should always keep in mind that the range of validity of perturbational methods is determined by the convergence domain of the resulting power series although this is difficult to access in practice.

Recently, various alternative approaches for performing approximate inference in intractable probabilistic models have been proposed by several authors. As for the Plefka expansion, most of these approaches rely in a perturbational expansion in a power series where each coefficient is evaluated in some tractable distribution, although there are exceptions to this, see e.g. [Leisink and Kappen 2000].

3.3.2 Variational cumulant expansions

In the *variational cumulant expansion* of [Barber and van de Laar 1999] the following modified energy function is considered

$$E_\xi = \xi E_1 + (1 - \xi) E_0 \ , \tag{3.103}$$

where E_1 is the energy function of the density of interest, E_0 is some tractable energy function and ξ is the perturbational parameter interpolating between the two corresponding densities. The log partition function of the modified system is then given by

$$\log Z_\xi = \log \left\langle e^{-E_0 - \xi(E_1 - E_0))} \right\rangle_1 = \log \left\langle Z_0 e^{-\xi(E_1 - E_0))} \right\rangle_{p_0} \tag{3.104}$$

$$= \log Z_0 + \log \left\langle e^{\xi(E_0 - E_1))} \right\rangle_{p_0} \ , \tag{3.105}$$

where the last term is recognized as the cumulant generating function associated to the moment generating function since

$$\frac{d^k}{d\xi^k} \left\langle e^{\xi(E_0 - E_1))} \right\rangle_{p_0} \Bigg|_{\xi=0} = \left\langle (E_0 - E_1)^k \right\rangle_{p_0} \ . \tag{3.106}$$

The partition function of the density of interest can now be found by Taylor expanding the cumulant generating function around $\xi = 0$,

$$\log Z_1 = \log Z_0 + \sum_{k=1}^{\infty} \frac{1}{k!} \kappa_{k,0} \tag{3.107}$$

$$= \log Z_0 + \sum_{k=1}^{l} \frac{1}{k!} \kappa_{k,0} + \frac{1}{(l+1)!} \kappa_{l+1,\xi} \ , \tag{3.108}$$

where $\kappa_{k,\xi}$ is the k'th cumulant of $E_0 - E_1$ with respect to p_ξ. The intractable remainder in the last equation follows from the mean value theorem where $0 \le \xi \le 1$. For the first order approximation we recover the KL variational bound through the Gibbs-Bogoliubov-Feynman inequality eq. (3.14),

$$\log Z_1 = \log Z_0 + \langle (E_0 - E_1) \rangle_{p_0} + \frac{1}{2} \kappa_{2,\xi} \tag{3.109}$$

$$\ge \log Z_0 + \langle (E_0 - E_1) \rangle_{p_0} \ , \tag{3.110}$$

where the inequality follows from the non-negativity of the variance $\kappa_{2,\xi}$ for any value of ξ. However, for higher order expansions it is not possible to obtain such a bound. Hence, maximizing the lower bound is no longer a reasonable approach for optimizing the variational parameters. Instead, it is possible to use an independence criterion due to the fact that the partition function of the density of interest is independent of the variational parameters $\{\lambda_i\}$, i.e.

$$\frac{d \log Z_1}{d\lambda_i} = 0 \tag{3.111}$$

In [Barber and van de Laar 1999] this approach is applied on the Boltzmann machine to recover the TAP solution eq. (3.102) as a second order expansion by imposing the additional constraint that the second order solution should be close to the first order solution.

3.3.3 The information geometrical viewpoint

In [Tanaka 2000] the mean field approximation is seen from an information geometrical point of view. Essentially this view boils down to considering the Gibbs free energy eq. (3.77) used in the Plefka expansion. In this case, however, the perturbational parameter is not a single scalar parameter ξ but instead the entire set of parameters mitigating interactions between dynamical variables, i.e. all parameters except bias parameters. The coefficients in the resulting power series expansion are found to be the cumulants of the approximating density. Taking on such information geometrical point of view makes it possible to gain some additional insight into to relationships between the naive mean field approach and more advanced approaches. It can be shown that the KL distance between a factorized density q_0 and a target density p can be expanded into two contributions

$$KL(q_0 \| p) = KL(q_0 \| q) + KL(q \| p) \ , \tag{3.112}$$

where q is a density which belongs to the same of the family of densities as p, i.e. shares the same canonical parameters, and furthermore satisfies $\langle S \rangle_q = \langle S \rangle_{q_0}$. The situation is illustrated in figure 3.4. The relationship eq. (3.112) quantifies the intuition that the naive mean field ansatz is a reasonable assumption when the target density is close to being factorized. In addition, it shows that higher order mean field approaches essentially takes into account the distance between the family (or manifold) of distributions, A, of which the target density is a member and the family of factorized distributions A_0. This means that the reactor terms in the higher order mean field methods arise from the $KL(q_0 \| q)$ term.

A closely related approach is found in [Kappen and Wiegerinck 2000] which makes use of the BKL and the fact that the naive mean field distribution, q_i, in this case is equal to the marginal distribution p_i. Since the marginal distribution p_i is intractable they propose expanding $\log p_i$ around q_i in terms of changes of all the parameters which give rise to interactions between dynamical variables.

3.3.4 The cavity approach and adaptive TAP

The adaptive TAP approach of [Opper and Winther 2000c] considers probabilistic models of the type

$$p(\boldsymbol{s}) = \frac{1}{Z}\rho(\boldsymbol{s})e^{\frac{1}{2}\sum_{i,j} s_i J_{ij} s_j + \sum \theta_i s_i} \ , \tag{3.113}$$

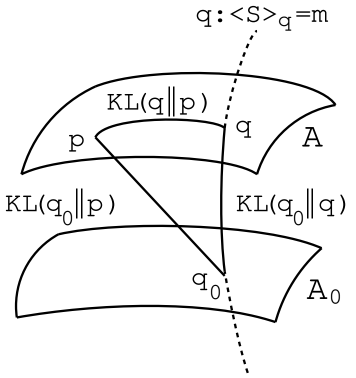


Fig. 3.4. Illustration of the information geometrical viewpoint of mean field approximations. The KL divergence between the approximating density q and the target density p, both of which lives on the same manifold of distributions A, is given by $KL(q \,\|\, p) = KL(q_0 \,\|\, p) - KL(q_0 \,\|\, q)$, where q_0 lives on the manifold of factorized densities A_0. To first order this yields the naive mean field approximation i.e. $KL(q \,\|\, p) \approx KL(q_0 \,\|\, p)$. The dashed line illustrates the distributions satisfying the constraint in eq. (3.72).

where the interaction-weights are symmetric, $J_{ij} = J_{ji}$, and $J_{ii} = 0$ such that all self-interactions are contained in the single variable constraint $\rho(\boldsymbol{s}) = \prod \rho(s_i)$. The derivation of the adaptive TAP equations is based on the cavity approach introduced by [Mezard et al. 1987]. The starting point of the cavity approach is the following exact equation for the marginal density of the dynamical variable S_i,

$$p(s_i) = \int d\boldsymbol{s}_{\setminus i} p(\boldsymbol{s}) \propto \rho(s_i) \int d\boldsymbol{s}_{\setminus i} e^{s_i(h_i+\theta_i)} p(\boldsymbol{s}_{\setminus i}) \ , \tag{3.114}$$

where $p(\boldsymbol{s}_{\setminus i})$ is the marginal distribution of all the remaining variables when S_i is excluded from the system. Since the dynamical variable S_i only interacts with the remaining variables through the field $h_i = \sum_j J_{ij} s_j$ it is convenient to introduce the *cavity distribution*, i.e. the distribution of the field h_i at the location of the missing variable S_i,

$$p(h_i) = \int d\boldsymbol{s}_{\setminus i} \delta(h_i - \sum_j J_{ij} s_j) p(\boldsymbol{s}_{\setminus i}) \ . \tag{3.115}$$

The marginal distribution of S_i can now be expressed in terms of the cavity distribution instead of the marginal $p(\boldsymbol{s}_{\setminus i})$, i.e.

$$p(s_i) = \frac{1}{Z_i} \rho(s_i) e^{s_i \theta_i} \left\langle e^{s_i h_i} \right\rangle_{\setminus i} = \frac{1}{Z_i} \rho(s_i) e^{s_i \theta_i} e^{-E_i} \ , \tag{3.116}$$

where Z_i is the partition function associated to the marginal and $\langle \cdot \rangle_{\setminus i}$ is the average with respect to the cavity distribution and the (minus) energy function is recognized as the cumulant generating function

$$-E_i(s_i) = \log \left\langle e^{s_i h_i} \right\rangle_{\setminus i} = \sum_k \frac{\kappa_k^{(i)}}{k!} s_i^k \ , \tag{3.117}$$

where $\kappa_k^{(i)}$ are the cumulants of the cavity distribution. The basic assumption of all cavity derivations of the TAP mean field theory is that all variables $\{S_i\}$ have only weak mutual dependencies. Mathematical expressed within the so-called clustering hypothesis this becomes equivalent to the vanishing of all cumulants $\kappa_k^{(i)}$ with $k > 2$ for fully connect systems [Opper and Winther 2000c]. The first two cumulants are given by $\kappa_1^{(i)} = \langle h_i \rangle_{\setminus i}$ and $V_i \equiv \kappa_2^{(i)} = \left\langle h_i^2 \right\rangle_{\setminus i} - \langle h_i \rangle_{\setminus i}^2$. Hence, under the assumption of vanishing higher order cumulants the marginal distribution of S_i is given by

$$p(s_i) = \frac{1}{Z_i} \rho(s_i) e^{\frac{1}{2} V_i s_i^2 + (\theta_i + \langle h_i \rangle_{\setminus i}) s_i} \tag{3.118}$$

The first set of TAP equations consists of the expectations of $\{S_i\}$ obtained by using the partition function, Z_i, of the marginal distribution eq. (3.118),

$$\langle S_i \rangle = \frac{\partial}{\partial \theta_i} \log Z_i \ . \tag{3.119}$$

To close the set of mean field equation we need to derive expressions for $\langle h_i \rangle_{\setminus i}$ and V_i. An expression of the average cavity field $\langle h_i \rangle_{\setminus i}$ is obtained by

$$\langle h_i \rangle = \frac{1}{Z_i} \int ds_i \rho(s_i) e^{s_i \theta_i} \frac{\partial}{\partial s_i} \left\langle e^{s_i h_i} \right\rangle_{\setminus i} = \langle h_i \rangle_{\setminus i} + V_i \langle S_i \rangle \ , \tag{3.120}$$

where the last equality follows from using the truncated power series expansion of the cumulant generating function eq. (3.117). The last term in eq. (3.120) is often referred to as the *Onsager reaction term* . While the naive mean field approach neglects the reaction term by setting $V_i = 0$ the adaptive TAP approach seeks to estimate the variance V_i by requiring self-consistency between two estimates of $\langle S_i^2 \rangle - \langle S_i \rangle^2$; one obtained using the TAP equations of the expectations directly and the other obtained by the linear response theorem from the naive mean field solution. The linear response theorem expresses the covariance matrix in terms of the mean, i.e.

$$\chi_{ij} = \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle = \frac{\partial \langle S_i \rangle}{\partial \theta_j} \ . \tag{3.121}$$

Provided that perturbations of the TAP equations leave the variances $\{V_i\}$ unchanged the linear response correction to the covariance matrix can be found by solving the set of linear equations given by

$$\chi_{ij} = \frac{\partial \langle S_i \rangle}{\partial \theta_j} = \frac{\partial \langle S_i \rangle}{\partial \theta_i}\frac{\partial \theta_i}{\partial \theta_j} + \frac{\partial \langle S_i \rangle}{\partial \langle h_i \rangle_{\setminus i}}\frac{\partial \langle h_i \rangle_{\setminus i}}{\partial \theta_j} \tag{3.122}$$

$$= \frac{\partial \langle S_i \rangle}{\partial \theta_i}\left(\delta_{ij} + \frac{\partial \langle h_i \rangle_{\setminus i}}{\partial \theta_j}\right) \tag{3.123}$$

$$= \frac{\partial \langle S_i \rangle}{\partial \theta_i}\left(\delta_{ij} + \sum_k (\theta_{ik} - V_k \delta_{ik})\chi_{kj}\right) , \tag{3.124}$$

where the last equality follows from using eq. (3.120). This set of linear equations can easily be solved with the result

$$\boldsymbol{\chi}^{LR} = (\boldsymbol{\Lambda} - \boldsymbol{J})^{-1} , \tag{3.125}$$

where $\boldsymbol{J} = \{J_{ij}\}$ is the matrix of interaction-weights and

$$\boldsymbol{\Lambda} \equiv \mathrm{diag}\,(\Lambda_1, \Lambda_2, \ldots, \Lambda_N) , \quad \text{where} \quad \Lambda_i \equiv V_i + \left(\frac{\partial \langle S_i \rangle}{\partial \theta_i}\right)^{-1} . \tag{3.126}$$

The set of TAP equations of the variances $\{V_i\}$ is obtained by requiring self-consistency in the estimates of $\langle S_i^2 \rangle - \langle S_i \rangle^2$, i.e. we equate the estimate χ_{ii}^{MF} obtained by direct use of the mean field equations and the linear response corrected estimate χ_{ii}^{LR},

$$\frac{\partial \langle S_i \rangle}{\partial \theta_i} = [(\boldsymbol{\Lambda} - \boldsymbol{J})^{-1}]_{ii} , \tag{3.127}$$

and solve for the variances $\{V_i\}$. The Gibbs free energy used in the Plefka expansion in section 3.3.1 was obtained by free-form minimization of a modified version of the variational free energy $F^*_{q,\xi}$ subject to the constraint the trail density q had to belong to the family of distributions for which $\langle \boldsymbol{S} \rangle_q = \boldsymbol{m}$. The TAP free energy is obtained in a similar way by imposing the additional constraint $\langle \boldsymbol{S}^2 \rangle_q = \boldsymbol{M}$, i.e.

$$G_\xi(\boldsymbol{m}, \boldsymbol{M}) = \min_q \{F^*_{q,\xi}\} \quad \text{s.t.} \quad \langle \boldsymbol{S} \rangle_q = \boldsymbol{m} \;\; \text{and} \;\; \langle \boldsymbol{S}^2 \rangle_q = \boldsymbol{M} , \tag{3.128}$$

where the modified variational free energy is defined by

$$F^*_{q,\xi} = -\left\langle \frac{1}{2}\boldsymbol{s}^T(\xi \boldsymbol{J})\boldsymbol{s} + \boldsymbol{s}^T\boldsymbol{\theta} + \log \rho(\boldsymbol{s}) \right\rangle_q + \langle \log q \rangle_q . \tag{3.129}$$

Again, the perturbational parameter ξ makes it possible to interpolate between the factorized trial density ($\xi = 0$) and the density of interest ($\xi = 1$). The optimizing trial density is readily found by free-form optimization of the modified variational free energy $F^*_{q,\xi}$, i.e.

$$q_{\xi,\boldsymbol{\gamma},\boldsymbol{\lambda}} \propto \rho(\boldsymbol{s}) e^{\frac{1}{2}\boldsymbol{s}^T(\xi \boldsymbol{J})\boldsymbol{s} + \boldsymbol{s}^T\boldsymbol{\theta} + \sum \gamma_i s_i + \frac{1}{2}\sum \lambda_i s_i^2} , \tag{3.130}$$

where $\{\gamma_i\}$ and $\{\lambda_i\}$ are the Lagrange multipliers enforcing the constraint of the means $\{m_i\}$ and second moments $\{M_i\}$, respectively. Substituting the optimizing trial density back into eq. (3.128) yields the TAP free energy of the modified system

$$G_\xi(\boldsymbol{m}, \boldsymbol{M}) = \sum_i \gamma_i m_i + \frac{1}{2}\sum_i \lambda_i M_i - \tag{3.131}$$

$$\log \int d\boldsymbol{s}\rho(\boldsymbol{s}) e^{\frac{1}{2}\boldsymbol{s}^T(\xi \boldsymbol{J}+\boldsymbol{\lambda})\boldsymbol{s}+\boldsymbol{s}^T(\boldsymbol{\theta}+\boldsymbol{\gamma})} \ , \tag{3.132}$$

where we have introduced the matrix $\boldsymbol{\lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and the vector $\boldsymbol{\gamma} = \{\gamma_i\}$ of variational parameters. By construction, the solution to $\partial_{\boldsymbol{\gamma}} G_\xi = 0$ and $\partial_{\boldsymbol{\lambda}} G_\xi = 0$ yields the fixed-point conditions $\boldsymbol{m} = \langle \boldsymbol{S} \rangle_{q_\xi}$ and $\boldsymbol{M} = \langle \boldsymbol{S}^2 \rangle_{q_\xi}$, respectively. To calculate the free energy of the system of interest, G_1, we make use of the following integral relation

$$G_1 - G_0 = \int_0^1 d\xi \frac{\partial G_\xi}{\partial \xi} = -\frac{1}{2}\int_0^1 d\xi \left\langle \boldsymbol{s}^T \boldsymbol{J} \boldsymbol{s} \right\rangle_{q_\xi} \tag{3.133}$$

$$= -\frac{1}{2}\int_0^1 d\xi \, \mathrm{Tr}\left((\boldsymbol{\chi}_\xi + \langle \boldsymbol{s} \rangle_{q_\xi} \langle \boldsymbol{s} \rangle_{q_\xi}^T) \boldsymbol{J} \right) \tag{3.134}$$

$$= -\frac{1}{2}\left(\boldsymbol{m}^T \boldsymbol{J} \boldsymbol{m} + \int_0^1 d\xi \, \mathrm{Tr}(\boldsymbol{\chi}_\xi \boldsymbol{J}) \right) \tag{3.135}$$

The remaining integral gives rise to the Onsager correction and can be evaluated using the linear response correction eq. (3.125) to the covariance [Opper and Winther 2000c],

$$\Delta G = -\frac{1}{2}\int_0^1 d\xi \, \mathrm{Tr}(\boldsymbol{\chi}_\xi \boldsymbol{J}) = -\frac{1}{2}\left(\log\det \boldsymbol{\chi} + \sum_i V_i \chi_{ii} - \sum_i \log \chi_{ii} \right) \tag{3.136}$$

Hence the TAP free energy of the system of interest is given by

$$G_1 = G_0 - \frac{1}{2}\boldsymbol{m}^T \boldsymbol{J} \boldsymbol{m} + \Delta G \ , \tag{3.137}$$

where the first term is the TAP free energy evaluated with the factorized trial density at $\xi = 0$,

$$G_0 = -\log \int d\boldsymbol{s}\rho(\boldsymbol{s}) e^{\frac{1}{2}\boldsymbol{s}^T \boldsymbol{\lambda} \boldsymbol{s} + \boldsymbol{s}^T(\boldsymbol{\theta}+\boldsymbol{\gamma})} + \sum_i \gamma_i m_i + \frac{1}{2}\sum_i \lambda_i M_i \tag{3.138}$$

The second term in eq. (3.137) is the naive mean field energy. As a sanity check, we should recover the adaptive TAP equations directly from the free energy eq. (3.137). Indeed, we see that the solution to $\partial_{\langle S_i \rangle} G_\xi = 0$ and $\partial_{\langle S_i^2 \rangle} G_\xi = 0$ yields the fixed-point conditions $\gamma_i = \langle h_i \rangle_{\backslash i}$ and $\lambda_i = V_i$, respectively.

4. Independent Component Analysis

In this chapter we develop mean field approaches for probabilistic independent component analysis (ICA). The sources are estimated from the mean of their posterior distribution and the mixing matrix (and noise level) is estimated by maximum a posteriori (MAP). The latter requires the computation of (a good approximation to) the correlations between sources. For this purpose we investigate three of the mean field methods considered in the previous chapter, namely the KL variational bound, linear response and adaptive TAP approach. These increasingly advanced mean field algorithms are tested on a number of problems. On synthetic data the advanced mean field approaches are able to recover the correct mixing matrix in cases where the variational mean field theory fails. For hand-written digits, sparse encoding is achieved using non-negative source and mixing priors. For speech, the mean field method is able to separate in the underdetermined (overcomplete) case of two sensors and three sources. One major advantage of the proposed method is its generality and algorithmic simplicity. Finally, we point out several possible extensions of the approaches developed here.

4.1 Introduction

Reconstruction of statistically independent source signals from linear mixtures is an active research field with numerous important applications, for background and references see e.g. [Lee 1998; Girolami 2000]. Blind signal separation in the face of additive noise typically involves four estimation problems: Estimation of source signals, source distribution, mixing coefficients, and noise distribution.

A full Bayesian treatment of the combined estimation problem is possible but requires extensive Monte Carlo sampling [Belouchrani and Cardoso 1995], therefore several authors have proposed variational (also known as mean field or ensemble) approaches in which the posterior distributions are either approximated by factorized Gaussians and/or integrals over the posteriors are evaluated by saddle point approximations [Attias 1999; Belouchrani and Cardoso 1995; Lewicki and Sejnowski 2000; Lappalainen and Miskin 2000; Hansen 2000; Rowe 1999; Knuth 1999]. The resulting algorithm is an Expectation-Maximization (EM) like procedure in which the four estimation problems are performed sequentially. One important problem with these approximations arises from the assumed posterior independence of sources. In particular, variational mean field theory using factorized trial distributions only treats "self-interactions" correctly, while producing trivial second moments, i.e. $\langle S_i S_i \rangle = \langle S_i \rangle \langle S_j \rangle$ for $i \neq j$. This is a poor approximation when estimating the mixing matrix and noise distribution since these estimates will typically depend upon correlations.

Recently, Kappen and Rodríguez [Kappen and Rodríguez 1998b] pointed out that for Boltzmann Machines this naive mean-field (NMF) approximation — introduced in this context by [Peterson and Anderson 1987] — may fail completely in some cases. They went on to propose an efficient learning algorithm based on linear response (LR) theory. Linear response theory gives a recipe for computing an improved approximation to the covariances directly from the solution to the NMF equations [Parisi 1988]. In this chapter, we give a general presentation of LR theory and apply it to the probabilistic ICA problem. We also briefly outline the supposedly more accurate adaptive TAP mean field theory [Opper and Winther 2000c] and compare this method to the NMF and LR approach. The actual derivation of the adaptive TAP mean field approach in context of the probabilistic ICA model is presented in chapter 5. Whereas estimates of correlations obtained from variational mean field theory and its linear response correction in general differ, adaptive TAP is constructed such that it is consistent with linear response theory.

