Chapter 7

Conclusion

7.1 Discussion

In this thesis we have shown how intractable Bayesian learning, inference, and model selection problems can be tackled using variational approximations. We have described a general framework for variational Bayesian learning and shown how it can be applied to several models of interest. We have demonstrated that it is an efficient and trustworthy approximation as compared to other more traditional approaches. Before summarising the contributions of this thesis, we spend the next few paragraphs discussing some of the evolving directions for model selection and variational Bayes, including the use of infinite models, inferring causal relationships using the marginal likelihood, other candidates for approximating the marginal likelihood, and lastly automated algorithm derivation procedures. These areas are expected to be active and fruitful future research directions. We conclude in section 7.2 with a summary of the main contributions of the thesis.

Infinite models

In this thesis we have focused on Bayesian learning in models that can be specified using a finite number of parameters. However, there are powerful arguments for entertaining models with infinitely many parameters, or at least as complex models as can be handled computationally. The process of Bayesian inference yields a unique answer. That is to say, given our prior beliefs, on observing some data all inference is automatic and there is one and only one answer to any prediction of the model. The problems of under- or overfitting by using too simple or too complex a model are simply not a concern if we integrate over all uncertain variables in the model, since applying Bayes' rule correctly at every step is guaranteed to result in coherent and optimal inferences given the prior beliefs. In this way the problem of model selection is no longer an issue, because the infinite model can entertain a continuum of models and average with respect to all of these simultaneously. This approach to modelling is discussed in Neal (1996) where, for example, neural networks with an infinite number of hidden units are shown (theoretically and empirically) to produce sensible predictions, and on some data sets state-of-the-art performance. In general it is difficult and sometimes impossible to entertain the limit of an infinite model, except where the mathematics lends itself to analytically tractable solutions — this is often the case for mixture models. Examples of Bayesian learning with infinite models include: the neural networks mentioned above, infinite mixtures of Gaussians (Rasmussen, 2000), infinite hidden Markov models (Beal et al., 2002), and infinite mixtures of Gaussian process experts (Rasmussen and Ghahramani, 2002). The basic idea of examining the infinite limit of finite models can be applied to a host of other as yet unexplored models and applications.

Unfortunately, a major drawback for these infinite models is that inference is generally intractable, and one has to resort to Monte Carlo sampling methods which can be computationally costly. Also, representing an infinite number of components in a mixture model, for example, can quickly become cumbersome; even elaborate Markov chain Monte Carlo approaches become very inefficient in models with many parameters. One further disadvantage of employing infinite models is that it is often difficult to find ways of encapsulating prior expert knowledge into the model. Methods such as examining the properties of data drawn from specific prior settings are illuminating but not always entirely satisfactory for designing the prior to articulate one's beliefs.

An alternative to grappling with the conceptual and implementational problems of infinite models is then to restrict ourselves to performing model inference, or selection amongst a finite set of finite-size models. Each individual model is then manageable and often simpler to interpret in terms of its structure. On the basis of the marginal likelihood we can obtain posterior distributions over the different candidate models. The problems discussed in this thesis have emphasised these model selection and structure learning tasks, as well as attempting to obtain full posterior distributions over model structures. We have examined a selection of statistical models, all of which contained hidden variables which cause the marginal likelihood computation to be intractable, and tackled this intractability using variational methods.

Bethe, Kikuchi, and cluster-variation methods

Variational Bayes, as described in this thesis, is just one type of variational approach that could be used to approximate Bayesian inference. It assumes simple forms for the posterior distributions over hidden variables and parameters, and then uses these forms to construct lower bounds on the marginal likelihood that are tractable. Algorithms for inference and learning are then derived as a result of optimising this lower bound by iteratively updating the parameters of these simplified distributions. Most of this thesis has concentrated on the ease with which the model parameters can be included in the set of uncertain variables to infer and integrate over, at least for the sub-class of conjugate-exponential models.

A promising alternative direction is to explore the Bethe and Kikuchi family of variational methods (Yedidia et al., 2001), sometimes called cluster-variational methods, which may be more accurate but do not provide the assurance of being bounds. These re-express the negative log marginal likelihood as a "free energy" from statistical physics, and then approximate the (intractable) entropy of the posterior distribution over latent variables by neglecting high order terms. In the Bethe approximation, the entropy is approximated with an expression which depends only on functions of single variables and pairs of variables. There are several procedures for minimising the Bethe free energy as a functional of the approximate posterior distributions to obtain estimates of the marginal likelihood. It turns out that for singly-connected graphs the fixed point equations that result from iterative minimisation of this free energy with respect to the single and pairwise functions correspond exactly to the messages that are passed in the junction tree and sum-product algorithms. Thus the Bethe free energy is exact for singly-connected graphs (trees). Interestingly, it has recently been shown that the belief propagation algorithm, even when run on multiply-connected graphs (i.e. 'loopy' graphs), has stable fixed points at the minima of the Bethe free energy (Heskes, 2003). While belief propagation on loopy graphs is not guaranteed