We expect that advanced mean field methods such as LR and TAP can be useful in the many contexts within neural computation, where variational mean field theory already have proven to be useful, e.g. for sigmoid belief networks [Saul et al. 1996]. In our experience, the main difference between variational mean field and the advanced methods lies in the estimates of correlations (often needed in algorithms of the EM-type) and the calculation of the likelihood of the data. We will, however, postpone the discussion of the latter to chapter 5 where a general method for computing the likelihood from the covariance matrix is presented. In ICA simulations, we find that the variational approach can fail typically by ignoring some of the sources and consequently overestimating the noise covariance. The LR and TAP approaches on the other hand succeed in all cases studied. However, we do not find a significant improvement using TAP (which is also somewhat more computationally intensive), suggesting that LR is close to being the optimal mean field approach for the probabilistic ICA model.

The derivation of the mean-field equations is valid for a general source prior (without temporal correlation) and tractable for priors that can be integrated analytically against a Gaussian kernel. This includes mixture of Gaussians, Laplacian and binary distributions. For other priors, one has to evaluate an extensive number of one dimensional integrals numerically. Alternatively, one can construct computationally tractable ICA algorithms using priors that are only defined implicitly. To illustrate this point we define one such algorithm which approximately corresponds to the prior having a power law tail.

To underline the flexibility and computational power of the probabilistic ICA framework and its mean field implementation, we give two quite different real world examples of recent interest that straightforwardly can be solved within this framework. The first example is that of separating speech in the overcomplete setting of two sensors and three sources [Lewicki and Sejnowski 2000] using a heavy tailed source prior such as a Laplacian or the (approximative) power law prior described above. The second real world problem considered in this chapter is that of feature extraction in images. For images, it is natural to work with a non-negativity constraint for the mixing matrix and sources as in [Lee and Seung 1999]. In the probabilistic framework this type of prior knowledge is readily build into the mixing matrix and source priors.

Throughout this chapter we confine ourselves to fixed source priors. There are, however, no theoretical problems in extending the EM algorithm to estimating hyperparameters. In fact, we will address this problem in chapter 5. Alternatively, see e.g. [Attias 1999] for a nice example of the methodology of estimating the source prior parameters within the EM framework. Hence, in this chapter we are mainly concerned with the inferential step of the learning problem, which in general is the hard part of any learning algorithm.

The chapter is organized as follows. In section 4.2 the basic probabilistic ICA model and the associated learning problem is stated. Section 4.3 concerns the inference part of the learning problem; we will see that variational mean field theory, linear response theory and the adaptive TAP approach can be seen as stepwise more refined ways of estimating correlations. Applying the advanced mean field methods to independent component analysis is the main contribution of this chapter. Another contribution is the generality of the framework. In section 4.4 we examine various types of explicitly given source priors which in turn leads us to define an implicitly given source prior. The impatient or application minded reader might consult section 4.4.1 which shows a table summarizing all priors considered in this chapter. Section 4.5 shows some simulation results on both synthetic data and on real world data. The pseudo-code for the algorithm is outlined in section 4.6 and some additional priors not directly used in this chapter are given in section 4.7. Finally, obvious ways to extend this work is outlined in the discussion given in section 4.8.

4.2 Probabilistic ICA

We formulate the ICA problem as follows [Hansen 2000]: The measurements are a collection of N temporal D-dimensional signals $\boldsymbol{X} = \{X_{dt}\}$, $d = 1, \dots, D$ and $t = 1, \dots, N$, where X_{dt} denotes the measurement at the

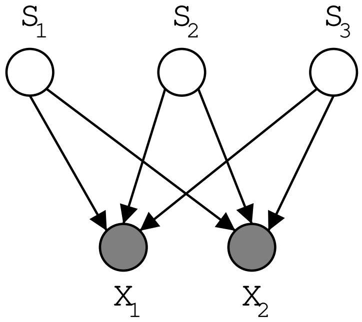


Fig. 4.1. The generative model for noisy ICA in the case of $I = 3$ sources and $D = 2$ sensors. A priori, the sources $\{S_i\}$ are mutually independent but a posteriori they become coupled through the observation $\{X_i\}$ due to "explaining away" effects.

dth sensor at time t. Similarly, let $\boldsymbol{S} = \{S_{it}\}$, $i = 1, \ldots, I$, denote a collection of I mutually statistical independent sources, where S_{it} is the ith source at time t. The measured signal $\boldsymbol{X}$ is assumed to be an instantaneous linear mixing of the sources corrupted with additive white Gaussian noise $\boldsymbol{\Gamma}$ that is

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{\Gamma} , \tag{4.1}$$

where $\boldsymbol{A}$ is a (time independent) mixing matrix and the noise is assumed to be without temporal correlations and with a time independent covariance matrix $\boldsymbol{\Sigma}$, i.e. we have $\overline{\Gamma_{dt}\Gamma_{d't'}} = \delta_{tt'}\Sigma_{dd'}$. Thus, we have the following likelihood for parameters and sources

$$p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}, \boldsymbol{S}) = (\det 2\pi\boldsymbol{\Sigma})^{-\frac{N}{2}} \, e^{-\frac{1}{2}\operatorname{Tr}(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})} \, . \tag{4.2}$$

Figure 4.1 shows the generative model for noisy ICA in the case of $I = 3$ hidden sources and $D = 2$ sensors. The aim of independent component analysis is to recover the unknown quantities given a set of observables; namely the sources $\boldsymbol{S}$, the mixing matrix $\boldsymbol{A}$ and the noise covariance $\boldsymbol{\Sigma}$.

The main difficulty is associated with the estimation of the source signals. The estimation problems for the mixing matrix and the noise covariance matrix are relatively simple, given the sufficient source statistics. Hence, our primary objective is to improve on the estimate of sufficient statistics from the posterior distribution of the sources. The mixing matrix $\boldsymbol{A}$ and the noise covariance $\boldsymbol{\Sigma}$ are then in turn estimated by maximum a posteriori (MAP) (or maximum likelihood II (ML-II)). This naturally leads to a EM-type algorithm where the expectation step amounts to finding the posterior mean and covariances of the sources and the maximization step is the MAP/ML-II estimation. Mean field methods especially the advanced ones are well suited for the non-trivial expectation step.

Given the likelihood eq. (4.2), the posterior distribution of the sources is readily given by

$$p(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{A},\boldsymbol{\Sigma}) = \frac{p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma},\boldsymbol{S})p(\boldsymbol{S})}{p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma})} \ , \tag{4.3}$$

where $p(\boldsymbol{S})$ is a prior on the sources which might include temporal correlations (although we will postpone this problem to a future contribution [Højen-Sørensen et al. 2001a]).

4.2.1 Estimation of mixing matrix and noise covariance

The likelihood of the parameters is given by

$$p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma}) = \int d\boldsymbol{S} p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma},\boldsymbol{S})\, p(\boldsymbol{S}) \ . \tag{4.4}$$

The problem of estimating the mixing matrix and noise covariance now amounts to finding the saddle-points of the likelihood eq. (4.4) with respect to the mixing matrix and noise covariance. We note that the saddle-points will be given in terms of averages over the source posterior. Alternatively, we could as our starting point have used the EM approach and considered the complete log-likelihood and the mean sufficient statistics directly. It is easily seen that the two approaches in this case are equivalent due to the Gaussian likelihood. It is the computation of mean sufficient statistics with respect to the posterior which pose the main challenge for mean field approaches since the sources will be coupled through the observations.

The mixing matrix $\boldsymbol{A}$ will be estimated by maximum a posteriori (MAP) and the noise by ML-II for convenience

$$\boldsymbol{A}_{\mathrm{MAP}} = \underset{\boldsymbol{A}}{\operatorname{argmax}}\, p(\boldsymbol{A}|\boldsymbol{X},\boldsymbol{\Sigma}) \tag{4.5}$$

$$\boldsymbol{\Sigma}_{\mathrm{MLII}} = \underset{\boldsymbol{\Sigma}}{\operatorname{argmax}}\, p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma}) \ , \tag{4.6}$$

where the posterior of $\boldsymbol{A}$ is given by $p(\boldsymbol{A}|\boldsymbol{X},\boldsymbol{\Sigma}) \propto p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma})p(\boldsymbol{A})$, where $p(\boldsymbol{A})$ is the prior on $\boldsymbol{A}$. For the optimization in eqs. (4.5) and (4.6), we need the derivatives of the likelihood term

$$\frac{\partial}{\partial \boldsymbol{A}} \log p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}\langle\boldsymbol{S}\rangle^T - \boldsymbol{A}\langle\boldsymbol{S}\boldsymbol{S}^T\rangle) \tag{4.7}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \log p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma}) = \frac{1}{2}\boldsymbol{\Sigma}^{-1}\langle(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})^T\rangle\boldsymbol{\Sigma}^{-1} - \frac{N}{2}\boldsymbol{\Sigma}^{-1} \ , \tag{4.8}$$

where $\langle\cdot\rangle = \langle\cdot\rangle_{\boldsymbol{S}|\boldsymbol{A},\boldsymbol{\Sigma},\boldsymbol{X}}$ denotes the posterior average with respect to the sources given the mixing matrix and noise covariance. Equating eq. (4.8) to zero leads to the well known result for $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma}_{\mathrm{MLII}} = \frac{1}{N}\langle(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})^T\rangle \ . \tag{4.9}$$

In the particular case of measurements with i.i.d. noise we can simplify the covariance $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, hence $\sigma^2 = \mathrm{Tr}\, \boldsymbol{\Sigma}_{\mathrm{MLII}}/D$, where D is the number of sensors.

For $\boldsymbol{A}$, we consider two factorized priors $p(\boldsymbol{A}) = \prod_{di} p(A_{di})$, a zero mean Gaussian $p(A_{di}) \propto \exp(-\alpha_{di} A_{di}^2/2)$ and the Laplace distribution $p(A_{di}) \propto \exp(-\beta_{di}|A_{di}|)$. Furthermore, we consider optimizing A_{di} both unconstrained and constrained to be non-negative. Clearly, the MAP approach offers a flexibility for encoding prior knowledge about $\boldsymbol{A}$ which is not available in the maximum likelihood II approach, i.e. one can encode sparseness [Hyvärinen and Karthikesh 2000] and non-negativeness (for e.g. images and text, see section 4.5 and [Lee and Seung 1999]).

Unconstrained mixing matrices. A straightforward calculation gives us the following iterative equation for the MAP estimate of $\boldsymbol{A}$,

$$\boldsymbol{A}^{(k+1)} = \left(\boldsymbol{X} \langle \boldsymbol{S} \rangle^T - \boldsymbol{\Sigma}(\alpha \boldsymbol{A}^{(k)} + \beta \mathrm{sign}(\boldsymbol{A}^{(k)})) \right) \langle \boldsymbol{S}\boldsymbol{S}^T \rangle^{-1} , \tag{4.10}$$

where we have included both priors and set $\alpha_{di} = \alpha$ and $\beta_{di} = \beta$. This equation can be solved explicitly for the Gaussian prior with equal noise variance on all sensors, i.e. $\beta = 0$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, and yields the result

$$\boldsymbol{A} = \boldsymbol{X} \langle \boldsymbol{S} \rangle^T \left(\langle \boldsymbol{S}\boldsymbol{S}^T \rangle + \alpha \sigma^2 \boldsymbol{I} \right)^{-1} . \tag{4.11}$$

The ML-II estimate is the special case obtained by setting $\alpha = 0$.

Non-negative mixing matrices. To enforce non-negative $\boldsymbol{A}$, we introduce a set of non-negative Lagrange multipliers $L_{di} \geq 0$ and maximize the modified cost: $\log p(\boldsymbol{A}|\boldsymbol{X}, \boldsymbol{\Sigma}) + \mathrm{Tr}\, \boldsymbol{L}^T \boldsymbol{A}$. Solving for the Lagrange multipliers we get

$$\mathbf{L} = \boldsymbol{\Sigma}^{-1}(\mathbf{A}\langle \mathbf{S}\mathbf{S}^T \rangle - \mathbf{X}\langle \mathbf{S} \rangle^T) + \alpha \mathbf{A} + \beta . \tag{4.12}$$

We can write down an iterative update rule for $A_{di} > 0$ using the Kuhn-Tucker condition $L_{di} A_{di} = 0$ [Luenberger 1984] together with the result for the Lagrange multipliers

$$A_{di}^{(k+1)} = \frac{[\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \langle \boldsymbol{S} \rangle^T]_{di}}{[\boldsymbol{\Sigma}^{-1} \boldsymbol{A}^{(k)} \langle \boldsymbol{S}\boldsymbol{S}^T \rangle]_{di} + \alpha A_{di}^{(k)} + \beta} A_{di}^{(k)} . \tag{4.13}$$

In the case of no prior knowledge i.e. $\alpha = 0$ and $\beta = 0$, we get an update rule similar to the image space reconstruction algorithm used in positron emission tomography (see e.g. [Pierro 1993] for references) or the more recently proposed non-negative matrix factorization procedure of [Lee and Seung 1999].

4.3 Mean Field Theory

We will present three different mean field approaches that give us estimates of the source second moment matrix of increasing quality. First, we derive mean field equations using the standard variational mean field theory. Next, using linear response theory, we obtain directly from the variational solution improved estimates of $\langle \boldsymbol{S}\boldsymbol{S}^T \rangle$ needed for estimating $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$. Finally, we present the adaptive TAP approach of Opper and Winther [Opper and Winther 2000c] which goes beyond the simple factorized trial distribution of variational mean field theory to give a theory which is self-consistent to within linear response corrections. From mean field theory we also get an approximation to the likelihood $p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma})$ which can be used for model selection [Hansen 2000].[1] In appendix 4.6, we summarize all mean field equations and give an EM-type recipe for solving them.

The following derivation is valid for any source prior without temporal correlations. Specific source priors are discussed in section 4.4. Although equations for the mean field estimates of the mean and covariance of the sources are written with equality in this section, it is to be understood that they are only approximations.

4.3.1 Variational Approach

We adopt the standard KL variational mean field theoretic approach and approximate the posterior distribution, $p(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{A},\boldsymbol{\Sigma})$, in a family of product distributions $q(\boldsymbol{S}) = \prod_{i,t} q(S_{it})$.[2] For a Gaussian likelihood $p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma},\boldsymbol{S})$, the optimal choice of $q(S_{it})$ is given by a Gaussian times the prior [Csató et al. 2000]

$$q(s_{it}) \propto p(s_{it}) e^{-\frac{1}{2}\lambda_{it}s_{it}^2 + \gamma_{it}s_{it}} , \tag{4.14}$$

where we use the canonical parameterization of the Gaussian density. This result is obtained straightforwardly by using of the functional form of the naive mean field distribution, eq. (3.16). Also, it turns out to be useful to consider the canonical parameterization of the Gaussian likelihood for the model parameters $\{\boldsymbol{A},\boldsymbol{\Sigma}\}$

$$p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{\Sigma},\boldsymbol{S}) = p(\boldsymbol{X}|\boldsymbol{J},\boldsymbol{h},\boldsymbol{S}) = \frac{1}{Z_L} e^{-\frac{1}{2}\operatorname{Tr}(\boldsymbol{S}^T\boldsymbol{J}\boldsymbol{S}) + \operatorname{Tr}(\boldsymbol{h}^T\boldsymbol{S})} , \tag{4.15}$$

where $\log Z_L = \frac{N}{2}\log\det 2\pi\boldsymbol{\Sigma} + \frac{1}{2}\operatorname{Tr}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ is the log partition function of the likelihood, and we have introduced the $M \times M$ interaction-matrix $\boldsymbol{J}$ and the data dependent external field $\boldsymbol{h}$ (having same dimension as $\boldsymbol{S}$) given respectively by,

$$\boldsymbol{J} = \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A} \tag{4.16}$$

$$\boldsymbol{h} = \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \ . \tag{4.17}$$

[1] The variational approximation is a lower bound to the exact likelihood whereas the TAP and LR approximations — not given here — are not bounds, but hopefully more accurate.

[2] Note that $q(S_{it})$ is also the variational mean field approximation to the marginal distribution $\int \prod_{i'\neq i,t'\neq t} dS_{i't'} p(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{A},\boldsymbol{\Sigma})$.

Note that $\boldsymbol{h}$ acts as an external field from which all moments of the sources can be obtained. This is the key property that we will make use of in the next section when we derive the linear response corrections. The starting point of the variational derivation of mean field equations is the Kullback-Leibler divergence between the product distribution $q(\boldsymbol{S})$ and the true source posterior, i.e.

$$\begin{aligned} KL(q\|p) &= \int d\boldsymbol{S} q(\boldsymbol{S}) \log \frac{q(\boldsymbol{S})}{p(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\Sigma})} \\ &= \log p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}) - \log p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}, \mathrm{NMF}) \quad (4.18) \\ \log p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}, \mathrm{NMF}) &= \sum_{i,t} \log \int ds_{it} p(s_{it}) e^{-\frac{1}{2}\lambda_{it} s_{it}^2 + \gamma_{it} s_{it}} \quad (4.19) \\ &\quad + \frac{1}{2} \sum_{it} (\lambda_{it} - J_{ii}) \langle S_{it}^2 \rangle + \mathrm{Tr}(\boldsymbol{h} - \boldsymbol{\gamma})^T \langle \boldsymbol{S} \rangle \\ &\quad + \frac{1}{2} \mathrm{Tr} \langle \boldsymbol{S}^T \rangle (\mathrm{diag}(\boldsymbol{J}) - \boldsymbol{J}) \langle \boldsymbol{S} \rangle - \log Z_L \ , \end{aligned}$$

where $p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}, \mathrm{NMF})$ is the naive mean field approximation to the likelihood and $\mathrm{diag}(\boldsymbol{J})$ is the diagonal matrix of $\boldsymbol{J}$. The Kullback-Leibler is zero when $q = p$ and positive otherwise. The parameters of q should consequently be chosen as to minimize $KL(q\|p)$. The saddle points define the mean field equations:[3]

$$\frac{\partial}{\partial \langle \mathbf{S} \rangle} KL(q\|p) = 0 : \qquad \boldsymbol{\gamma} = \mathbf{h} - (\mathbf{J} - \mathrm{diag}(\mathbf{J})) \langle \mathbf{S} \rangle \tag{4.20}$$

$$\frac{\partial}{\partial \langle S_{it}^2 \rangle} KL(q\|p) = 0 : \qquad \lambda_{it} = J_{ii} \ . \tag{4.21}$$

The remaining two equations depend explicitly on the source prior, $p(\boldsymbol{S})$;

$$\begin{aligned} \frac{\partial}{\partial \gamma_{it}} KL(q\|p) = 0 : \quad \langle S_{it} \rangle &= \frac{\partial}{\partial \gamma_{it}} \log \int ds_{it} p(s_{it}) e^{-\frac{1}{2}\lambda_{it} s_{it}^2 + \gamma_{it} s_{it}} \\ &\equiv m(\gamma_{it}, \lambda_{it}) \end{aligned} \tag{4.22}$$

$$\frac{\partial}{\partial \lambda_{it}} KL(q\|p) = 0 : \quad \langle S_{it}^2 \rangle = -2 \frac{\partial}{\partial \lambda_{it}} \log \int ds_{it} p(s_{it}) e^{-\frac{1}{2}\lambda_{it} s_{it}^2 + \gamma_{it} s_{it}} \ . \tag{4.23}$$

The variational mean $m(\gamma_{it}, \lambda_{it})$ plays a crucial role in defining the mean field algorithm since all dependence upon the prior is implicit in m (as well as in $\frac{\partial m}{\partial \gamma}$ for the advanced mean field methods). In section 4.4, we calculate

[3] The requirement that we should be at a local minima of $\log p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}, \mathrm{NMF})$ is fulfilled when the covariance matrix eq. (4.27) is positive definite. To test whether we are at the global minima is harder. However, when the model is well-matched to the data, we expect the problem to be convex.

$m(\gamma_{it}, \lambda_{it})$ for some of the prior distributions found in the ICA literature. Finally, by inserting the saddle points into eq. (4.19), the naive mean field approximation to the likelihood reduces to

$$\log p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}, \text{NMF}) = \sum_{i,t} \log \int ds_{it} p(s_{it}) e^{-\frac{1}{2}\lambda_{it} s_{it}^2 + \gamma_{it} s_{it}} \tag{4.24}$$

$$+\frac{1}{2} \operatorname{Tr}\langle \boldsymbol{S}^T \rangle (\boldsymbol{J} - \operatorname{diag}(\boldsymbol{J})) \langle \boldsymbol{S} \rangle - \log Z_L \ . \tag{4.25}$$

This approximation to the likelihood can be used to determine the number of latent sources. We will, however, postpone this subject until chapter 5.

4.3.2 Linear Response Theory

So far we have not discussed how to obtain mean field approximations to the covariances

$$\chi_{ii'}^{tt'} \equiv \langle S_{it} S_{i't'} \rangle - \langle S_{it} \rangle \langle S_{i't'} \rangle \ .$$

Since variational mean field theory uses a factorized trial distribution, the covariances between different variables is trivially predicted to be zero. However, using linear response theory, we can improve the variational mean field solution. As mentioned earlier, $\boldsymbol{h}$ acts as an external field. This makes it possible to calculate the means and covariances as derivatives of $\log p(\boldsymbol{X}|\boldsymbol{J}, \boldsymbol{h})$, i.e.

$$\langle S_{it} \rangle = \frac{\partial \log p(\boldsymbol{X}|\boldsymbol{J}, \boldsymbol{h})}{\partial h_{it}} \tag{4.26}$$

$$\chi_{ii'}^{tt'} = \frac{\partial^2 \log p(\boldsymbol{X}|\boldsymbol{J}, \boldsymbol{h})}{\partial h_{i't'} \partial h_{it}} = \frac{\partial \langle S_{it} \rangle}{\partial h_{i't'}} \ . \tag{4.27}$$

These relations are exact when using the exact likelihood. However, we can also use the NMF likelihood through the mean field equations (4.20), (4.21) and (4.22) to derive an approximate equation for $\chi_{ii'}^{tt'}$,

$$\begin{aligned} \chi_{ii'}^{tt'} &= \frac{\partial m(\gamma_{it}, \lambda_{it})}{\partial \gamma_{it}} \frac{\partial \gamma_{it}}{\partial h_{i't'}} \\ &= \frac{\partial m(\gamma_{it}, \lambda_{it})}{\partial \gamma_{it}} \left(- \sum_{i'', i'' \neq i} J_{ii''} \chi_{i''i'}^{tt} + \delta_{ii'} \right) \delta_{tt'} \ . \end{aligned} \tag{4.28}$$

As a direct consequence of the lack of temporal correlations in the present setting, the $\boldsymbol{\chi}$-matrix factorizes in time, i.e. $\chi_{ii'}^{tt'} = \delta_{tt'} \chi_{ii'}^{t}$. We can straightforwardly solve for $\chi_{ii'}^{t}$

$$\chi_{ii'}^{t} = \left[(\boldsymbol{\Lambda}_t + \boldsymbol{J})^{-1} \right]_{ii'} \ , \tag{4.29}$$

where we have defined the diagonal matrix

$$\boldsymbol{\Lambda}_t = \operatorname{diag}(\Lambda_{1t}, \ldots, \Lambda_{It}), \qquad \Lambda_{it} \equiv \left(\frac{\partial m(\gamma_{it}, \lambda_{it})}{\partial \gamma_{it}}\right)^{-1} - J_{ii} \; . \tag{4.30}$$

For comparison, the naive mean field result is $\chi_{ii'}^{t,\mathrm{NMF}} = \delta_{ii'} \frac{\partial \langle S_{it} \rangle}{\partial h_{it}}$ which follows directly from eq. (4.23).

Why is the covariance matrix obtained by linear response more accurate? Here, we give an argument that can be found in Parisi's book on statistical field theory [Parisi 1988]: Let us assume (as always implicit in any mean field theory) that the approximate and exact distribution are close in some sense, i.e. $q(\boldsymbol{S}) - p(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\Sigma}) = \varepsilon$. Then by direct application of the factorized distribution we have $\langle S_{it} S_{i't} \rangle_{\mathrm{Exact}} = \langle S_{it} S_{i't} \rangle_{\mathrm{NMF}} + \mathcal{O}(\varepsilon)$. On the other hand since $KL(q \parallel p)$ is non-negative the NMF theory log-likelihood gives a lower bound on the log-likelihood, see eq. (4.18). Consequently, the linear term vanishes in the expansion of the log-likelihood: $\log p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}) = \log p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}, \mathrm{NMF}) + \mathcal{O}(\varepsilon^2)$. Obtaining moments of the variables through derivatives of the approximate log-likelihood, i.e. by linear response, is therefore more precise than to use the trial distribution directly.

For some specific cases it is possible to demonstrate the improvement directly. Consider the Gaussian prior[4] $p(s_{it}) \propto \exp(-s_{it}^2/2)$. In this case the variational mean field, eq. (4.22) is given by $f(\gamma, \lambda) = \gamma/(1+\lambda)$. Thus, the variational mean field theory predicts

$$\chi_{ii'}^{t,\mathrm{NMF}} = \delta_{ii'} \frac{\partial \langle S_{it} \rangle}{\partial h_{it}} = 1/(1 + \lambda_{it}) = 1/(1 + J_{ii}) \; . \tag{4.31}$$

However, the linear response estimate eq. (4.29) gives $\chi_{ii'}^{t,\mathrm{LR}} = \left[(\boldsymbol{I} + \boldsymbol{J})^{-1}\right]_{ii'}$ and hence reconstructs the full covariance matrix identical with the exact result obtained by direct integration.

4.3.3 Adaptive TAP Approach

So far we have derived two different estimates of the covariance matrix from variational mean field theory: $\chi_{ii'}^{t,\mathrm{NMF}} = \delta_{ii'} \frac{\partial \langle S_{it} \rangle}{\partial h_{it}}$ and $\chi_{ii'}^{t,\mathrm{LR}} = \left[(\boldsymbol{\Lambda}_t + \boldsymbol{J})^{-1}\right]_{ii'}$. Obviously there is no guarantee that the two estimates are identical. Variational mean field theory is thus not self-consistent within linear response corrections. The adaptive TAP approach [Opper and Winther 2000c] on the other hand goes beyond the factorized trial distribution and requires self-consistency for the covariances estimated by linear response. This is achieved by introducing a set of IT additional mean field (or variational) parameters, the variances λ_{it} in the marginal distribution eq. (4.14), such that the diagonal term $\chi_{ii}^{t,\mathrm{TAP}}$ obeys

[4] It is noted that a Gaussian source prior is not suitable for doing source separation. We merely use it here to show that the linear response correction in this case recovers the exact result.

$$\frac{\partial \langle S_{it} \rangle}{\partial h_{it}} = \left[(\boldsymbol{\Lambda}_t + \boldsymbol{J})^{-1} \right]_{ii} , \tag{4.32}$$

where Λ_{it} and γ_{it} now depend upon λ_{it} through the relations

$$\Lambda_{it} = \left(\chi_{ii}^t \right)^{-1} - \lambda_{it} \tag{4.33}$$

$$\gamma_{it} = h_{it} - \sum_{i'} (J_{ii'} - \lambda_{i't} \delta_{ii'}) \langle S_{i't} \rangle \ . \tag{4.34}$$

To recover the variational mean field equations (4.30) and (4.20), we just let $\lambda_{it} = J_{ii}$. It is beyond the scope of this chapter to derive the adaptive TAP mean field theory. Instead, the reader is referred to section 3.3.4 for a derivation of the adaptive TAP approach valid for models with quadratic interactions and general variable prior. In chapter 5, the mean field equations for the noisy ICA model have been derived using the cavity approach instead of the KL variational bound approach. We have chosen to present and test the resulting theory here because it offers the most advanced (and hopefully the most precise) mean field approximation for this type of model.

4.4 Source Models

In this section we calculate for various source priors the variational mean m, eq. (4.22) and the derivative $\partial m / \partial \gamma$ needed for the linear response correction and adaptive TAP updates. The priors that we are considering are all chosen such that the variational mean can be calculated using tables of standard integrals, e.g. [Gradshteyn and Ryzhik 1980]. It turns out to be convenient to introduce the Gaussian kernel D with unit variance and its associated cumulative distribution function (cdf.) Φ in order to keep the following expressions of a manageable size, i.e.

$$D(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x^2 \right) \ , \qquad D'(x) = -x D(x) \tag{4.35}$$

$$\Phi(x) = \int_{-\infty}^{x} D(t) dt \ , \qquad \Phi'(x) = D(x) \ . \tag{4.36}$$

4.4.1 Summary of source priors

Table 4.1 summarizes the variational means and response functions corresponding to the priors described in this paper. It should be mentioned that this is by no means a complete list of all priors for which it is possible to calculate these quantities, e.g. the Rayleigh distribution is one such prior.

Source Prior	$p(s)$	Mean Function $m(\gamma,\lambda)=\langle S\rangle$	Response Func. $\frac{\partial\langle S\rangle}{\partial\gamma}$
Binary	$\frac{1}{2}\delta(s-1)+\frac{1}{2}\delta(s+1)$	$\tanh(\gamma)$	$1-\langle S\rangle^2$
Gaussian Mix.	eq. (4.37)	eqs. (4.39) & (4.41)	
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-s^2/2)$	$\gamma/(1+\lambda)$	$1/(1+\lambda)$
Heavy tail	not analytic	$\frac{\gamma}{\lambda}-\alpha\frac{\gamma}{\lambda\alpha+\gamma^2}$	$\frac{1}{\lambda}+\alpha\frac{\gamma^2-\lambda\alpha}{(\lambda\alpha+\gamma^2)^2}$
Uniform	$\frac{1}{b-a}\Theta(s-a)\Theta(b-s)$	eq. (4.65)	eq. (4.66)
Laplace	$\frac{1}{2}\exp(-\|s\|)$	eq. (4.43)	eq. (4.45)
Pos. Gauss	$\sqrt{\frac{2}{\pi}}\exp(-s^2/2)\Theta(s)$	eq. (4.60)	eq. (4.62)
Exponential	$\exp(-s)\Theta(s)$	eq. (4.47)	eq. (4.48)

Table 4.1. The variational mean and response function corresponding to various source priors. The three first rows describe source priors having negative, zero and positive kurtosis, respectively. The fourth row express non-negative priors. The step–function is defined as $\Theta(s)=1$ for $s>0$ and zero otherwise.

4.4.2 Mixture of Gaussians source prior

In this section we consider a general mixture of Gaussians, i.e.

$$p(s|\boldsymbol{\mu},\boldsymbol{\sigma})=\sum_{k=1}^{N_i}\pi_k p(s|\mu_k,\sigma_k)\ , \qquad s\in\mathbb{R} \tag{4.37}$$

where each of the N_i individual mixture components are parametrized by

$$p(s|\mu_k,\sigma_k)=\frac{1}{\sqrt{2\pi\sigma_k^2}}e^{-\frac{1}{2}(s-\mu_k)^2/\sigma_k^2}\ . \tag{4.38}$$

Using this source prior the generative ICA model becomes the independent factor analysis model proposed in [Attias 1999]. Since the main scope of this chapter is concerned with reliable inferring mean sufficient statistics with respect to the sources we will in contrary to [Attias 1999] always regard the source parameters as fixed, e.g. we are at no times adapting the source priors to data. However, it is straightforward to extend the proposed methodology to allow for this possibility, e.g. in a EM setting where the improved mean field solutions are being used in the posterior expectation of the complete log-likelihood.

Trivial but tedious calculations shows that the variational mean $m(\gamma,\lambda)$ of a mixture of Gaussians is given by

$$m(\gamma,\lambda)=\frac{\sum_{k=1}^{N_i}\kappa_k\frac{\gamma\sigma_k^2+\mu_k}{\lambda\sigma_k^2+1}e^{\xi_k}}{\sum_{k=1}^{N_i}\kappa_k e^{\xi_k}}\ , \tag{4.39}$$

where we have introduced

$$\kappa_k = \frac{\pi_k}{\sqrt{\lambda\sigma_k^2 + 1}} \ , \ \text{and} \ \xi_k = -\frac{1}{2}\left((\mu_k/\sigma_k)^2 - \frac{(\gamma\sigma_k + \mu_k/\sigma_k)^2}{\lambda\sigma_k^2 + 1}\right) . \quad (4.40)$$

The derivative with respect to γ is easy to obtain but are left out in the interest of space. For the special case of a mixture of two Gaussians ($N_i = 2$) with common variance σ^2 and means $\mu_k = \pm\mu$ we get

$$m(\gamma, \lambda) = \frac{1}{\lambda\sigma^2 + 1}\left(\gamma\sigma^2 + \mu \tanh(\frac{\gamma\mu}{\lambda\sigma^2 + 1})\right) \ . \quad (4.41)$$

For $\sigma^2 = 0$ and $\mu = 1$, we recover the variational mean for the binary source $p(s) = \frac{1}{2}\delta(s-1) + \frac{1}{2}\delta(s+1)$: $m = \tanh(\gamma)$. This particular choice of the bi-Gaussian source distribution (eq. 4.41) which is also known as the symmetric Pearson mixture density, was proposed in [Girolami 1998] as a simple way of achieving a negative kurtosis (sub-Gaussian) density function. To become familiar with the m-function and its derivative, consider the variational mean of the bi-Gaussian with $\sigma^2 = 1$ shown in figure 4.2(a,b) for two values of μ; namely $\mu = 1$, for which the density function is uni-modal and $\mu = 4$ for which the density function is significantly bimodal. We see that the more bimodal the source distribution is the more compact the region of high curvature becomes. By introducing additional mixture components it is possible to form the region of high curvature, which is illustrated in figure 4.2(g) in the case of a mixture of five Gaussians.

4.4.3 Laplace source prior

Although a sub-Gaussian distribution may be a reasonable source prior for some applications, e.g. telecommunications (discrete priors, see e.g. [van der Veen 1997]) or processing of functional magnetic resonance images [Petersen et al. 2000], there are, however, a large class of interesting real world signals, such as speech, which have heavier tails than the Gaussian distribution. We therefore need to consider source priors which have positive kurtosis (super-Gaussian). One such choice which have been widely used in the ICA community is $p(s) = 1/(\pi \cosh s)$ [Bell and Sejnowski 1995; MacKay 1996]. Using this prior, however, it is not possible to calculate the variational mean analytically. Instead, we consider the Laplace or double exponential distribution which is very similar. The Laplace density is given by

$$p(s) = \frac{\eta}{2} e^{-\eta|s|} \ , \qquad s \in \mathbb{R}, \ \eta > 0. \quad (4.42)$$

where $\eta > 0$ is the decay rate of the Laplacian. The variational mean can be calculated as

$$m(\gamma, \lambda) = \frac{1}{\sqrt{\lambda}} \frac{\xi_+\kappa_+ + \xi_-\kappa_-}{\kappa_+ + \kappa_-} \ , \quad (4.43)$$

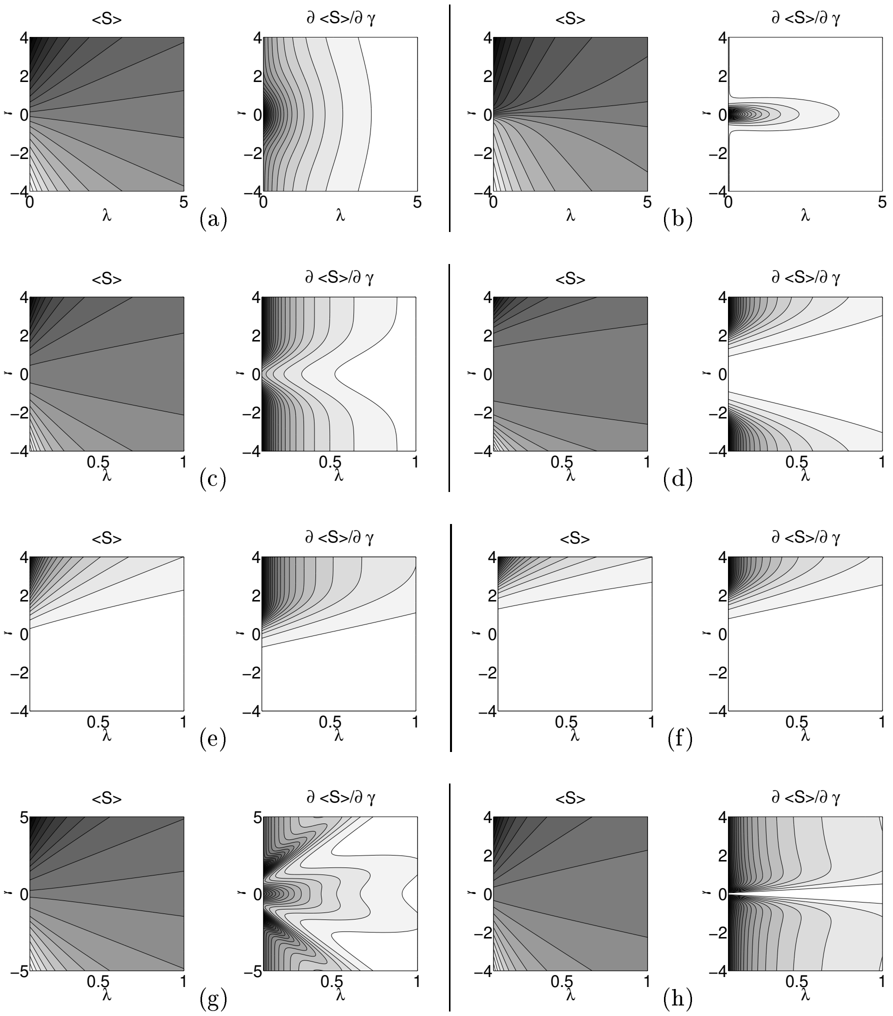


Fig. 4.2. Shows the variational mean m (left row) and its derivative m' (right row) as a function of γ and λ. (a) and (b) shows the bi-Gaussian case with $\sigma^2 = 1$ for $\mu_i = \pm 1$ and $\mu_i = \pm 4$, respectively. (c) and (d) shows the Laplacian prior for decay rates $\eta = 1/2$ and $\eta = 2$, respectively. (e) and (f) shows the exponential prior for decay rates $\eta = 1/2$ and $\eta = 2$, respectively.; (g) shows the variational mean m and the derivative m' of a mixture of five Gaussian with mixing proportions $\pi_i = 1/5$, means $\mu_i = \{-4, -1, 0, 1, 4\}$ and standard deviations $\sigma_i = \{1, 2, 4, 2, 1\}$. (h) shows the heavy tailed prior eq. (4.51) with $\alpha = 1$.

where we have introduced

$$\xi_\pm = \frac{\gamma \mp \eta}{\sqrt{\lambda}} \ , \qquad \text{and} \qquad \kappa_\pm = \Phi(\pm\xi_\pm) D(\xi_\mp) \ . \tag{4.44}$$

Using eqs. (4.35) and (4.36), the derivative is found to be

$$\frac{\partial m}{\partial \gamma} = \frac{1}{\lambda}\left(1 - \xi_-\xi_+ + D(\xi_+)D(\xi_-)\frac{\xi_+ - \xi_-}{\kappa_+ + \kappa_-} + \sqrt{\lambda}\frac{(\xi_+\kappa_- + \xi_-\kappa_+)}{(\kappa_+ + \kappa_-)} m\right) \ . \tag{4.45}$$

Figure 4.2(c,d) shows the variational mean and its derivative for a slowly decaying ($\eta = 0.5$) and a fast decaying ($\eta = 2$) Laplacian prior. The Laplacian prior have, contrary to the bi-Gaussian source, its region of high curvature for numerical large values of γ.

4.4.4 Exponential source prior

Some application domains naturally restrict the possible range of the hidden sources and the mixing matrix due to the physical interpretation of these quantities in the generative model. This is for instance the case when the measured signal is known to be a positive superposition of latent counting numbers or intensities. Positivity constrains are relevant, e.g., in "parts based representations" of natural images, deconvolution of the power spectrum of nuclear magnetic resonance (NMR) spectrometers and latent semantic analysis in text mining [Lee and Seung 1999]. In this section we consider the exponential source prior parameterized by

$$p(s) = \eta e^{-\eta s} \ , \qquad s \in \mathbb{R}_+ \ , \ \eta > 0 \tag{4.46}$$

where $\eta > 0$ is the decay rate of the exponential density. The variational mean and response function associated to this source prior are given by

$$m(\gamma, \lambda) = \frac{1}{\sqrt{\lambda}} \frac{\xi\Phi(\xi) + D(\xi)}{\Phi(\xi)} \tag{4.47}$$

$$\frac{\partial m}{\partial \gamma} = \frac{1}{\lambda} + \frac{D(\xi)}{\sqrt{\lambda}\Phi(\xi)} m \ , \tag{4.48}$$

where we, as for the Laplacian case, have introduced

$$\xi = \frac{\gamma - \eta}{\sqrt{\lambda}} \ . \tag{4.49}$$

Figure 4.2(e,f) shows the variational mean and its derivative for the exponential source prior. It is verified that the exponential variational mean is non-negative. At this point we will make some short remarks on some algorithmic issues when the normal cdf. Φ appears in the denominator of the variational mean. Special care has to be taken when $\xi \to -\infty$, e.g. when

$\gamma - \eta < 0$ and λ is small, i.e. for small self-interactions. Using l'Hospital's rule together with eqs. (4.35) and (4.36), it is seen that

$$\frac{D(\xi)}{\Phi(\xi)} \to -\xi \quad \text{for} \quad \xi \to -\infty \ , \tag{4.50}$$

which in turn implies that the variational mean $m \to 0$ and its derivative $(\partial m / \partial \gamma) \to 1/\lambda$ for $\xi \to -\infty$. In section 4.5.4, we will use this prior to learn a set of sparse localized basis functions in images. The source priors considered until now are just some examples of priors where the variational mean can be computed analytically. However, in section 4.7 we simply state some additional examples of priors for which this calculation can be carried out analytically.

4.4.5 Power law tail prior

In the previous sections we have only considered source priors for which it was possible to carry out the integration eq. (4.22) analytically. For arbitrary source priors, however, the one dimensional integral may be solved using standard approaches for numerical integration. Alternatively, we could simply use the insight gained in the previous sections, where we considered the functional form of the variational mean of various source priors, to come up with computationally tractable m functions directly. To give an example of this, we will construct an m which for large $|\gamma|/\sqrt{\lambda}$ corresponds to a distribution with a power law tail $p(s) \propto |s|^{-\alpha}$ for $|s|$ large. In this limit the integral in eq. (4.22) is dominated by its saddle-point. The saddle-point value of s is $s_0 = \frac{\gamma}{2\lambda}(1 + \sqrt{1 - \frac{4\alpha\lambda}{\gamma^2}}) \approx \frac{\gamma}{\lambda} - \frac{\alpha}{\gamma}$. This gives the behavior of the mean function for large γ. We can now straightforwardly construct a mean function that has this asymptotic behavior and is well-defined for small values of γ,

$$m(\gamma, \lambda) = \frac{\gamma}{\lambda} - \frac{\alpha\gamma}{\alpha\lambda + \gamma^2} \ . \tag{4.51}$$

Figure 4.2(h) shows the heavy tail m-function as a function of γ and λ. Figure 4.3 shows for a fixed $\lambda = 1$ the variational mean and derivative for some of the unconstrained source priors considered so far. For $\gamma \to \infty$, the Gaussian and the uniform (improper) prior give respectively the the lower and upper value for m for the priors considered. The variational means and derivatives for the priors considered in this chapter are summarized in the table in section 4.4.1.

4.5 Simulations

In this section we compare the performance of the different mean field approaches described in the previous sections, i.e. NMF, LR correction and

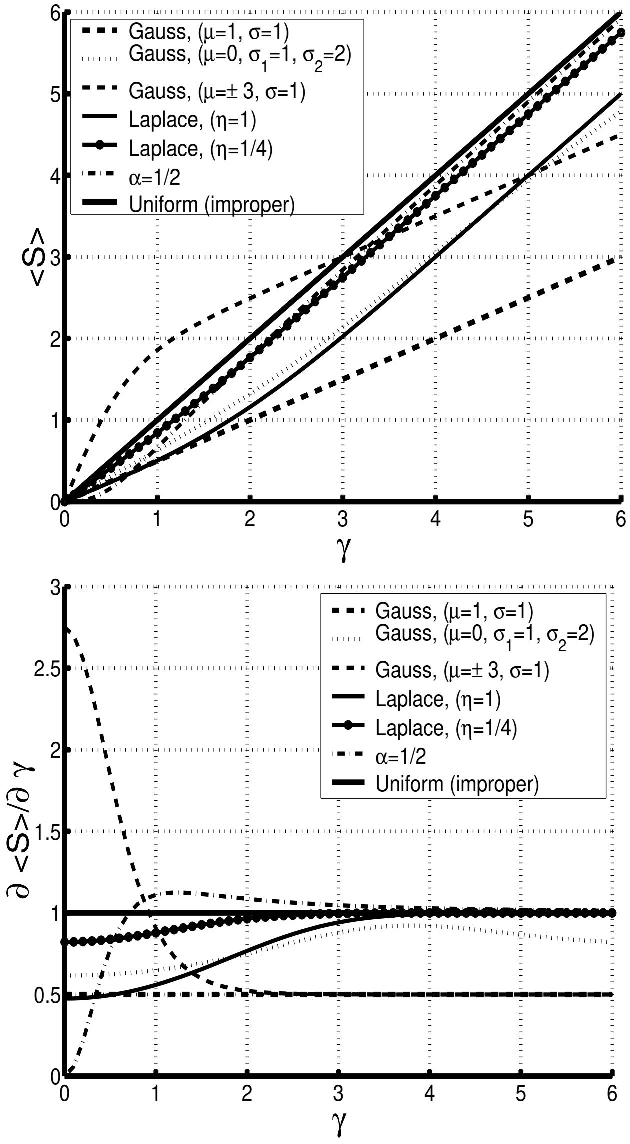


Fig. 4.3. Shows the variational mean (top) and derivative (lower) as a function of γ for various source priors and fixed $\lambda = 1$. From top to bottom the legends are; [- -] Gaussian with unit mean and variance; [$\cdots$] Mixture of two Gaussians with 0 mean and std. 1 and 2; [- -] Mixture of two Gaussians with unit variance and mean at ± 3; [—] and [-•-] Laplacian with $\eta = 1$ and $\eta = 1/4$, respectively; [-·-] Heavy tail with $\alpha = 1/2$; [- -] Uniform (improper) distribution.

adaptive TAP. To begin with, we conduct two experiments with artificial generated data. The source priors used in these experiments are equal to the source prior which generated the dataset. We consider both the complete case in which 2 binary sources are mixed into 2 sensors and the overcomplete case of 3 continuous sources mixed into 2 sensors. Finally, we apply the linear response corrected mean field approach to perform ICA on two real world datasets; namely speech signals and parts of the MNIST handwritten digit database.

4.5.1 Synthetic binary sources in a complete setting

Independent component analysis of binary sources has been considered e.g. in data transmission using binary modulation schemes such as MSK or biphase codes [van der Veen 1997]. Here, we consider a binary source $S = \{\pm 1\}$ with prior distribution $p(s) = \frac{1}{2}[\delta(s-1) + \delta(s+1)]$. In this case we recover the well known mean field equations $\langle S \rangle = \tanh(\gamma)$. Figure 4.4(a) shows the column vectors of the mixing matrix and 1000 samples generated from the ICA generative model using a fairly low noise variance, $\sigma^2 = 0.3$. Ideally, the noise-less measurements would consist of the four combinations (with sign) of the columns in the mixing matrix. However, due to the noise, the measurements will be scattered around these prototype observations (shown as $+$ in figure 4.4(a)). Figure 4.4(b) shows, for each of the mean field approaches, the variance as a function of iteration number. At these moderate noise variances an improvement in the convergence rate is obtained by using the linear response corrected mean field solution. The adaptive TAP approach, on the other hand, is seen to have a slower convergence rate and only a marginal improvement in the estimated noise variance and mixing matrix is obtained. This is due to the fact that this approach is critically sensitive to how well the variational parameters have been determined.

Figure 4.4(c,d,e) shows, for the different mean field approaches, the trajectories of the fix-point iterations. All the methods use the same initial conditions ('$\times$') and the final point in the trajectory is marked '$\circ$'. The dashed lines are $+/-$ the column vectors of the true mixing matrix. In this case there is no significant difference in the mixing matrix estimated using the different mean field approaches.

We now increase the noise variance to $\sigma^2 = 1$. In this case it is hard to identify the prototype signals from the measured data (see figure 4.5(a)). The naive mean field approach fails in recovering the mixing matrix. Figure 4.5(c) shows that one of the directions in the mixing matrix vanishes during the fixed-point iterations which in turn result in the noise variance being overestimated (see figure 4.5(b)). However, the linear response corrected mean field approach and adaptive TAP recover the true mixing matrix.

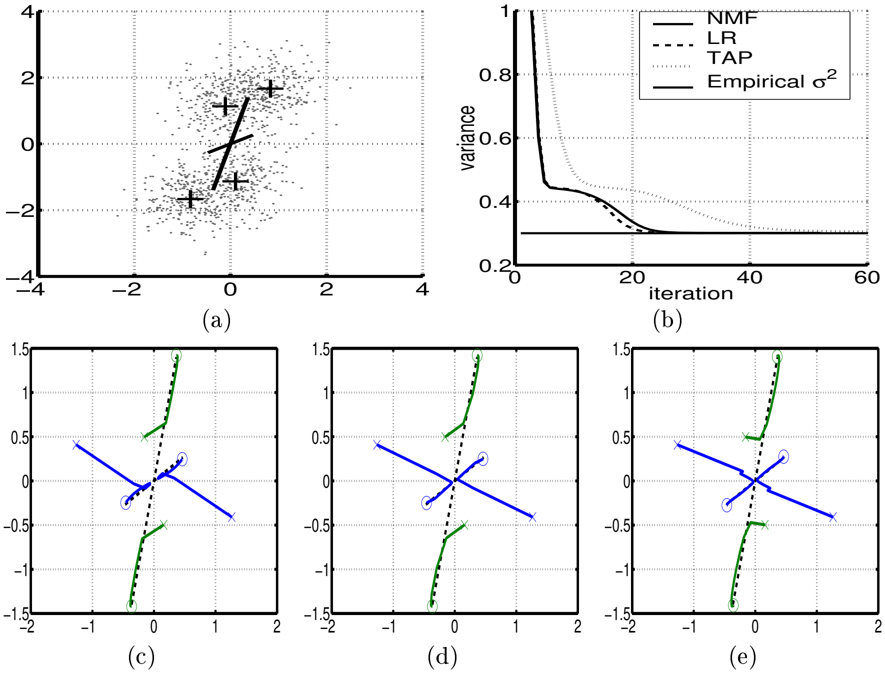


Fig. 4.4. Binary source recovery for a low noise variance, $\sigma^2 = 0.3$. (a) Shows 1000 measurements (scatter plot), $+/-$ the column vectors of the true mixing matrix (the solid axis) and the measurement prototypes ($+$) for the noise-less case. (b) Shows the estimated variance for NMF, LR and TAP as a function of iterations. The thick solid line is the true empirical noise variance. The empirical variance is the variance of the 1000 random noise contributions. The trajectories of the fixed-point iteration using (c) NMF, (d) LR and (e) adaptive TAP. The initial condition is marked '$\times$' and the final point '$\circ$'. The dashed lines are the true mixing matrix.

4.5.2 Continuous sources in an overcomplete setting

In this section the problem is to recover more sources than sensors; in particular we consider mixing 3 source into 2 sensors. The source used in this experiment is the symmetric Pearson mixture eq. (4.41) with $\mu = 1$. A total of 2000 samples were generated from the generative model (see figure 4.6(a)) and the three mean field approaches were used to learn the mixing matrix. The trajectories plot in figure 4.6(c) shows that the naive mean field approach fails in recovering the mixing matrix. Similar to the binary case with high variance, one of the directions in the mixing matrix vanishes (see figure 4.6). Only the dominant direction in the data-space is captured whereas the two remaining directions collapse into one “mean” direction. However, both the linear response corrected and the adaptive TAP mean field approaches succeed in estimating the mixing matrix. We will restrict ourselves to the LR approach in the next real world examples since NMF has turned out to fail in

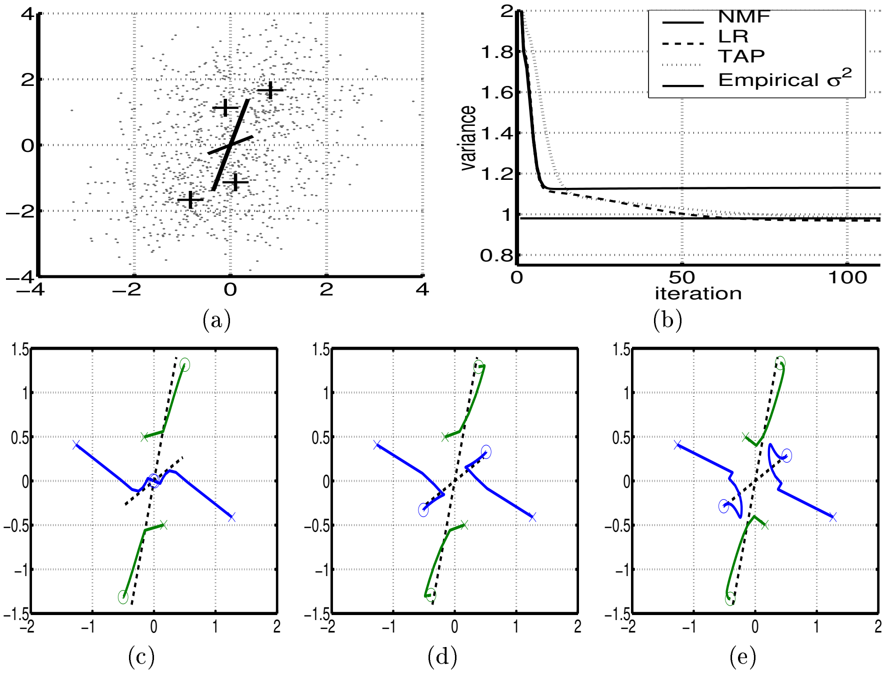


Fig. 4.5. Binary source recovery for a high noise variance, $\sigma^2 = 1$. (a) Shows 1000 measurements (scatter plot), $+/-$ the column vectors of the true mixing matrix (the solid axis) and the measurement prototypes ($+$) for the noise-less case. (b) Shows the estimated variance for NMF, LR and TAP as a function of iterations. The thick solid line is the true empirical noise variance. The trajectories of the fixed-point iteration using (c) NMF, (d) LR and (e) adaptive TAP. The initial condition is marked '$\times$' and the final point '$\circ$'. The dashed lines are the true mixing matrix.

some cases and TAP is considerably more computationally expensive while giving comparable performance.

4.5.3 Separating 3 speakers from 2 microphones

In this section we consider the problem of separating three speakers from two microphones. At hand we have the three original speech signals, each having a duration of 1 second and sampled at 8 kHz. The speech signals are then instantaneously linearly mixed into 2 microphones. Figure 4.7(a) shows a scatter plot of the 8000 samples in the measurement (microphone) space. The fact that natural speech has a heavy tailed distribution makes this over-complete problem somewhat easier in the sense that the hidden directions of the mixing matrix reveal themselves clearly in the scatter plot. The linear response corrected mean field approach was used in performing ICA with the computationally tractable variational mean eq. (4.51) with $\alpha = 1$. The

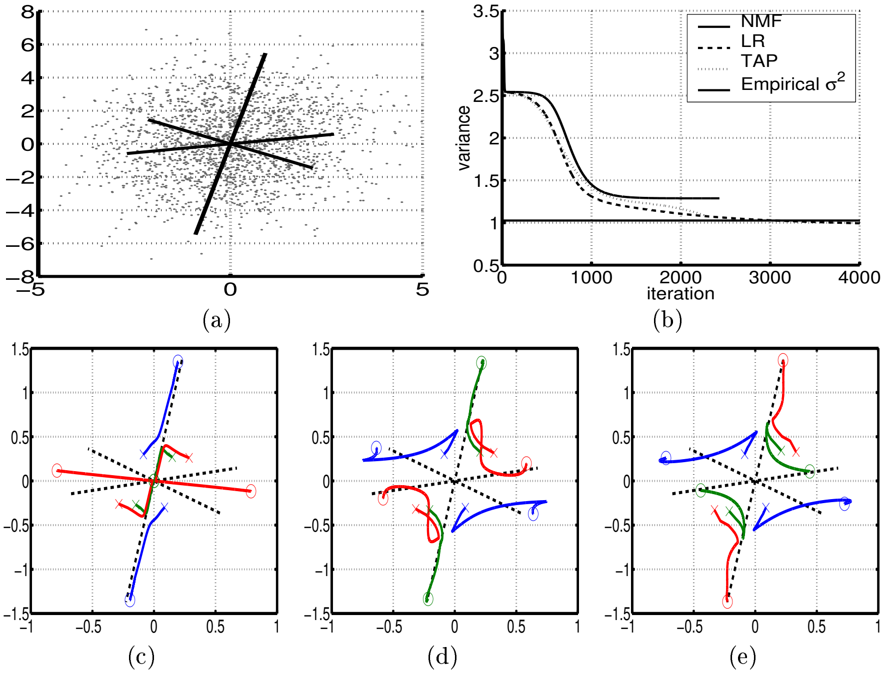


Fig. 4.6. Overcomplete continuous source recovery with $\sigma^2 = 1$. (a) Shows 2000 measurements (scatter plot), $+/-$ the column vectors (4 times axis) of the true mixing matrix (the solid axis). (b) Shows the estimated variance for NMF, LR and TAP as a function of iterations. The thick solid line is the true empirical noise variance. The trajectories of the fix-point iteration using (c) NMF, (d) LR and (e) adaptive TAP. The initial condition is marked '$\times$' and the final point '$\circ$'. The dashed lines are the true mixing matrix.

initial mixing matrix was randomly picked (shown as the dotted axis in figure 4.7(a)). Figure 4.7(b) shows the convergence of the algorithm in terms of the angle between the estimated directions and the true directions (the dashed lines in figure 4.7(a)). Figure 4.7(a) shows that the algorithm converges rapidly to a mixing matrix which is very close to the one that actually mixed the speech signals. Figure 4.8 shows each of the inferred sources plotted against each of the true sources. We see that the three recovered sources are nicely correlated with exactly one of the true sources and (more or less) uncorrelated with the remaining sources. Notice that any relabelling of the sources and corresponding perturbation of the columns of the mixing matrix leaves the solution of the ICA problem invariant.

4.5.4 Local feature extraction with sparse positive encoding

In this section we apply the linear response corrected ICA algorithm to the problem of finding a small set of localized images representing parts of the

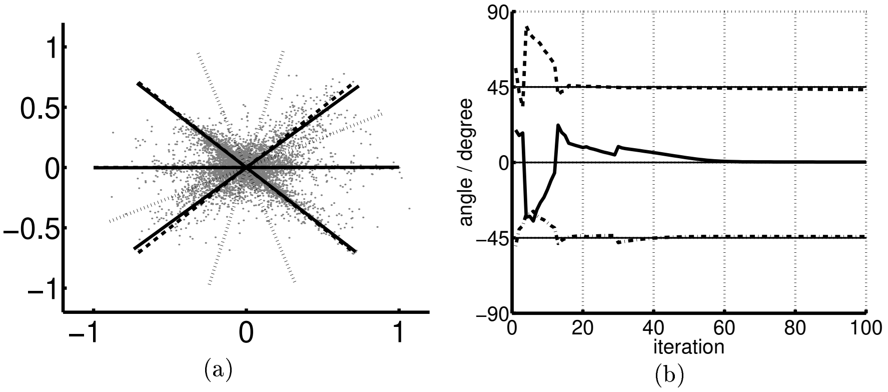


Fig. 4.7. Overcomplete speech separation (3-in-2) using the heavy tailed m-function eq. (4.51) with $\alpha = 1$; see figure 4.2(h). (a) scatter plot of 1 sec. of the mixed speech (@8 kHz), the true $\boldsymbol{A}$ (dashed lines), the initial $\boldsymbol{A}$ (black dotted) and the estimated $\boldsymbol{A}$. (b) shows the estimated angle as a function iteration. The horizontal lines illustrate true angles at 0 and ± 45 degrees.

digit images in the MNIST handwritten digit database. For illustration purposes we will only consider a small sub-set of the database, namely the first 500 cases of the handwritten digit "3". Figure 4.9(a) shows 25 examples from this sub-set of the dataset. As mentioned already in section 4.4.4 it is natural to consider positive constraints on latent variables (say pixels) when dealing with images. However, such constraints are usually ignored by most of the commonly used preprocessing models e.g. the principal component analysis (PCA) generative model which simply amounts to sequentially finding orthogonal directions (components) with maximum variance in the data space. Ignoring such constraints is problematic since for an unconstrained model to yield positive digit images there has to be an interaction between positive and negative regions in different components and it is therefore not obvious what the set of components represents visually.

To illustrate these points we conduct two ICA experiments using the exponential prior $p(s) \propto e^{-s}$, $s \in \mathbb{R}_+$. In the first experiment we do not constrain the mixing matrix whereas in the second experiment the mixing matrix is constrained to be positive. For both experiments we assume that there are 25 hidden images. Figure 4.9(c) shows the 25 hidden images obtained using ICA with positively constrained sources but unconstrained mixing matrix. Although the sources in this case are positively constrained, the fact that hidden images are allowed to be subtracted in order to obtain a positive image leads to non-local hidden images which are hard to interpret visually. Figure 4.9(d) shows the 25 hidden images obtained by performing ICA which enforces the positive constraint on the mixing matrix. In this case the hidden images clearly represent local features, in particular the different handwriting styles/strokes in the various parts of the written digit. For comparison we

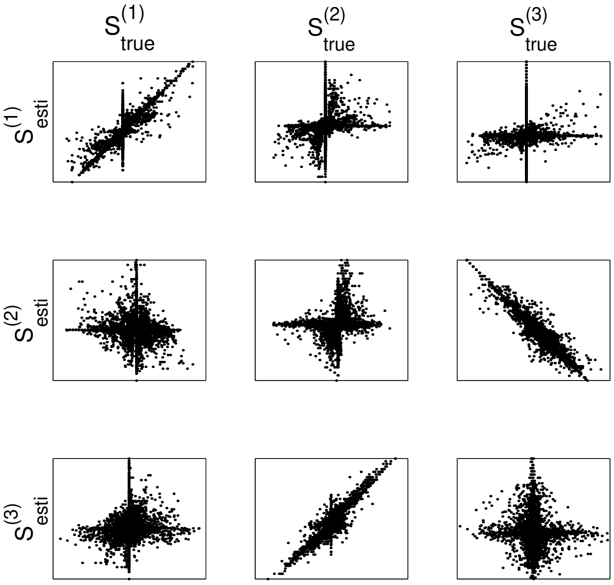


Fig. 4.8. Overcomplete speech separation (3-in-2) using the heavy tailed m-function eq. (4.51) with $\alpha = 1$. Shows the scatter plots of the ICA estimated sources $S^{(i)}_{esti}$ versus the true sources $S^{(i)}_{true}$, $i = 1, 2, 3$.

show the eigenimages obtained by the square loss version of the non-negative matrix factorization algorithm (NNMF) of [Lee and Seung 2001]. Although the positively constraint ICA model and the NNMF roughly yield the same results one major drawback of the non-negative matrix factorization is, however, its lack of probabilistic interpretation.

4.6 Algorithmic recipe

In table 4.2, we give an EM recipe for solving the mean field equations and the equations for the mixing matrix and the noise covariance. It is indicated in the table which equations that have been used. Here, we have given the equations for the adaptive TAP approach. Linear response theory is obtained by omitting the updating step for λ_{it}, i.e. by setting $N_{\boldsymbol{\lambda}} := 0$. Furthermore, setting $\chi^t_{ii'} := \delta_{ii'} m'(\gamma_{it}, \lambda_{it})$ instead of $\boldsymbol{\chi}^t := (\boldsymbol{\Lambda}_t + \boldsymbol{J})^{-1}$ leads to the naive mean field algorithm.

In the table, we have given the update rule for the non-negative mixing matrix eq. (4.13). To get to the unconstrained mixing matrix, the unconstrained update rule eq. (4.10) should be used.

Note that we use a greedy update step for all variables but the expectations $\langle \boldsymbol{S} \rangle$. Especially adaptive TAP is quite sensitive to the choice of the

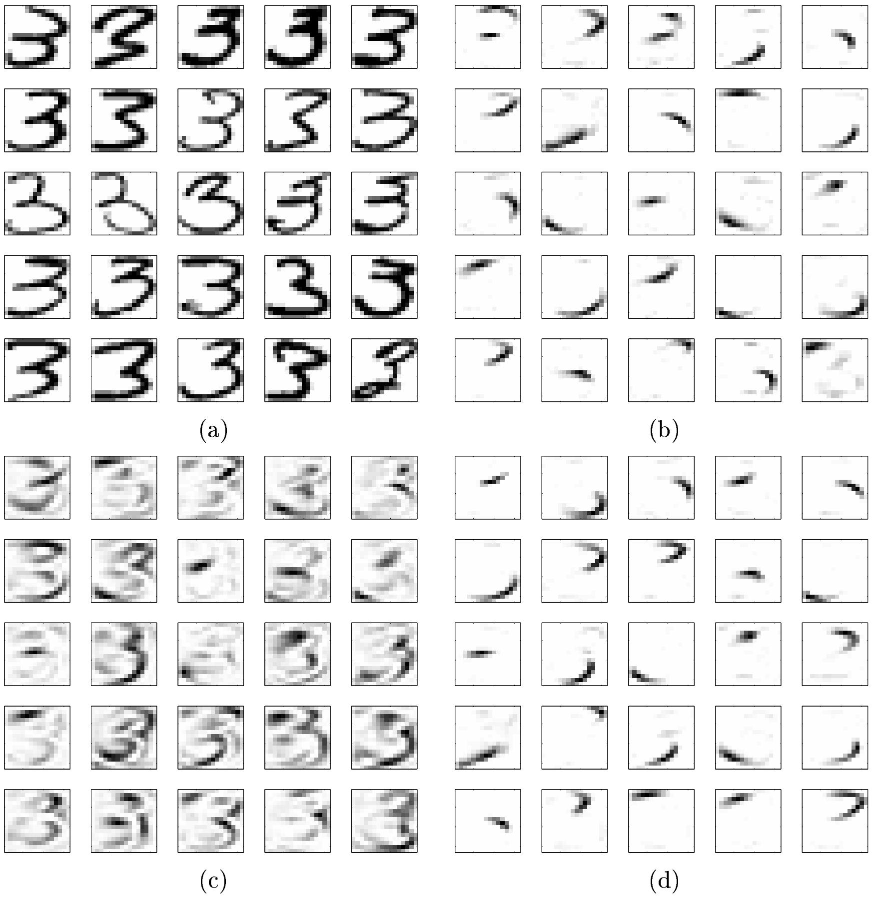


Fig. 4.9. Feature extraction of 500 cases of the handwritten digit "3" from the MNIST handwritten digit database. (a) shows 25 cases from the dataset. (b) shows the eigenimages obtained by the square loss version of the non-negative matrix factorization algorithm. Eigenimages obtained by the noisy ICA model using a exponential prior with $\eta = 1$ and (c) unconstrained mixing matrix and (d) positive constrained mixing matrix.

learning rate η. It is therefore made adaptive such that it is increased with a factor of 1.1 if the sum of the squared deviations $\sum_{i,t} |\delta\langle S_{it}\rangle|^2$ decreases compared to the previous update. Otherwise it is decreased with a factor 2. Our experience with the TAP equations also indicates that running with variable number of updates of $\langle \boldsymbol{S} \rangle$ could be helpful. However, in the simulations described here we kept the number of iterations fixed.

Initialization: Eqs. (4.16),(4.17) and (4.21)
    $\boldsymbol{J} := \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A}$
    $\boldsymbol{h} := \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}$
    $\langle \boldsymbol{S} \rangle := 0$ (or small random values if 0 is a fixed point)
    for $m := 1, \ldots, M$ and $t := 1, \ldots, N$:
        $\lambda_{mt} := J_{mm}$
    endfor
    $N_{\langle \boldsymbol{S} \rangle} := 20, N_{\boldsymbol{\lambda}} := 10, N_{\boldsymbol{A}} := 10, N_{\boldsymbol{\Sigma}} := 1, \text{ftol} := 10^{-5}$
do:
    Expectation-step:
        for $N_{\langle \boldsymbol{S} \rangle}$ iterations, eqs. (4.34) and (4.22)
            for $m := 1, \ldots, M$ and $t := 1, \ldots, N$:
                $\gamma_{mt} = h_{mt} - \sum_{m'} (J_{mm'} - \lambda_{m't} \delta_{mm'}) \langle S_{m't} \rangle$
                $\delta \langle S_{mt} \rangle := f(\gamma_{mt}, \lambda_{mt}) - \langle S_{mt} \rangle$
            endfor
            $\langle \boldsymbol{S} \rangle := \langle \boldsymbol{S} \rangle + \eta \delta \langle \boldsymbol{S} \rangle$
        endfor
        for $N_{\boldsymbol{\lambda}}$ iterations, eqs. (4.33) and (4.32)
            for $m := 1, \ldots, M$ and $t := 1, \ldots, N$:
                $\Lambda_{mt} := \lambda_{mt} + \frac{1}{f'(\gamma_{mt}, \lambda_{mt})}$
            endfor
            for $m := 1, \ldots, M$ and $t := 1, \ldots, N$:
                $\delta \lambda_{mt} := \frac{1}{\left[(\boldsymbol{\Lambda}_t + \boldsymbol{J})^{-1} \right]_{mm}} - \frac{1}{f'(\gamma_{mt}, \lambda_{mt})}$
                $\lambda_{mt} := \lambda_{mt} + \delta \lambda_{mt}$
            endfor
        endfor
        for $t := 1, \ldots, N$, eq. (4.29)
            $\boldsymbol{\chi}^t := (\boldsymbol{\Lambda}_t + \boldsymbol{J})^{-1}$
        endfor
    Maximization-step
        for $N_{\boldsymbol{A}}$ iterations, eq. (4.13) or (4.10)
            for $d := 1, \ldots, D$ and $m := 1, \ldots, M$:
                $\delta A_{dm} := \frac{[\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \langle \boldsymbol{S} \rangle^T]_{dm}}{[\boldsymbol{\Sigma}^{-1} \boldsymbol{A} \langle \boldsymbol{S} \boldsymbol{S}^T \rangle]_{dm} + \alpha A_{dm} + \beta} A_{dm} - A_{dm}$
                $A_{dm} := A_{dm} + \delta A_{dm}$
            endfor
        endfor
        for $N_{\boldsymbol{\Sigma}}$ iterations, eq. (4.9)
            $\delta \boldsymbol{\Sigma} := \frac{1}{N} \langle (\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S})(\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S})^T \rangle - \boldsymbol{\Sigma}$
            $\boldsymbol{\Sigma} := \boldsymbol{\Sigma} + \delta \boldsymbol{\Sigma}$
        endfor
        $\boldsymbol{J} := \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A}$
        $\boldsymbol{h} := \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}$
while $\max(|\delta \langle S_{mt} \rangle|^2, |\delta \lambda_{mt}|^2, |\delta A_{dm}|^2, |\delta \Sigma_{dd'}|^2) > \text{ftol}$

Table 4.2. Pseudo-code for the mean field ICA algorithms.

4.7 Some additional analytical source priors

In this section we derive the variational mean and response function for some additional analytical source priors which have not been directly used in this chapter. We show these calculations in some details since they are of the same type as the one we carried out in deriving the variational mean of the sources in section 4.4.

4.7.1 Positively constrained Gaussian source prior

Calculating the variational mean eq. (4.22) in general involves the calulation of an intergral of the form

$$\int dsp(s)e^{-\frac{1}{2}\lambda s^2+\gamma s} \ , \tag{4.52}$$

where $p(s)$ is the source prior. The source priors considered in this chapter are all of such a form that this integral can reparameterized into a integral over a Gaussian kernel. For this reason it is useful to have at hand an expression for the integral of a Gaussian kernel, i.e.

$$\int_{-\infty}^{x} dse^{-\frac{1}{2}\lambda s^2+\gamma s} = (\sqrt{2\pi}D(\frac{\gamma}{\sqrt{\lambda}}))^{-1}\int_{-\infty}^{x} dse^{-\frac{1}{2}\lambda(s-\frac{\gamma}{\lambda})^2} \tag{4.53}$$

$$= (\sqrt{\lambda}\sqrt{2\pi}D(\frac{\gamma}{\sqrt{\lambda}}))^{-1}\int_{-\infty}^{\xi} dse^{-\frac{1}{2}s^2} \tag{4.54}$$

$$= \frac{\Phi(\xi)}{\sqrt{\lambda}D(\frac{\gamma}{\sqrt{\lambda}})} \ , \tag{4.55}$$

where $\xi = \sqrt{\lambda}(x - \gamma/\lambda)$. The first equality follows from completing squares and introducing the Gaussian probability density function, eq. (4.35). The second equality follows by changing the integration variable whereas the final equality follows by introducing the Gaussian cdf., eq. (4.36). We can now calculate the following integral

$$\int_{0}^{+\infty} dse^{-\frac{1}{2}\lambda s^2+\gamma s} = \int_{-\infty}^{+\infty}(\cdot) - \int_{-\infty}^{0}(\cdot) = \frac{1-\Phi(-\frac{\gamma}{\sqrt{\lambda}})}{\sqrt{\lambda}D(\frac{\gamma}{\sqrt{\lambda}})} = \frac{\Phi(\frac{\gamma}{\sqrt{\lambda}})}{\sqrt{\lambda}D(\frac{\gamma}{\sqrt{\lambda}})} \ . \tag{4.56}$$

Suppose we are interested in calculating the variational mean of a density having eq. (4.56) as partition function. It is remembered that any factor of proportionality independent of γ is not needed in calculating the variational mean, i.e.

$$m(\gamma, \lambda) = (\frac{\Phi}{D})^{-1}\frac{\Phi'_{\gamma}D - \Phi D'_{\gamma}}{D^2} = (\frac{\Phi}{D})^{-1}\frac{D^2/\sqrt{\lambda} + \gamma/\lambda\Phi D}{D^2} \tag{4.57}$$

$$= \frac{\gamma}{\lambda} + \frac{D(\frac{\gamma}{\sqrt{\lambda}})}{\sqrt{\lambda}\Phi(\frac{\gamma}{\sqrt{\lambda}})} \ . \tag{4.58}$$

We can now return to the problem of calculating the variational mean of a positively constrained Gaussian parameterized by

$$p(s|\mu, \sigma) \propto e^{-\frac{1}{2}(s-\mu)^2/\sigma^2} \ , \qquad s \in \mathbb{R}_+ \ , \tag{4.59}$$

where μ and σ^2 are the mean and variance, respectively. Multiplying the source prior onto the Gaussian kernel and identifying terms it is seen that the product can be written as a Gaussian with $\lambda := \lambda + 1/\sigma^2$ and $\gamma := \gamma + \mu/\sigma^2$. Substituting back into eq. (4.58) we directly obtain the variational mean

$$m(\gamma, \lambda) = \frac{\gamma + \mu/\sigma^2}{\lambda + 1/\sigma^2} + \frac{1}{\sqrt{\lambda + 1/\sigma^2}} \frac{D(\kappa)}{\Phi(\kappa)} \ , \tag{4.60}$$

where we have introduced

$$\kappa = \frac{\gamma + \mu/\sigma^2}{\sqrt{\lambda + 1/\sigma^2}} \tag{4.61}$$

and the response function can be readily derived

$$\frac{\partial m}{\partial \gamma} = \frac{\mu/\sigma^2}{\lambda + 1/\sigma^2} \left(1 - \kappa \frac{D(\kappa)}{\Phi(\kappa)} - \left(\frac{D(\kappa)}{\Phi(\kappa)} \right)^2 \right) \ . \tag{4.62}$$

We now turn to consider the variational mean and the response function associated to the uniform source prior.

4.7.2 Uniform source prior

In this section we consider the uniform prior parametrized by

$$p(s) = \frac{1}{b-a} \ , \qquad s \in [a; b] \ , \tag{4.63}$$

where $b \geq a$. By reusing the calculations made in section 4.7.1 we directly obtain

$$\int_a^b ds e^{-\frac{1}{2}\lambda s^2 + \gamma s} = \int_{-\infty}^b (\cdot) - \int_{-\infty}^a (\cdot) = \frac{\Phi(\kappa_b) - \Phi(\kappa_a)}{\sqrt{\lambda} D(\frac{\gamma}{\sqrt{\lambda}})} \ , \tag{4.64}$$

where $\kappa_x = \sqrt{\lambda}(x - \gamma/\lambda) = \sqrt{\lambda}x - \frac{\gamma}{\sqrt{\lambda}}$. Here, we have again left out the normalizing constant since it is of no importance in the calculation of the variational mean

$$m(\gamma, \lambda) = \frac{\gamma}{\lambda} + \frac{1}{\sqrt{\lambda}} \frac{D(\kappa_a) - D(\kappa_b)}{\Phi(\kappa_b) - \Phi(\kappa_a)} \ , \tag{4.65}$$

and the response function

$$\frac{\partial m}{\partial \gamma} = \frac{1}{\lambda}\left(1 + \frac{\kappa_a D(\kappa_a) - \kappa_b D(\kappa_b)}{\Phi(\kappa_b) - \Phi(\kappa_a)} - \left(\frac{D(\kappa_a) - D(\kappa_b)}{\Phi(\kappa_a) - \Phi(\kappa_b)}\right)^2\right) . \qquad (4.66)$$

This and the previous section showed some illustrative examples of the calculation needed in deriving the variational mean and response functions for the source priors considered in this chapter.

4.8 Discussion

In this chapter, we have presented a probabilistic (Bayesian) approach to ICA. Sources are estimated by their posterior mean while maximum a posteriori estimates are used for the mixing matrix and the noise covariance matrix. By this procedure we derived an EM-type algorithm. The expectation step is carried out using different mean field (MF) approaches namely variational (also known as ensemble learning or naive MF), linear response and adaptive TAP. The MF theories produce estimates of posterior source correlations of increasing quality. These are needed for the maximization step in the estimate for the mixing matrix and the noise covariance matrix.

The importance of a good estimate of correlations is seen for specific examples where in fact the simplest variational approach fails. The general applicability of the formalism and its MF implementation is demonstrated on local feature extraction in images (using non-negative mixing matrix and source priors) and in overcomplete separation of speech (using heavy tailed source priors). The good performance of the mean field approach supports the belief that we get fair estimates of the posterior means and covariances. However, a rigorous test requires either explicit numerical integration which is possible only for low dimensional problems or Monte Carlo sampling (which may also be inaccurate in complex cases).

In the following, we will discuss a number of possible extensions of this work. One obvious extension is the modeling of temporal correlations. The most general formulation of the model with temporal correlation leads to the consideration of the junction tree algorithm. We are currently working on a mean field algorithm for online belief propagation on the junction tree [Højen-Sørensen et al. 2001a].

Optimization of the hyperparameters of the prior can be performed by trivially extending the current EM algorithm. The mean field approach can also be used to derive leave-one-out estimators [Opper and Winther 2000b; Opper and Winther 2000c] that can be used both for optimization of hyperparameters and model selection. Model selection can also be performed using the (approximate mean field) likelihood of a test set. In chapter 5, the optimization of the hyperparameters of the prior and the model selection problem is considered in an analysis of functional neuroimages.

Finally, it could be interesting to relax some of the basic requirement of the model. Firstly, that of statistical independence of the sources. Our formalism can be extended to treat a priori Gaussian correlations between (the non-Gaussian) sources. We should be able to estimate these correlations effectively by for example the linear response technique. Secondly, the model can be extended to nonlinear mixing by for example introducing a sigmoidal squashing of the mixed signal. This situation can also, with some increase in the computational complexity be, included in the mean field framework [Opper and Winther 2000c].

5. Analysis of Functional Neuroimages

The low signal-to-noise ratio and the many possible sources of variability makes recording from non-invasive functional neuroimaging techniques a most challenging data analysis problem. In this chapter we present a computationally efficient mean field algorithm for noisy independent component analysis (ICA) with adaptive binary sources for exploratory analysis of functional magnetic resonance imaging (fMRI) data. The number of hidden sources is determined using the Bayesian information criterion (BIC) in which the TAP free energy is used as an approximation to the likelihood. The developed algorithm is applied to both an artificial data set and a set of functional neuroimages from a visual activation study. In chapter 4 the starting point of the derivation of the mean field algorithm for probabilistic ICA was the KL variational bound. In this chapter, however, we derive the mean field algorithm by using the cavity approach which leads to the adaptive TAP algorithm examined experimentally in chapter 4.

5.1 Introduction

Functional magnetic resonance imaging (fMRI) is the common name for a large selection of non-invasive techniques that enables indirect measures of neuronal activity in the working human brain. Common for these techniques are that they utilize the MRI technique to detect and measure the regionally localized physiological changes which accompany neuronal activation. The most common fMRI technique is based on an image contrast induced by temporal shifts in the relative concentration of oxyhemoglobin and deoxyhemoglobin; this is known as the blood oxygenation level dependent (BOLD) contrast . The working model which relates to neuronal activity to the measured BOLD contrast goes as follows [Bandettini and Wong 1998]. An increase in neuronal activity causes local vasodilatations which in turn causes an increase in blood flow; the so-called *hemodynamic response*. This results in an excess of oxygenated hemoglobin beyond the metabolic need, thus reducing the proportion of paramagnetic deoxyhemoglobin in the vasculature. This in turn leads to a reduction in susceptibility differences in the vicinity of veinules, veins and red blood cells within veins, which thereby causes an increase in spin coherence and therefore an increase in the measured signal in $T2/T2^*$ weighted sequences. For the uninitiated, a nice discussion on how brain metabolism relates to the BOLD signal can be found in [Barinaga 1997]. The typical fMRI activation study consists of maintaining a healthy volunteer in controlled mental states; typically a baseline control state and an active task state. The temporal course of the baseline/activation paradigm is denoted the *reference function*. The BOLD signal is the regional hemodynamic response to focal neuronal activation. Being dispersed in both space and time, the hemodynamic response severely confounds the interpretation of the neuronal activation from the BOLD signal. Needless to say, the neuronal response may itself consist of multiple components besides the activity induced by the stimulus.

Bandettini et al. [Bandettini et al. 1993] analyzed the correlation between a binary reference function and the BOLD signal. Lange and Zeger [Lange and Zeger 1997] discuss a parameterized hemodynamic response adapted by a least squares procedure whereas multivariate strategies have been pursued in [Worsley et al. 1997] and [Hansen et al. 1999]. Several explorative strategies have been proposed for finding spatio-temporal activation patterns without explicit reference to the activation paradigm. McKeown et al. [McKeown et al. 1998] used the independent component analysis algorithm of [Bell and Sejnowski 1995] to identify *spatially* independent patterns and found several types of activations including components with "transient task related" response, i.e., responses that could not simply be accounted for by the paradigm; such components would typically not be identified by a simple principal component analysis (PCA) since they would tend not to be orthogonal to the task. Other authors have argued for identifying *temporally* independent patterns, e.g. [Hansen 2000]; a comparative study of temporal and spatial ICA approaches for analyzing fMRI were carried out in [Petersen et al. 2000]. While previous ICA approaches succeeded in finding task related components, the ICA schemes applied were not well-matched to the binary on/off character of the stimulus (see also [Petersen et al. 2000] for further discussion of this point). Hence, in this chapter we will focus on the more appropriate binary source distribution assumption. Furthermore, it is interesting to note that when analyzing the signals for spatially independent components, the binary source assumption corresponds to another popular approach for exploratory analysis of fMRI, namely wave form clustering. The binary spatial component can be viewed as the *binary association* of pixels with the corresponding time course; see e.g. [Goutte et al. 99] for use of clustering techniques for fMRI analysis. A extensive discussion of the statistical limitations in functional neuroimaging can be found in [Petersson et al. 1999].

The ICA technology invoked here for analyzing fMRI makes use of advanced mean field (or alternatively variational) methods to carry out approximative posterior inference in the generative ICA model; this is described in greater detail in chapter 4. To the best of our knowledge all generative ICA models which until now have been used for analyzing functional neuroimages have been assuming unconstrained continuous hidden sources. An elaborate list of references of such ICA algorithms and applications can be found in e.g. [Lee 1998; Girolami 2000].

The chapter is organized as follows. In section 5.2 and 5.3, we develop a computationally efficient algorithm for learning and inference in a noisy ICA model with adaptive binary $\{0,1\}$-sources, based on the cavity mean field approach of [Opper and Winther 2000c]. In section 5.4, the number of hidden sources is determined by the Bayesian information criterion (BIC) using the TAP free energy as an approximation to the likelihood. In section 5.5 and 5.6 we show some results on an artificial data set and a real fMRI data set, respectively. We end this chapter with a discussion in section 5.7.

5.2 The generative model for noisy ICA

Spatial and temporal ICA of functional neuroimages are both based on the decomposition of the spatio-temporal observations in terms of a sum of pairwise "outer-products" of a set of characteristic images and time series. Here we follow [Hansen 2000; Højen-Sørensen et al. 2001b] and consider noisy mixtures. For simplicity we present the theory for the case of temporal ICA, hence assuming that the observation can be considered as a sum of characteristic images activated by a set of corresponding independent times series. However, the choice is arbitrary and in the experimental evaluation we use the theory for both spatial and temporal analysis of fMRI.

As for the model from chapter 4, the generative model considered in this chapter assumes the D observations $\boldsymbol{x}_t = \{x_{dt}\}$ at time $t = 1, \ldots, N$ to be an instantaneously mixing of I hidden independent sources $\boldsymbol{s}_t = \{s_{it}\}$ corrupted with additive white Gaussian noise $\boldsymbol{\varepsilon}_t$ with covariance $\boldsymbol{\Sigma}$, i.e.

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{\epsilon}_t \ , \tag{5.1}$$

where $\boldsymbol{A}$ is the mixing matrix. It is useful to consider the canonical parameterization of the likelihood of the model parameters $\boldsymbol{\Omega}_L = \{\boldsymbol{A}, \boldsymbol{\Sigma}\}$ and sources $\boldsymbol{s}_t$,

$$p(\boldsymbol{x}_t|\boldsymbol{\Omega}_L, \boldsymbol{s}_t) = p(\boldsymbol{x}_t|\boldsymbol{J}, \boldsymbol{\theta}_t, \boldsymbol{s}_t) = Z_{L,t}^{-1} e^{-\frac{1}{2}\boldsymbol{s}_t^T \boldsymbol{J} \boldsymbol{s}_t + \boldsymbol{\theta}_t^T \boldsymbol{s}_t} \ . \tag{5.2}$$

where $\log Z_{L,t} = \frac{1}{2} \log \det 2\pi\boldsymbol{\Sigma} + \frac{1}{2}\boldsymbol{x}_t^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t$ is the log partition function, and we have introduced the interaction-matrix $\boldsymbol{J}$ and the data dependent external field $\boldsymbol{\theta}_t$ given respectively by

$$\boldsymbol{J} = \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} \qquad \text{and} \qquad \boldsymbol{\theta}_t = \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t \ . \tag{5.3}$$

The aim of independent component analysis is to recover the unknown quantities; the sources $\boldsymbol{S} = \{\boldsymbol{s}_t\}$, the mixing matrix $\boldsymbol{A}$ and the noise covariance $\boldsymbol{\Sigma}$

given an observed set of data $\boldsymbol{X} = \{\boldsymbol{x}_t\}$. This can be done using the generalized EM algorithm in which the expected sufficient statistics $\langle \boldsymbol{s}_t \rangle$ and $\langle \boldsymbol{s}_t \boldsymbol{s}_t^T \rangle$ are computed based on an approximation to the posterior density of the hidden sources. The quality of the expected sufficient statistics depends on the mean field method used in approximating the source posterior. A comparative study of the solution to the inference problem provided by a naive mean field (variational) approach and the more advanced linear response correction and adaptive TAP approach was carried out in [Højen-Sørensen et al. 2001c] in context of the noisy ICA model. In contrast to [Højen-Sørensen et al. 2001c] we will in this chapter use the cavity approach [Opper and Winther 2000c] to obtain a set of mean field equations as well as an approximation to the probability of the observed data, i.e. the minus free energy. Since this chapter deals with recovering hidden sources from functional neuroimages, i.e. very large data sets, we will, however, due to the computational complexity of the linear response corrections and the adaptive TAP approach only be considering the expected sufficient statistics provided by the naive mean field approach.

5.3 Cavity mean field and adaptive TAP for noisy ICA

We obtain the set of mean field equations for the noisy ICA model following the adaptive TAP approach of [Opper and Winther 2000c] which considers probabilistic models of the type

$$p(\boldsymbol{s}_t) \propto \rho(\boldsymbol{s}_t) e^{\frac{1}{2}\boldsymbol{s}_t^T \tilde{\boldsymbol{J}} \boldsymbol{s}_t + \boldsymbol{\theta}_t^T \boldsymbol{s}_t} \ , \tag{5.4}$$

where the interaction-weights are symmetric, $\tilde{J}_{ij} = \tilde{J}_{ji}$, and $\tilde{J}_{ii} = 0$ such that all self-interactions are contained in the single variable constraint $\rho(\boldsymbol{s}_t) = \prod_i \rho(s_{i,t})$. Hence, for the noisy ICA we first have to remove all self-interaction term in the likelihood into the single variable constraint. For notational simplicity consider a particular example $\boldsymbol{x} = \{x_i\}$, i.e. we will in the following suppress the subscript t. Furthermore, let $\boldsymbol{\Omega}_S = \{\boldsymbol{\mu}_i\}$ denote the parameters of the source model consisting of I hidden sources parametrized by $\boldsymbol{\mu}_i$ and let $\boldsymbol{\Omega} = \{\boldsymbol{\Omega}_L, \boldsymbol{\Omega}_S\}$ denote the entire set of model parameters. The adaptive TAP approach was derived in section 3.3.4 for models with probability densities given by eq. (5.4). In context of the probabilistic ICA model the distribution of interest is the source posterior

$$p(\boldsymbol{s}|\boldsymbol{\Omega}, \boldsymbol{x}) \propto p(\boldsymbol{s}) Z_L^{-1} e^{-\frac{1}{2}\boldsymbol{s}^T \boldsymbol{J} \boldsymbol{s} + \boldsymbol{\theta}^T \boldsymbol{s}} \ . \tag{5.5}$$

Hence, all we need to do now is to express $\rho(\boldsymbol{s})$ and $\tilde{\boldsymbol{J}}$ in terms of $p(\boldsymbol{s})$ and $\boldsymbol{J}$. Clearly, we have

$$\rho(\boldsymbol{s}) = p(\boldsymbol{s}) Z_L^{-1} e^{-\frac{1}{2}\boldsymbol{s}^T \mathrm{diag}(\boldsymbol{J}) \boldsymbol{s}} \tag{5.6}$$

$$\tilde{\boldsymbol{J}} = \mathrm{diag}(\boldsymbol{J}) - \boldsymbol{J} \ . \tag{5.7}$$

Hence, using the cavity approach we directly get the following approximation to the marginal source posterior

$$p(s_i|\boldsymbol{\Omega},\boldsymbol{x}) \propto p(s_i)e^{\frac{1}{2}(V_i-J_{ii})s_i^2+(\theta_i+\gamma_i)s_i} \ , \tag{5.8}$$

where $p(s_i)$ is the prior distribution associated to the i'th source and the mean of the cavity field is given by

$$\gamma_i = -\sum_j (1-\delta_{ij})J_{ij}m_j - V_i m_i \ , \tag{5.9}$$

where V_i is the variance of the cavity field and m_i is the posterior mean of the hidden sources s_i. The first set of mean field equations are obtained directly from eq. (5.8) by noticing that θ_i acts as an external field from which all posterior cumulants of the sources can be obtained. Using the linear response theorem an improved estimate of the posterior source covariance $\boldsymbol{\chi}$ can be obtained

$$\boldsymbol{\chi}^{LR} = (\boldsymbol{\Lambda}+\boldsymbol{J})^{-1} \ , \tag{5.10}$$

where $\boldsymbol{J} = \{J_{ij}\}$ is the matrix of interaction-weights and

$$\boldsymbol{\Lambda} \equiv \mathrm{diag}\,(\Lambda_1,\Lambda_2,\ldots,\Lambda_N)\,, \quad \text{where} \quad \Lambda_i \equiv V_i + \left(\frac{\partial m_i}{\partial \theta_i}\right)^{-1} - J_{ii} \ . \tag{5.11}$$

In the adaptive TAP approach proposed by [Opper and Winther 2000c] the variance of the cavity field is estimated by requiring consistency in the posterior variance χ_{ii}^{MF} obtained by the naive mean field approach and the posterior variance χ_{ii}^{LR} obtained through linear response correction which provide the second set of mean field equations. To carry out any form of model selection we need to be able to compute the likelihood, i.e. the probability of the observed data given the model parameters $\boldsymbol{\Omega}$ associated to a model $\mathcal{M}$. This is clearly intractable since we have to marginalize the hidden sources which are coupled through the observed data. However, following section 3.3.4 we obtain an approximation to the (minus) log-likelihood given by the TAP free energy

$$\begin{aligned} -\log p(\boldsymbol{x}|\boldsymbol{\Omega}) = &-\sum_i \log \int ds_i p(s_i) e^{\frac{1}{2}(V_i-J_{ii})s_i^2+(\theta_i+\gamma_i)s_i} \\ &-\frac{1}{2}\boldsymbol{m}^T(\boldsymbol{J}-\mathrm{diag}(\boldsymbol{J}))\boldsymbol{m} + \log Z_L \\ &-\frac{1}{2}\sum_i V_i\chi_{ii} - \frac{1}{2}\log\det\boldsymbol{\chi} + \frac{1}{2}\sum_i \log\chi_{ii} \end{aligned} \tag{5.12}$$

Although the presented mean field theory is feasible for a large class of source priors (see chapter 4 for details) we will in this particular application restrict ourself to consider binary $\{0,1\}$–sources parameterized by

$$p(s_{i,t}) = \mu_i^{s_{i,t}}(1-\mu_i)^{1-s_{i,t}} \ , \tag{5.13}$$

where μ_i is the mean of the binary sources. The set of mean field equations $\{m_i\}$ are readily obtained by utilizing the external fields $\{\theta_i\}$, i.e.

$$m_i = \langle S_i \rangle = \frac{\partial}{\partial \theta_i} \log Z_i = \frac{\mu_i e^{\frac{1}{2}(V_i - J_{ii}) + (\theta_i + \gamma_i)}}{\mu_i e^{\frac{1}{2}(V_i - J_{ii}) + (\theta_i + \gamma_i)} + (1-\mu_i)} \ , \tag{5.14}$$

where Z_i is the partition function associated to the posterior $p(s_i|\boldsymbol{\Omega}, x_i)$, i.e. the likelihood $p(x_i|\boldsymbol{\Omega})$, obtained when using the eq. (5.8) as an approximation to the complete likelihood. The response function needed to improve the estimate for posterior source covariance is then given by

$$\frac{\partial m_i}{\partial \theta_i} = (1 - m_i) m_i \ . \tag{5.15}$$

In section 5.5 and 5.6 we will only consider expected sufficient statistics obtained by the naive mean field approach which is then used in the E-step to get improved estimates of the mixing matrix $\boldsymbol{A}$, noise covariance matrix $\boldsymbol{\Sigma}$ and source parameters $\{\boldsymbol{\mu_i}\}$. Using the naive mean field ansatz the approximation to the log-likelihood is readily obtained from the TAP free energy eq. (5.12), i.e.

$$\begin{aligned} \log p(\boldsymbol{x}|\boldsymbol{\Omega}) = & \sum_i \log \int ds_i p(s_i) e^{-\frac{1}{2} J_{ii} s_i^2 + (\theta_i + \gamma_i) s_i} \\ & + \frac{1}{2} \boldsymbol{m}^T (\boldsymbol{J} - \mathrm{diag}(\boldsymbol{J})) \boldsymbol{m} - \log Z_L \ . \end{aligned} \tag{5.16}$$

In section 5.4, this approximation to the log-likelihood is used together with the Bayesian Information Criterion (BIC) to determining the number of independent sources in the observed data.

5.4 Estimating the number of sources

The above development was predicated at ICA with a fixed number of sources. When applying the model to real world data the number of latent components is unknown. In [Hansen et al. 1999; Hansen 2000], the number of components were determined for PCA and ICA, respectively using test set methods, e.g. testing how well a fitted model on one set fMRI data generalizes to another independent set. This approach, however, can suffer from basic violations of the underlying statistical assumptions of stationarity across multiple runs. Here we suggest to use an approximate Bayesian approach. The *Bayesian information criterion* (BIC) [Schwarz 1978] is an approximation to the log marginal likelihood $p(\boldsymbol{X}|\mathcal{M})$, where $\mathcal{M}$ denotes the model. In context of the noisy ICA model eq. (5.1) the model $\mathcal{M}$ is simply determined by the number of hidden sources I. The log marginal likelihood is obtained by marginalizing the model parameters $\boldsymbol{\Omega}$, i.e.

$$p(\boldsymbol{X}|\mathcal{M}) = \int d\boldsymbol{\Omega} e^{f(\boldsymbol{\Omega})} p(\boldsymbol{\Omega}|\mathcal{M}) \ , \tag{5.17}$$

where $f(\boldsymbol{\Omega}) = \log p(\boldsymbol{X}|\boldsymbol{\Omega}, \mathcal{M})$ denotes the log likelihood associated to model $\boldsymbol{M}$. As the amount of the data, N, grows to infinity the log likelihood becomes more sharply peaked in the neighborhood around its maximum $\tilde{\boldsymbol{\Omega}}$. Hence, a reasonable approach is to evaluate the integral by *Laplace integration*, i.e. to approximate the log likelihood by

$$f(\boldsymbol{\Omega}) \approx f(\hat{\boldsymbol{\Omega}}) + \frac{1}{2}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}})^T \boldsymbol{H}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}) \ , \tag{5.18}$$

and evaluate the integral as

$$\begin{aligned} p(\boldsymbol{X}|\mathcal{M}) &\approx e^{f(\hat{\boldsymbol{\Omega}})} \int d\boldsymbol{\Omega} e^{\frac{1}{2}(\boldsymbol{\Omega}-\hat{\boldsymbol{\Omega}})^T \boldsymbol{H}(\boldsymbol{\Omega}-\hat{\boldsymbol{\Omega}})} p(\boldsymbol{\Omega}|\mathcal{M}) && (5.19)\\ &\approx e^{f(\hat{\boldsymbol{\Omega}})} p(\hat{\boldsymbol{\Omega}}|\mathcal{M}) \int d\boldsymbol{\Omega} e^{\frac{1}{2}(\boldsymbol{\Omega}-\hat{\boldsymbol{\Omega}})^T \boldsymbol{H}(\boldsymbol{\Omega}-\hat{\boldsymbol{\Omega}})} && (5.20)\\ &= \frac{e^{f(\hat{\boldsymbol{\Omega}})} p(\hat{\boldsymbol{\Omega}}|\mathcal{M})}{\sqrt{\det(-2\pi \boldsymbol{H})}} \ , && (5.21) \end{aligned}$$

where the second approximation follows from the assumption that the prior $p(\boldsymbol{\Omega}|\mathcal{M})$ can be regarded as constant in the vicinity of the sharp peak at $\hat{\boldsymbol{\Omega}}$ and the equality follows from assuming that $-\boldsymbol{H}$ is positive definite. The log marginal likelihood is then approximated by

$$\begin{aligned} \log p(\boldsymbol{X}|\mathcal{M}) &\approx \log p(\boldsymbol{X}|\hat{\boldsymbol{\Omega}}, \mathcal{M}) + \log p(\hat{\boldsymbol{\Omega}}|\mathcal{M}) - \frac{1}{2} \log \det(-2\pi \boldsymbol{H}) && (5.22)\\ &\approx \log p(\boldsymbol{X}|\hat{\boldsymbol{\Omega}}, \mathcal{M}) - \frac{|\boldsymbol{\Omega}|}{2} \log N \ , && (5.23) \end{aligned}$$

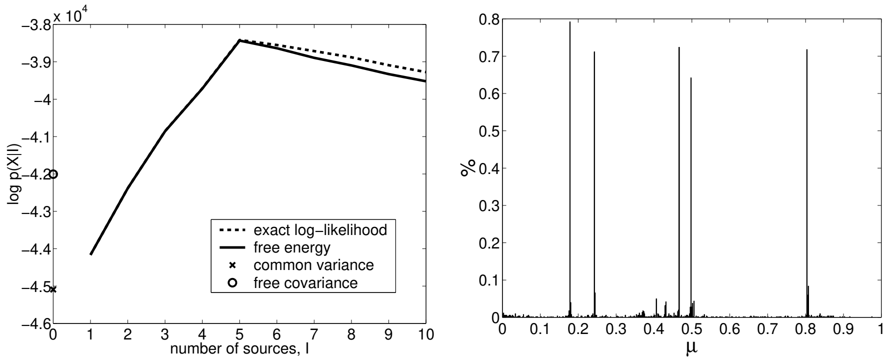
where $|\boldsymbol{\Omega}|$ denotes the number of model parameters associated to model $\mathcal{M}$ which in this case is the number of sources. The last approximation follows from retaining only those terms that increase with the sample size, i.e. the log-likelihood which increases linearly with N and $\log \det(-\boldsymbol{H})$ which increases as $|\boldsymbol{\Omega}| \log N$. The approximation eq. (5.23) is the Bayesian information criterion (BIC). We see that the BIC combine the likelihood with some penalty relating to the complexity of the model. Furthermore, it obviates the need for parameter priors although the derivation assumes that the prior is non-zero around $\hat{\boldsymbol{\Omega}}$. Notice that the BIC is only strictly valid for models with complete data, i.e. where the values of all variables are specified in each training case. In the presence of hidden variables the log-likelihood does not necessarily tend toward a peak as the sample size increases. In fact, it turns out that the model penalizing term in the BIC in that case is the rank of the Jacobian matrix of the transformation between the parameters of the network and the parameters of the observable variables [Geiger et al. 1998]. However, for the present model we have an approximation for the likelihood of the model parameters, i.e. we have been able to marginalize the hidden variables and hence taken into account the volume contribution arising from that integral. Hence, provided the TAP free energy yields a reasonable approximation to the likelihood we are in a perfectly valid position to use the BIC for complete data.

5.5 Analysis of an artificial data set

In this section we test the algorithm on a data set with known ground truth. A total of 500 samples, each consisting of 50 observations, was drawn from the generative model with 5 hidden sources with mean $\boldsymbol{\mu} = (.2, .25, .5, .5, .8)$ and noise covariance $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, where $\sigma^2 = 1$. In all the experiments presented in this chapter we have assumed diagonal noise covariance matrix with common variance σ^2. Figure 5.1(a) shows as a function of the number of hidden sources I, the log marginal likelihood $p(\boldsymbol{X}|I)$ computed using the BIC with both the exact likelihood and the approximation provided by the free energy. The log marginal likelihood for the hypothesis that no non-trivial independent components are present in the data is shown assuming Gaussian noise with free covariance and common variance. The log marginal likelihood shown in the figure is the average of 50 random parameter initializations. We see that the BIC is successful in detecting the right number of hidden sources. Although the free energy provides a lower bound on the log-likelihood, it is worth noticing that the free energy indeed provides a very accurate estimate of the log-likelihood when the data has more independent components than the generative model being fitted. However, when the data has fewer independent components than the fitted model the free energy tends to underestimate the log-likelihood significantly and hence also the log marginal likelihood. As mentioned in section 5.4 this happens when there are many likely configurations of the model parameters and the latent sources; this is especially the case when the data is fitted by a too flexible model. Figure 5.1(b) shows a histogram of the estimated source parameters $\boldsymbol{\mu}$ for 500 random parameter initializations. This figure clearly suggests that it is possible to recover the source parameters from the observed data.

5.6 Analysis of a fMRI data set

In this section we analyze a fMRI data set acquired during a visual activation study. The data set was acquired by Dr. Egill Rostrup, Danish Center of Magnetic Resonance Research. A single slice fMRI scan were acquired every 330 millisecond in a para-axial orientation parallel to calcarine sulcus



(a) (b)

Fig. 5.1. Analysis of an artificial data set with $D = 50$, $M = 5$ and $N = 500$. In this experiment the source parameters were $\boldsymbol{\mu} = (.2, .25, .5, .5, .8)$. (a) shows the log marginal likelihood as a function of the number of hidden sources. The log marginal likelihood is computed using both the exact likelihood and the free energy. At $I = 0$ the log marginal likelihood is calculated for a Gaussian with both a free covariance ($\circ$) and a diagonal covariance with common variance ($\times$). (b) shows the histogram of the estimated sources parameters from 500 random initializations. The histogram clearly reflects the true statistics of the underlying sources.

using $T2^*$-weighted EPI (@ 1.5 T). Each run of the experimental paradigm consisted of 30 scans of fixation, 30 scans of stimulation and 60 scans of post-stimulus fixation and was repeated 10 times. The visual activation consisted of an annular full-field checkerboard reversing at 8 Hz. This is a very strong stimulus of the primary visual areas and we expect these to respond with a notable on-off activation.

To reduce the computational cost we extracted the voxels belonging to the brain using a simple morphological masking procedure on the single slice. We perform spatial and temporal ICA for the resulting 4196x120 matrix of observations, both with varying number of components.

In figure 5.2 we show in panel (a) the log marginal likelihood as function of the number of components for *temporal* ICA. A generative ICA model with a single latent component is optimal in this case, and in panel (b) we show the inferred posterior mean of the source and the on-off binary reference function (with values 0 and 0.5 for clarity). The ICA time series shows a few scattered activations and a large contiguous activation beginning approximately 3 seconds (10 scans) after stimulus onset. This is consistent with typical hemodynamic delays found in primary visual cortex [Bandettini and Wong 1998]. The spatial pattern associated to the inferred source is presented in panels (c) and (d). In (c) we show the 2.5 % most positive (white) and negative (black) activation "hot spots" superimposed on an anatomical background which has the same spatial resolution as the data. In panel (d) we provide a quantitative representation of the spatial pattern. The spatial

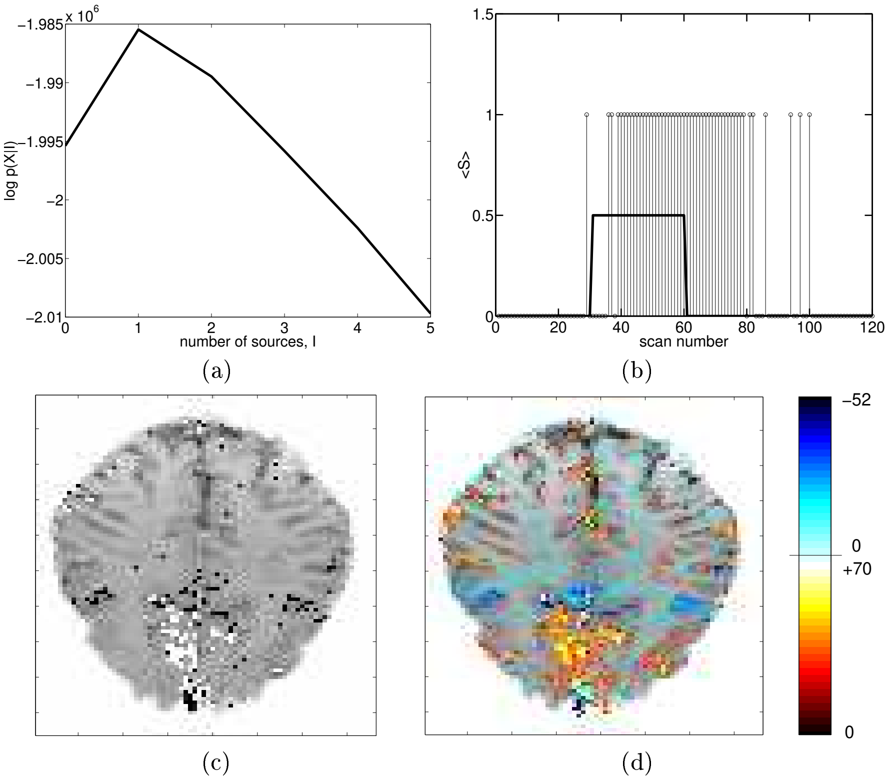


Fig. 5.2. Analysis of a fMRI data set using temporal binary ICA. (a) shows the log marginal likelihood as a function of the number of hidden sources. At $I = 0$ the log marginal likelihood is calculated for a Gaussian with common variance. (b) shows the experimental paradigm (in solid) and the inferred sources. (c) shows the .025 (black) and .975 (white) fractiles of the values in the eigenimage superimposed on an anatomical reference. (d) shows a quantitative representation of the spatial pattern retaining only the 0.1 and 0.9 fractiles of the eigenimage.

pattern is dominated by a large cluster of pixels in the primary visual areas. We note two relatively weak, but bilateral negative activations that could be auditive regions that are processing audible scanner noise when the subject is not attending to visual input, as suggested in [Petersen et al. 2000] for the same data set.

In figure 5.3 we present the results of searching for *spatially* independent components. In panel (a) we show the log marginal likelihood with nine being the most probable number of latent components. This is consistent with [McKeown et al. 1998] who found a high number of interpretable components using spatial ICA. In panel (b) and (c) we show the binary images and the associated time series. The time series have been globally post-normalized such that the maximal value is 1. Notice that some of the binary images are

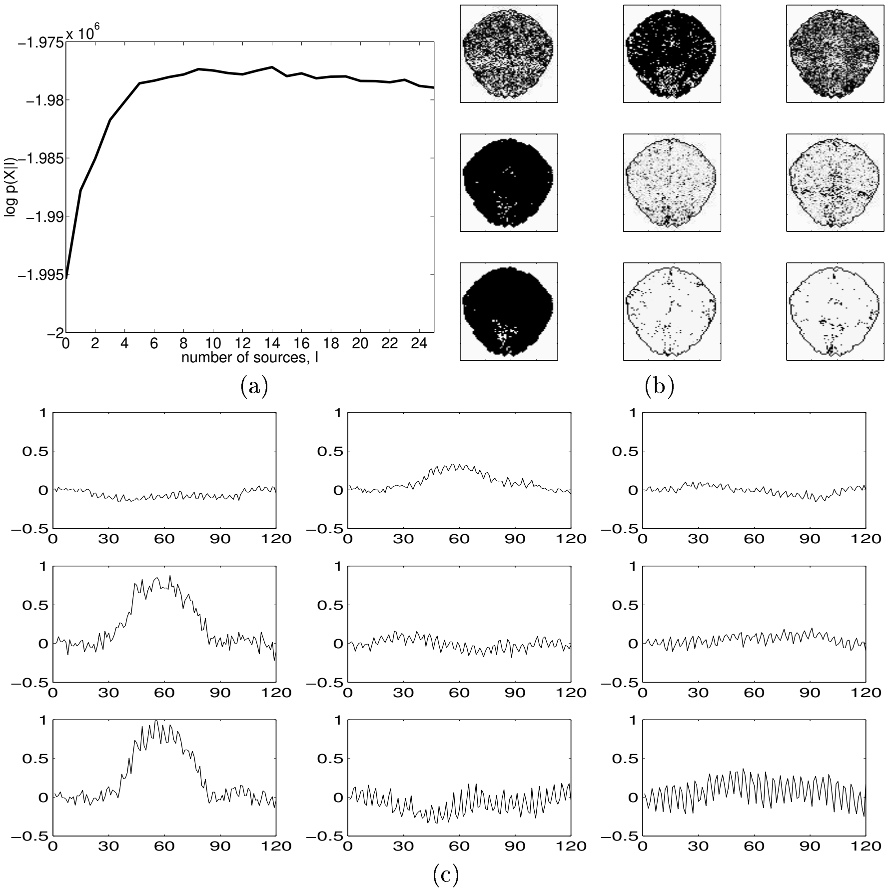


Fig. 5.3. Analysis of a fMRI data set using spatial binary ICA. (a) shows the log marginal likelihood as a function of the number of hidden sources. At $I = 0$ the log marginal likelihood is calculated for a Gaussian with common variance. The log marginal likelihood is the average of 30 random parameter initializations. (b) shows the posterior mean of the nine hidden binary images. (c) shows the corresponding responses associated to each of the nine hidden images.

mainly in the off-state, while others are mainly in the on-state. In particular, two components (4 and 7) have strong responses which are highly correlated with the stimulus. Furthermore, these two component time series are active only in the expected regions for visual stimulation. The weaker signals are confounds that are found globally (components 1,3,5,6,8,9) or locally (components 2).

5.7 Discussion

In this chapter we presented a general approach to ICA with binary sources and applied to a fMRI data set from a visual activation study. It was argued that a binary source assumption is indeed appropriate for explorative analysis of the on-off type of stimulation commonly used in fMRI activation studies. We proposed using the BIC as an approximation for the log marginal likelihood and used this to determine the number of hidden sources. We applied the scheme for both spatial and temporal ICA. Interestingly, the two models lead to rather different "optimal" descriptions. While both agree on the form of the image and time series for the activation components, the spatial ICA suggest to invoke eight additional independent components with weaker signal strengths in the optimal model. Future studies should be aimed at understanding of the significance of this observation.

6. Approximate Message Passing

In this chapter we describe a possible approach for carrying out approximate message passing in probabilistic models for which marginalization of some of the clique potentials turns out to be intractable. The approach is illustrated on a generative model for on-line classification as well as on a simple extension of the previously considered probabilistic ICA model which takes into account temporal correlations between sources. Common for these two probabilistic models is that they can be specified in terms of the DAG for the state-space model considered in section 2.4.2. Although the approach, at least in principle, is generally applicable, the two different examples suggest that the success of this methodology is highly dependent of the particular probabilistic model under consideration.

6.1 Moment passing scheme

The easiest way to illustrate the moment passing scheme is to apply it on a specific model. Figure 6.1 shows a fragment of the junction tree associated to the state-space model analyzed in section 2.4.2. With reference in this fragment, let us consider the forward-pass in which messages are being passed towards the root clique. Provided the separator potentials $\psi^*(x_t)$ and $\tilde{\psi}^*(x_{t+1})$ have already been updated, we can now update the separator potential $\psi^*(x_{t+1})$ using the marginal-propagation update eq. (2.15), i.e.

$$\psi^*(x_{t+1}) = \int dx_t \psi^*(x_t)\tilde{\psi}^*(x_{t+1})\phi(x_t, x_{t+1}) \; . \tag{6.1}$$

However, for general potentials it is not possible to analytically evaluate the integral of the product of separator and clique potentials. In this specific example, suppose that it is the updated separator potential $\psi^*(x_t)$ which renders an analytically evaluation of the integral impossible. Instead, provided the evaluation of the integral becomes doable, simply by replacing $\psi^*(x_t)$ with an appropriate Gaussian approximation, we could then carry out the approximate marginal-propagation update

$$\psi^*(x_{t+1}) = \int dx_t \psi_g^*(x_t)\tilde{\psi}^*(x_{t+1})\phi(x_t, x_{t+1}) \; , \tag{6.2}$$

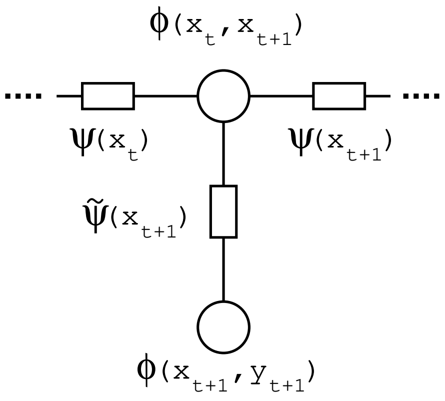


Fig. 6.1. Illustration of the moment passing scheme. Shows a fragment of the junction-tree associated to the state-space model. In this example all the separator potentials $\{\psi^*(x_t)\}$ are being approximated by Gaussians in the BKL sense.

where $\psi_g^*(x_t)$ is a Gaussian approximation to the separator potential $\psi^*(x_t)$. In order to proceed further with the message passing recursion we need to calculate a Gaussian approximation to the newly updated separator potential $\psi^*(x_{t+1})$. One way to obtain the mean and variance of this Gaussian approximation is to calculate the cumulant generating function

$$\mathcal{K}(\xi) = \log \int dx_{t+1} \psi^*(x_{t+1}) e^{\xi x_{t+1}}, \tag{6.3}$$

and find the first two cumulants, κ_1 and κ_2, using the relation

$$\kappa_k = \left[\frac{\partial^k}{\partial \xi^k} \mathcal{K}(\xi) \right] \Bigg|_{\xi=0} . \tag{6.4}$$

Besides of providing the mean and variance directly, another advantage of using the cumulant generating function instead of the moment generating function is that we do not need to know the normalization of the product of potentials. In the next section we illustrate this scheme on a model for on-line classification.

6.2 A generative model for on-line classification

Although the following generative model seems somewhat artificial it is closely related to a Bayesian approach for on-line learning known as *assumed-density filtering* (e.g. see [Opper 1998] or more recently [Minka 2001]). In the Bayesian approach for on-line learning, the cases in the training set are being processed sequentially in such a way that the posterior after one training example acts as the prior for the next. In this section we consider a generative model for on-line classification (with labels $y_t = \pm 1$) given by

$$p(\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_0) \tag{6.5}$$
$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma}) \tag{6.6}$$
$$p(y_t|\boldsymbol{x}_t) = \Phi(y_t \boldsymbol{w}^T \boldsymbol{x}_t) \ . \tag{6.7}$$

Since this model possesses the same conditional independence relations as the state-space model from section 2.4.2 we can reuse the junction-tree from figure 2.8. In that case the potentials are given by

$$\phi(\boldsymbol{x}_0, y_0) = \Phi(y_0 \boldsymbol{w}^T \boldsymbol{x}_0)\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_0) \tag{6.8}$$
$$\phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{x}_t, \boldsymbol{\Sigma}) \tag{6.9}$$
$$\phi(\boldsymbol{x}_t, y_t) = \Phi(y_t \boldsymbol{w}^T \boldsymbol{x}_t) \ . \tag{6.10}$$

In this section we will only consider the forward recursion since the backward recursion essentially follows along the same lines. Assume that we have already made a Gaussian approximation to the separator potential $\psi^*(\boldsymbol{x}_t)$; in other words we assume that we have at hand

$$\psi_g^*(\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}) \ . \tag{6.11}$$

The marginal-propagation update for the next separator potential $\psi(\boldsymbol{x}_{t+1})$ is then given by

$$\psi^*(\boldsymbol{x}_{t+1}) = \tilde{\psi}^*(\boldsymbol{x}_{t+1}) \int d\boldsymbol{x}_t \psi_g^*(\boldsymbol{x}_t)\phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) \tag{6.12}$$
$$= \tilde{\psi}^*(\boldsymbol{x}_{t+1})\mathcal{N}(\boldsymbol{\mu}_{t+1|t}, \boldsymbol{\Sigma}_{t+1|t}) \ , \tag{6.13}$$

where the mean $\boldsymbol{\mu}_{t+1|t}$ and the covariance matrix $\boldsymbol{\Sigma}_{t+1|t}$ of the one-step-ahead predictor can be computed by

$$\mathcal{N} = \left\langle \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})\mathcal{N}(\boldsymbol{x}_{t+1}; \boldsymbol{A}\boldsymbol{x}_t, \boldsymbol{\Sigma})\right\rangle_{1|\boldsymbol{x}_{t+1}} \tag{6.14}$$
$$= \left\langle \mathcal{N}_* \left(\begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t+1} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\Lambda}_{t|t}^T \boldsymbol{\mu}_{t|t} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_{t|t} + \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A} & -\boldsymbol{A}^T \boldsymbol{\Lambda} \\ -\boldsymbol{\Lambda}\boldsymbol{A} & \boldsymbol{I} \end{bmatrix} \right) \right\rangle_{1|\boldsymbol{x}_{t+1}} \tag{6.15}$$
$$= \mathcal{N}_*(\boldsymbol{x}_{t+1}; \boldsymbol{\gamma}_{t+1|t}, \boldsymbol{\Lambda}_{t+1|t}) \ . \tag{6.16}$$

Here we have made use of the canonical parameterization of the Gaussian density to find the canonical parameters for the one-step-ahead density

$$\boldsymbol{\gamma}_{t+1|t} = \boldsymbol{\Lambda}\boldsymbol{A} \left(\boldsymbol{\Lambda}_{t|t} + \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A}\right)^{-1} \boldsymbol{\Lambda}_{t|t}^T \boldsymbol{\mu}_{t|t} \tag{6.17}$$
$$\boldsymbol{\Lambda}_{t+1|t} = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{A} \left(\boldsymbol{\Lambda}_{t|t} + \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^T \boldsymbol{\Lambda} \ . \tag{6.18}$$

By applying the matrix inversion lemma eq. (A.9) we can obtain the corresponding mean and covariance matrix

$$\boldsymbol{\mu}_{t+1|t} = \boldsymbol{A}\boldsymbol{\mu}_{t|t} \tag{6.19}$$
$$\boldsymbol{\Sigma}_{t+1|t} = \boldsymbol{A}\boldsymbol{\Sigma}_{t|t}\boldsymbol{A}^T + \boldsymbol{\Sigma} \ , \tag{6.20}$$

which we alternatively could have derived directly by taking the conditional expectation and variance of $\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\varepsilon}$, where the conditioning is on the observed sequence $\{y_1, y_2, \ldots, y_t\}$ and $\boldsymbol{\varepsilon}$ is Gaussian noise with covariance matrix $\boldsymbol{\Sigma}$.

To complete the update of the separator potential $\psi(\boldsymbol{x}_{t+1})$ we need to consider the effect of multiplying the label driven potential $\tilde{\psi}^*(\boldsymbol{x}_{t+1})$ onto the Gaussian one-step-ahead predictor. Since this is obviously trivial for the case where the class label Y_{t+1} is not observed we only consider here the case where $Y_{t+1} = \dot{y}_{t+1}$. In this case the cumulant generating function for the updated separator potential is given by

$$\mathcal{K}(\boldsymbol{\xi}) = \log \int d\boldsymbol{x}_{t+1} \Phi(\dot{y}_{t+1}\boldsymbol{w}^T\boldsymbol{x}_{t+1}) \mathcal{N}(\boldsymbol{\mu}_{t+1|t}, \boldsymbol{\Sigma}_{t+1|t}) e^{\boldsymbol{\xi}^T\boldsymbol{x}_{t+1}} \tag{6.21}$$

$$= \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{\xi} + \boldsymbol{\mu}_{t+1|t}^T\boldsymbol{\xi} + \log L(\boldsymbol{\xi}) \ , \tag{6.22}$$

where

$$L(\boldsymbol{\xi}) = \int d\boldsymbol{x}_{t+1} \Phi(\dot{y}_{t+1}\boldsymbol{w}^T\boldsymbol{x}_{t+1}) \mathcal{N}(\tilde{\boldsymbol{\mu}}_{t+1|t}, \boldsymbol{\Sigma}_{t+1|t}) \tag{6.23}$$

$$= \int ds d\boldsymbol{x}_{t+1} \delta(s - \dot{y}_{t+1}\boldsymbol{w}^T\boldsymbol{x}_{t+1}) \Phi(s) \mathcal{N}(\tilde{\boldsymbol{\mu}}_{t+1|t}, \boldsymbol{\Sigma}_{t+1|t}) \ , \tag{6.24}$$

and $\tilde{\boldsymbol{\mu}}_{t+1|t} = \boldsymbol{\mu}_{t+1|t} + \boldsymbol{\Sigma}_{t+1|t}\boldsymbol{\xi}$. In eq. (6.24) we have introduced the new scalar variable $s = \dot{y}_{t+1}\boldsymbol{w}^T\boldsymbol{x}_{t+1}$ by making use of the delta-function trick. Since the distribution of this new variable is clearly Gaussian we can find the mean and variance by direct calculations

$$\mu_s = \langle s \rangle = \dot{y}_{t+1}\boldsymbol{w}^T \langle \boldsymbol{x}_{t+1} \rangle = \dot{y}_{t+1}\boldsymbol{w}^T\tilde{\boldsymbol{\mu}}_{t+1|t} \tag{6.25}$$

$$\langle s^2 \rangle = \boldsymbol{w}^T \langle \boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^T \rangle \boldsymbol{w} = \boldsymbol{w}^T(\boldsymbol{\Sigma}_{t+1|t} + \tilde{\boldsymbol{\mu}}_{t+1|t}\tilde{\boldsymbol{\mu}}_{t+1|t}^T)\boldsymbol{w} \tag{6.26}$$

$$\sigma_s^2 = \langle (s - \langle s \rangle)^2 \rangle = \langle s^2 \rangle - \langle s \rangle^2 = \boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{w} \ . \tag{6.27}$$

Hence we have

$$L(\boldsymbol{\xi}) = \int ds \Phi(s) \mathcal{N}(\mu_s, \sigma_s^2) \tag{6.28}$$

$$= \sqrt{\frac{\lambda_s}{2\pi}} \exp\left(-\frac{1}{2}\lambda_s\mu_s^2\right) \int ds \Phi(s) \exp\left(-\frac{1}{2}\lambda_s s^2 + \gamma_s s\right) \tag{6.29}$$

$$= \Phi\left(\gamma_s / \sqrt{\lambda_s(\lambda_s + 1)}\right) \ , \tag{6.30}$$

where $\lambda_s = 1/\sigma_s^2$ and $\gamma_s = \lambda_s\mu_s$. To obtain the last equation we have made use of the following result

$$\int dx \exp\left(-\frac{1}{2}ax^2 + bx\right) \Phi(cx + d) = \frac{\Phi(\frac{bc+ad}{\sqrt{a^2+ac^2}})}{\sqrt{a}D(\frac{b}{\sqrt{a}})} \ , \quad a > 0 \ . \tag{6.31}$$

The cumulant generating function associated to the updated separator potential $\psi^*(\boldsymbol{x}_{t+1})$ therefore reduces to

$$\mathcal{K}(\boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{\xi}^T \boldsymbol{\Sigma}_{t+1|t}\boldsymbol{\xi} + \boldsymbol{\mu}_{t+1|t}^T\boldsymbol{\xi} + \log \Phi\left(\kappa_{t+1}\right) \ , \tag{6.32}$$

where we for notational convenience have introduced

$$\kappa_{t+1} = \frac{\dot{y}_{t+1}\boldsymbol{w}^T(\boldsymbol{\mu}_{t+1|t} + \boldsymbol{\Sigma}_{t+1|t}\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{w}}} \ . \tag{6.33}$$

The mean is then found by taking the derivative

$$\frac{\partial \mathcal{K}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}^T} = \boldsymbol{\xi}^T\boldsymbol{\Sigma}_{t+1|t} + \boldsymbol{\mu}_{t+1|t}^T + \frac{D(\kappa_{t+1})}{\Phi(\kappa_{t+1})}\frac{\partial \kappa_{t+1}}{\partial \boldsymbol{\xi}^T} \tag{6.34}$$

$$= \boldsymbol{\xi}^T\boldsymbol{\Sigma}_{t+1|t} + \boldsymbol{\mu}_{t+1|t}^T + \frac{D(\kappa_{t+1})}{\Phi(\kappa_{t+1})}\frac{\dot{y}_{t+1}\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}}{\sqrt{1+\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{w}}} \ , \tag{6.35}$$

and evaluate it in $\boldsymbol{\xi} = 0$ which then yields the mean of the newly updated separator potential

$$\boldsymbol{\mu}_{t+1|t+1} = \boldsymbol{\mu}_{t+1|t} + \frac{D(\kappa_{t+1}^{(0)})}{\Phi(\kappa_{t+1}^{(0)})}\frac{\dot{y}_{t+1}\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}}{\sqrt{1+\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{w}}} \ , \tag{6.36}$$

where $\kappa_{t+1}^{(0)}$ denotes κ_{t+1} evaluated in $\boldsymbol{\xi} = 0$, that is

$$\kappa_{t+1}^{(0)} = \frac{\dot{y}_{t+1}\boldsymbol{w}^T\boldsymbol{\mu}_{t+1|t}}{\sqrt{1+\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{w}}} \ . \tag{6.37}$$

Similarly, to derive the covariance matrix we start by calculating the Hessian

$$\frac{\partial^2 \mathcal{K}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}^T} = \boldsymbol{\Sigma}_{t+1|t} + \left(\frac{\partial}{\partial \boldsymbol{\xi}}\frac{D(\kappa_{t+1})}{\Phi(\kappa_{t+1})}\right)\frac{\dot{y}_{t+1}\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}}{\sqrt{1+\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{w}}} \tag{6.38}$$

where we have introduced

$$\frac{\partial}{\partial \boldsymbol{\xi}}\frac{D(\kappa_{t+1})}{\Phi(\kappa_{t+1})} = -H(\kappa_{t+1})\frac{\partial \kappa_{t+1}}{\partial \xi} \ , \tag{6.39}$$

and

$$H(\kappa_{t+1}) = \left(\kappa_{t+1} + \frac{D(\kappa_{t+1})}{\Phi(\kappa_{t+1})}\right)\frac{D(\kappa_{t+1})}{\Phi(\kappa_{t+1})} \ . \tag{6.40}$$

Evaluating the Hessian in $\boldsymbol{\xi} = 0$ yields the covariance matrix of the newly updated separator potential

$$\boldsymbol{\Sigma}_{t+1|t+1} = \boldsymbol{\Sigma}_{t+1|t} - H(\kappa_{t+1}^{(0)})\frac{\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}}{1+\boldsymbol{w}^T\boldsymbol{\Sigma}_{t+1|t}\boldsymbol{w}} \ . \tag{6.41}$$

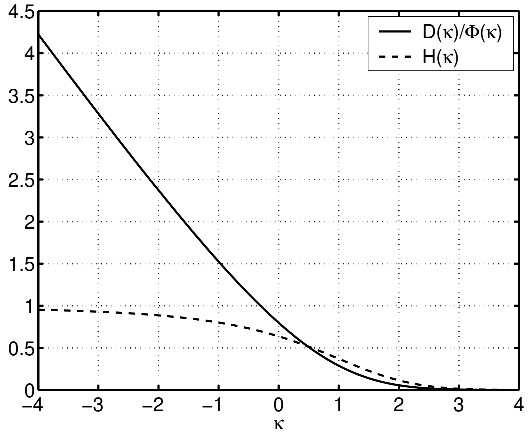


Fig. 6.2. Gain factors for the parameter updates in the on-line classification model. The gain factor for the mean and covariance matrix is $D(\kappa)/\Phi(\kappa)$ and $H(\kappa)$, respectively.

The basis case for the forward recursion is readily identified from the ordinary datum update recursion, since the cumulant generating function in the basis case basically is of the same type as eq. (6.21). Hence the basis case for the forward recursion is given by

$$\boldsymbol{\mu}_{0|0} = \sqrt{\frac{2}{\pi}} \frac{\dot{y}_0 \boldsymbol{w}^T \boldsymbol{\Sigma}_0}{\sqrt{1 + \boldsymbol{w}^T \boldsymbol{\Sigma}_0 \boldsymbol{w}}} \tag{6.42}$$

$$\boldsymbol{\Sigma}_{0|0} = \boldsymbol{\Sigma}_0 - \frac{2}{\pi} \frac{\boldsymbol{\Sigma}_0 \boldsymbol{w} \boldsymbol{w}^T \boldsymbol{\Sigma}_0}{1 + \boldsymbol{w}^T \boldsymbol{\Sigma}_0 \boldsymbol{w}} \ . \tag{6.43}$$

Figure 6.2 shows as a function of κ_{t+1} the gain factor for both the mean and covariance updates. Essentially this figure shows that we tend to make the most radical parameter changes when the observed labels are very unlikely under our one-step-ahead predictive distribution. Similarly, we tend not to make any changes to the parameters as long as we are predicting well; in other words we simply discount the value of unsurprising labels.

After having processed the entire dataset sequentially using these forward recursions we are left with filtered state estimates. Obviously, we could carry out a subsequent set of backward recursions which would provide us with smoothened state estimates. However, it should always be kept in mind that both the filtered and smoothened state estimate is just an approximation to the true posterior density.

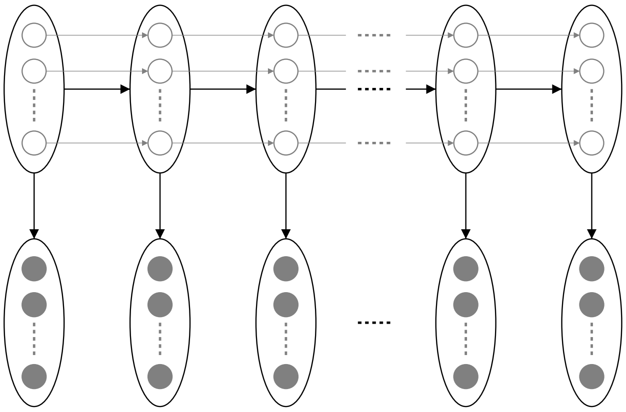


Fig. 6.3. Graphical representation of a generative model for ICA with temporal correlated sources. The connections from each of the sources in $\boldsymbol{X}_t$ to every microphone in $\boldsymbol{Y}_t$ are not shown explicitly but simply implied by a single connection between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$. The gray horizontal arrow shows the temporal dependence of each source as well as the a priori independence between sources.

6.3 ICA with temporal correlated sources

In this section we consider the implication of performing the moment passing scheme on a generative model for ICA with temporal correlated sources. The model, which is a simple extension of the ICA model considered in chapter 4, takes the form

$$p(\boldsymbol{x}_0) = \prod_m \frac{1}{2}\eta_0^{(m)} e^{-\eta_0^{(m)}|x_0^{(m)}|} \tag{6.44}$$

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \prod_m \frac{1}{2}\eta^{(m)} e^{-\eta^{(m)}|x_t^{(m)}-x_{t-1}^{(m)}|} \tag{6.45}$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{A}\boldsymbol{x}_t, \boldsymbol{\Sigma}) \ , \tag{6.46}$$

where $\eta_0^{(m)}$ and $\eta^{(m)}$ is the Laplacian decay rates associated to the m'th source. The graphical representation of this model is shown in figure 6.3. Whereas the intractable part of the inferential step for the on-line classification model was the measurement update, it is in this case the time update which is computationally intractable. We define the following set of potentials

$$\phi(\boldsymbol{x}_0, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{A}\boldsymbol{x}_0, \boldsymbol{\Sigma}) \prod_m \frac{1}{2}\eta_0^{(m)} e^{-\eta_0^{(m)}|x_0^{(m)}|} \tag{6.47}$$

$$\phi(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) = \prod_m \frac{1}{2}\eta^{(m)} e^{-\eta^{(m)}|x_t^{(m)}-x_{t-1}^{(m)}|} \tag{6.48}$$

$$\phi(\boldsymbol{x}_t, \boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{A}\boldsymbol{x}_t, \boldsymbol{\Sigma}) \ . \tag{6.49}$$

Again with reference in figure 6.1, assume that we have already obtained $\psi_g^*(\boldsymbol{x}_t)$ which is a Gaussian approximation to the separator potential $\psi^*(\boldsymbol{x}_t)$.

We now follow along the same lines as in section 6.2 and calculate the cumulant generating function of the updated potential $\psi^*(\boldsymbol{x}_{t+1})$. Since the likelihood is Gaussian we will, however, this time, restrict ourself to find a Gaussian approximation for the time update, i.e. we consider the cumulant generating function

$$\mathcal{K}(\boldsymbol{\xi}) = \log \int d\boldsymbol{x}_t \psi_g^*(\boldsymbol{x}_t) \int d\boldsymbol{x}_{t+1} \phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) e^{\boldsymbol{\xi}^T \boldsymbol{x}_{t+1}} \tag{6.50}$$

$$= \sum_m \frac{(\eta^{(m)})^2}{(\eta^{(m)})^2 - (\xi^{(m)})^2} + \log \int d\boldsymbol{x}_t \psi_g^*(\boldsymbol{x}_t) e^{\boldsymbol{\xi}^T \boldsymbol{x}_t} \ . \tag{6.51}$$

The first term in the last equation is simply a sum of cumulant generating functions for zero mean Laplacian distributions with variances $2/(\eta^{(m)})^2$ whereas the last term is the moment generating function of the Gaussian approximation $\psi_g^*(\boldsymbol{x}_t)$. This is, however, kind of a depressing result since it shows that the mean and the covariance matrix of the approximating Gaussian for the time update $\psi^*(\boldsymbol{x}_{t+1})$ is simply given by $\boldsymbol{\mu}_{t+1|t} = \boldsymbol{\mu}_{t|t}$ and $\boldsymbol{\Sigma}_{t+1|t} = \boldsymbol{\Sigma}_{t|t} + \boldsymbol{I}\boldsymbol{\eta}$, respectively. Using this approximation the remaining recursions are identical to the Kalman filter (and RTS) updates. This approach for temporal ICA clearly seems dubious. However, it should be remembered that the usual requirement of non-Gaussian of the source is not required when dealing with temporal correlated sources.

7. Conclusion

The Road goes ever on and on
Out from the door where it began.
Now far ahead the Road has gone,
Let others follow it who can!
Let them a journey new begin,
But I at last with weary feet
Will turn towards the lighted inn,
My evening-rest and sleep to meet.

J.R.R. Tolkien

This chapter summarizes the work presented in the thesis and outlines possible conclusions. Furthermore, in the end of this chapter there is given some suggestions for possible directions one could go to carry on this work.

7.1 Summary of the work

In this thesis we investigated mean field methods for carrying out approximate inference in intractable graphical models. In particular, we applied increasingly advanced mean field methods on a generative model for independent component analysis where the sources are instantaneously mixed. For a specific family of source priors we derived analytical expressions for the posterior mean and covariance of the sources which are needed for carrying out learning by expectation-maximization. In fact, it was shown that simply guessing a reasonable functional form for these quantities is sufficient to perform source separation. This is in a sense related to the concept of using contrast-functions in blind source separation. For algorithmic design, this is a useful property since it gives us some freedom to make ICA algorithms that only makes use of functions that are easy to evaluate. It was experimentally shown that overcomplete ICA is not possible within the naive mean field approach. However, by simply improving the naive mean field estimate of the posterior source covariance by a linear response correction step it turned out that separation of more sources than sensors was indeed possible.

We carried out a exploratory analysis of a fMRI dataset from a visual activation study using the computationally efficient ICA algorithm that one gets by assuming binary sources. With reference in the experimental setting in functional activation studies, we motivated the seemingly naive choice of using binary $\{0,1\}$-sources in this context. We would like to emphasize that we by no means are claiming that the presented generative model of the measured BOLD signal is solidly founded from a physiological point of view. However, this particular choice of sources naturally leads to a simple interpretation of the observed signal; something which is entirely missed by many of the "of the shelf" ICA approaches used in this context. In spite of its simplicity, this method indeed seems to infer reasonable brain activation patterns.

In the end of the thesis we considered a way for carrying out approximate message passing. There we used the moment matching property of the BKL distance instead of the KL divergence. We considered two examples; a generative model for on-line classification with binary labels and a generative model for ICA with temporal correlated sources. However, whether or not these models are useful remains to be investigated.

The results presented in this thesis shows that we for some type of probabilistic models indeed obtain better results by using simple tricks such a linear response correction or more advanced mean field methods like e.g. TAP. However, since the computational cost typically increases dramatically for the advanced mean field methods the improvement of the quality of the solution should be of a considerable size before invoking these methods in practical applications.

7.2 Suggestions for future work

The advanced mean field methods considered in this text are applicable to many of the models which have been proposed in the machine learning community and for which the naive mean field approach has proven to be efficient. It is still an open question whether the advanced mean field methods are feasible for any of those models. For instance, the naive mean field approach was used to make approximate inference in the Factorial Hidden Markov Model [Ghahramani and Jordan 1997] and it is likely that the advanced methods would be able to improve the quality of the inferential step; except for the distribution of the hidden states, this model is essentially identical to the generative model for temporal ICA considered in section 6.3.

At this time the persistent reader might have guessed that the subject of this text is deterministic methods (as opposed to sample based methods) for approximating inference in intractable probabilistic models. One obvious way for future work is of course that of combining the two approximating modalities into one efficient method. In my opinion, a must more interesting question is how the naive mean field approximation and sample based methods are related too each other as suggested in section 3.2.3. Essentially we need to know the answer to this question in order to do better than just combining the two modalities in a more or less ad hoc fashion.

Another really interesting direction of research is that of approximate inference by loopy belief propagation and its connection to ordinary mean field theory. In the recent years there has been quite a lot of progress in this field of research both experimentally and theoretically (e.g. see [Yedidia 2000]).

A. The multivariate Gaussian density

This appendix introduces the two parameterizations for the multivariate Gaussian density which are used in this thesis. Furthermore, for these two parameterizations we summarize the well known results for marginalizing and conditioning on variables in the Gaussian density since these operations happens to appear frequently in many highly celebrated statistical models, e.g. Gaussian linear state space models and Gaussian processes (also known as Kriging or optimal prediction/interpolation models).

The moment parameterization of the Gaussian density is given by

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = Z^{-1}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) , \tag{A.1}$$

where $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\Sigma}$ is the covariance matrix and $Z = |2\pi\boldsymbol{\Sigma}|^{1/2}$ is the normalizing constant. The canonical parameterization of the Gaussian density is given by

$$\mathcal{N}_*(\boldsymbol{x};\boldsymbol{\gamma},\boldsymbol{\Lambda}) = Z_*^{-1}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x}+\boldsymbol{\gamma}^T\boldsymbol{x}\right) , \tag{A.2}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\gamma} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are known as the canonical parameters and $Z_* = Z\exp\left(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$ is the normalizing constant expressed in terms of the parameters associated to the moment parameterization. In the canonical parameterization, the mean and covariance matrix is obviously given by $\boldsymbol{\mu} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$, respectively.

In order to state the standard results for marginalization and conditioning of Gaussian random variables $\boldsymbol{x}$, it is convenient to partition the vector $\boldsymbol{x}$ into two subvectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and consider the joint densities

$$p(\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right) \tag{A.3}$$

$$= \mathcal{N}_*\left(\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix};\begin{bmatrix}\gamma_1\\ \gamma_2\end{bmatrix},\begin{bmatrix}\boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12}\\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22}\end{bmatrix}\right) . \tag{A.4}$$

The following standard expressions essentially follows from block-diagonalizing the partitioned covariance matrix $\boldsymbol{\Sigma}$ and the partitioned inverse covariance matrix $\boldsymbol{\Lambda}$ as well as using knowledge about Gaussian integrals and the trick of completing squares. The block-diagonalizing step can be carried out using a Gauss-elimination procedure in which suitable matrices are premultiplied and postmultiplied onto the partitioned matrix.

In short, by conditioning on $\boldsymbol{x}_2$ it can be shown that the conditional density expressed in terms of the two parameterizations is given by

$$\begin{aligned} p(\boldsymbol{x}_1|\boldsymbol{x}_2) &= \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) && \text{(A.5)} \\ &= \mathcal{N}_*(\boldsymbol{x}_1; \boldsymbol{\gamma}_1 - \boldsymbol{\Lambda}_{12}\boldsymbol{x}_2, \boldsymbol{\Lambda}_{11}) \; . && \text{(A.6)} \end{aligned}$$

Similarly, it can be shown that marginalizing with respect to $\boldsymbol{x}_2$ yields the marginal density

$$\begin{aligned} p(\boldsymbol{x}_1) &= \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) && \text{(A.7)} \\ &= \mathcal{N}_*(\boldsymbol{x}_1; \boldsymbol{\gamma}_1 - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\gamma}_2, \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}) \; . && \text{(A.8)} \end{aligned}$$

Equations (A.5)–(A.8) show that there is no way of saying which of the two parameterizations are the better in terms of the computational cost. In fact, while one parameterization is advantageous for one type of Gaussian manipulations it turns out to be disadvantageous for the other type of Gaussian manipulations. In other words, for both parameterizations we must at some time face the computationally expensive problem of inverting a matrix.

When changing between the two parameterizations it is convenient to make use of the *matrix inversion lemma*[Anderson and Moore 1979] which establish the identity

$$\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \qquad \text{(A.9)}$$

Furthermore, using the matrix inversion lemma and the fact that

$$\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} = \boldsymbol{I} - \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)\boldsymbol{D}^{-1} \; , \qquad \text{(A.10)}$$

we can now derive another useful identity

$$\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \; . \qquad \text{(A.11)}$$

The identity provided by the matrix inversion lemma in a sense falls out during the derivation of the marginal and conditional Gaussian densities. To be specific, let $\boldsymbol{M}$ denote the partitioned matrix and let $\boldsymbol{U}$ and $\boldsymbol{L}$ denote an upper and lower triangular matrix, respectively. Clearly we can block-diagonalize $\boldsymbol{M}$ using both $\boldsymbol{UML}$ and $\boldsymbol{LMU}$ and use both results to derive expressions for the inverse $\boldsymbol{M}^{-1}$. Obviously, the two approaches should yield the same inverse $\boldsymbol{M}^{-1}$. The matrix inversion lemma is then obtained by simply equating entries in the two expressions for the inverse $\boldsymbol{M}^{-1}$. Since the matrix inversion lemma is derived by considering the inversion of block matrices it seems reasonable that this identity is indeed useful when juggling between the two parameterizations of the Gaussian density.

B. UAI*2001 submission

This appendix contains the paper:

de Freitas J. F. G., **Højen-Sørensen P. A. d. F. R.**, Jordan M. I. and Russell S.: Variational MCMC. Submitted to the *17th Conference on Uncertainty in Artificial Intelligence.* (2001).

C. Neurocomputing*2001 submission

This appendix contains the paper:

Højen-Sørensen P. A. d. F. R., Winther O. and Hansen L. K.: Analysis of Functional Neuroimages using ICA with Adaptive Binary Sources. Submitted to *Journal of Neurocomputing* (2001).

D. Neural Computation*2000 submission

This appendix contains the paper:

Højen-Sørensen P. A. d. F. R., Winther O. and Hansen L. K.: Mean Field Approaches to Independent Component Analysis. Submitted to *Journal of Neural Computation* (2000).

E. NIPS*2000 contribution

This appendix contains the paper:

Højen-Sørensen P. A. d. F. R., Winther O. and Hansen L. K.: Ensemble Learning and Linear Response Theory for ICA. In *Advances in Neural Information Processing Systems*, (NIPS 13), NIPS*00. (2000).

F. NNSP*2000 contribution

This appendix contains the paper:

Højen-Sørensen P. A. d. F. R., de Freitas J. F. G. and Fog T.: On-line Probabilistic Classification with Particle Filters. In *Proceeding of IEEE International Workshop on Neural Networks for Signal Processing.* NNSP*00, (2000).

G. NIPS*1999 contribution

This appendix contains the paper:

Højen-Sørensen P. A. d. F. R., Hansen L. K. and Rasmussen C. E.: Bayesian modelling of fMRI time series. In *Advances in Neural Information Processing Systems*, (NIPS 12), NIPS*99. (1999).

H. HBM*1999 contribution

This appendix contains the paper:

Højen-Sørensen P. A. d. F. R., Hansen L. K. and Rostrup E.: A Bayesian approach for estimating activation in fMRI time series. In *Proceedings of the 5th Int. Conf. on Functional Mapping of the Human Brain*. NeuroImage (1999).

Bibliography

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J.

Attias, H. (1999). Independent factor analysis. *Neural Computation*, 11(4):803–851.

Attias, H. (2000). A variational Bayesian framework for graphical models. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 209–215. MIT Press.

Bandettini, P. A., Jesmanowicz, A., Wong, E. C., and Hyde, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine*, 30:161–173.

Bandettini, P. A. and Wong, E. C. (1998). Echo–planar magnetic resonance imaging of human brain activation. In Schmitt, F., Stehling, M. K., Turner, R., and P., M., editors, *Echo–Planar Imaging*. Springer-Verlag, New York.

Barber, D. and van de Laar, P. (1999). Variational cumulant expansions for intractable distributions. *Journal of Artificial Intelligence Research*, 10:435–455.

Barinaga, M. (1997). What makes brain neurons run? *Science*, 276(11):196–198.

Bell, A. J. and Sejnowski, T. J. (1995). An information–maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Belouchrani, A. and Cardoso, J.-F. (1995). Maximum likelihood source separation by the expectation-maximization technique: deterministic and stochastic implementation. In *In Proc. NOLTA*, pages 49–53.

Bhattacharyya, C. and Keerthi, S. S. (1999). Mean-field methods for stochastic connectionist networks. Technical report, Dept. of Computer Science & Automation, Indian Institute of Science.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons, New York.

Cowell, R. G., Dawid, A. P., L., L. S., and J., S. D. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

Csató, L., Fokoué, E., Opper, M., Schottky, B., and Winther, O. (2000). Efficient approaches to Gaussian process classification. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 251–257. MIT Press.

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

Frey, B. J. (1998). *Graphical Models for Machine Learning and Digital Communication*. MIT Press, Cambridge, MA.

Geiger, D., Heckerman, D., and Meek, C. (1998). Asymptotic model selection for directed networks with hidden variables. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 461–477. Kluwer Academic Publishers, Dordrecht.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, 20(5):507–534.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Ghahramani, Z. and Beal, M. J. (2000). Variational inference for bayesian mixture of factor analysers. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 449–455. MIT Press.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, 29:245–275.

Girolami, M. (1998). An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*.

Girolami, M., editor (2000). *Advances in Independent Components Analysis*. Springer-Verlag, Berlin.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å., and Hansen, L. K. (99). On clustering fMRI time series. *NeuroImage*, 9:298–310.

Gradshteyn, I. S. and Ryzhik, I. M. (1980). *Tabel of Integrals, Series, and Products*. Academic Press, New York, corrected and enlarged edition edition.

Haft, M., Hofmann, R., and Tresp, V. (1999). Model-independent mean field theory as a local method for approximate propagation of information. *Network: Computation in Neural Systems*, 10:93–105.

Hansen, L. K. (2000). Blind separation of noisy image mixtures. In Girolami, M., editor, *Advances in Independent Components Analysis*, pages 165–187. Springer-Verlag, Berlin.

Hansen, L. K., Larsen, J., Nielsen, F. Å., Strother, S. C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J. J., and Svarer, C. (1999). Generalizable patterns in neuroimaging: How many principal components? *NeuroImage*, 9:534–544.

Hertz, J. A., Krogh, A. S., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

Højen-Sørensen, P. A. d. F. R., Winther, O., and Hansen, L. K. (2001a). In preparation.

Højen-Sørensen, P. A. d. F. R., Winther, O., and Hansen, L. K. (2001b). Ensemble learning and linear response theory for ICA. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Højen-Sørensen, P. A. d. F. R., Winther, O., and Hansen, L. K. (2001c). Mean field approaches to independent component analysis. *Neural Computation*. submitted.

Huang, C. and Darwiche, A. (1996). Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning*, 15:225–263.

Hyvärinen, A. and Karthikesh, R. (2000). Sparse priors on the mixing matrix in independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 477–452, Helsinki, Finland.

Jaakkola, T. S. and Jordan, M. I. (1998). Improving the mean field approximation via the use of mixture distributions. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 163–173. Kluwer Academic Publishers, Dordrecht.

Jaakkola, T. S. and Jordan, M. I. (1999). Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research*, 10:291–322.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks.* UCL Press, London.

Jordan, M. I., editor (1998). *Learning in Graphical Models.* Kluwer Academic Publishers, Dordrecht.

Jordan, M. I. and Bishop, C. M. (2001). *An Introduction to Graphical Models.*

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). An introduction to variational methods for graphical models. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 105–161. Kluwer Academic Publishers, Dordrecht.

Kappen, H. J. and Rodríguez, F. B. (1998a). Boltzmann machine learning using mean field theory and linear response correction. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems*, volume 10, pages 280–286. MIT Press.

Kappen, H. J. and Rodríguez, F. B. (1998b). Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10:1137–1156.

Kappen, H. J. and Wiegerinck, W. (2000). Mean field theory for graphical models. In Opper, M. and Saad, D., editors, *Advanced Mean Field Methods - Theory and Practice.* MIT Press. In press.

Kirkpatrick, S., Gelatt Jr., C. D., and Vecchi, M. P. (1983). Optimization by simmulated annealing. *Science*, 220:671–680.

Kjærulff, U. (1990). Triangulation of graphs – algorithms giving small total state space. Technical report, Dept. of Math. and Comp. Sci., Aalborg University, Denmark.

Knuth, K. (1999). A bayesian approach to source separation. In Cardoso, J.-F., Jutten, C., and Loubaton, P., editors, *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 283–288, Aussios, France.

Lange, N. and Zeger, S. L. (1997). Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging. *Journal of the Royal Statistical Society - Series C Applied Statistics*, 46(1–30).

Lappalainen, H. and Miskin, J. W. (2000). Ensemble learning. In Girolami, M., editor, *Advances in Independent Components Analysis*. 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.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Lee, T.-W. (1998). *Independent Component Analysis: Theory and Applications*. Kluwer Academic Publishers, Boston.

Leisink, M. A. R. and Kappen, H. J. (2000). A tighter bound for graphical models. *Neural Computation*. submitted.

Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, 12(2):337–365.

Li, S. Z. (1995). *Markov Random Field Modeling in Computer Vision*. Springer-Verlag, New York.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA., second edition.

MacKay, D. J. C. (1996). Maximum likelihood and covariant algorithms for independent component analysis. Technical report, University of Cambridge, Cavendish Laboratory. Draft 3.7.

MacKay, D. J. C. (1998). Introduction to monte carlo methods. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 175–204. Kluwer Academic Publishers, Dordrecht.

Marsden, J. E. and Hoffman, M. J. (1987). *Basic Complex Analysis*. W. H. Freeman and Company, New York, 2 edition.

McKeown, M., Jung, T.-P., Makeig, S., Brown, G., Kindermann, S. S., Lee, T.-W., and Sejnowski, T. J. (1998). Spatially independent activity patterns in functional MRI data during the Stroop color-naming task. In *Proc. Natl. Acad. Sci. USA*, volume 95, pages 803–810.

Mezard, M., Parisi, G., and Virasoro, M. (1987). *Spin Glass Theory and Beyond*, volume 9 of *Leture Notes in Physics*. World Scientific.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology.

Murphy, K. (1998). Filtering and smoothing in linear dynamical systems using the junction tree algorithm. Technical report, Computer Science Division, University of California, Berkeley, USA.

Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods. Technical report, Department of Computer Science, University of Toronto. CRG-TR-93-1.

Neal, R. M. and Hinton, G. E. (1998). 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. Kluwer Academic Publishers, Dordrecht.

Opper, M. (1998). A Bayesian approach to on-line learning. In Saad, D., editor, *On-line Learning in Neural Networks*, chapter 16, pages 363–378. Cambridge University Press.

Opper, M. and Winther, O. (2000a). From naive mean field theory to the TAP equations. In Opper, M. and Saad, D., editors, *Advanced Mean Field Methods - Theory and Practice*. MIT Press. In press.

Opper, M. and Winther, O. (2000b). Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12(11):2655–2684.

Opper, M. and Winther, O. (2000c). Tractable approximations for probabilistic models: The adaptive tap mean field approach. *Phys. Rev. Lett.* submitted.

Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley.

Petersen, K. S., Hansen, L. K., Kolenda, T., Rostrup, E., and Strother, S. (2000). On the independent components in functional neuroimages. In Pajunen, P. and Karhunen, J., editors, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 615–620, Helsinki, Finland.

Peterson, C. and Anderson, J. R. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019.

Peterson, C. and Söderberg, B. (1989). A new method for mapping optimization problems onto neural networks. *International Journal of Neural Systems*, 1:3–21.

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, A. P. (1999). Statistical limitations in functional neuroimaging. i. non-inferential methods and statistical models. *Philosophical Transactions of the Royal Society - Ser B - Biological Sciences*, 354(1387):1239–1260.

Pierro, A. R. (1993). On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Transactions on Medical Imaging*, 12(2):328–333.

Plefka, T. (1982). Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *J. Phys. A: Math. Gen.*, 15:1971–1978.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Rowe, D. (1999). Bayesian blind source separation. *IEEE Trans. Signal Processing*. submitted.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neaural Computation*, 11(2):305–345.

Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76.

Saul, L. K. and Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 486–492. MIT Press.

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

Shachter, R. D. (1998). Bayes-Ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference*, pages 480–487, San Francisco, CA. Morgan Kaufmann.

Sherrington, D. and Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Physical review letters*, 35(26):1792–1796.

Tanaka, T. (2000). Information geometry of mean-field approximation. *Neural Computation*, 12(8):1951–1968.

Thouless, D. J., Anderson, P. W., and Palmer, R. G. (1977). Solution of "solvable model of a spin glass". *Philosophical Magazine*, 35(3):593–601.

van der Veen, A.-J. (1997). Analytical method for blind binary signal separation. *IEEE Trans. on Signal Processing*, 45(4):1078–1082.

Worsley, K. J., Poline, J. B., Friston, K. J., and Evans, A. C. (1997). Characterizing the response of PET and fMRI data using multivariate linear models (MLM). *NeuroImage*, 6:305–319.

Yedidia, J. S. (2000). An idiosyncratic journey beyond mean field theory. In Opper, M. and Saad, D., editors, *Advanced Mean Field Methods - Theory and Practice*. MIT Press. In press.

Zhang, J. (1996). The application of the Gibbs-Bogoliubov-Feynman inequality in mean field calculations for Markov random fields. *IEEE Trans. on Image Processing*, 5(7):1208–1214.

Index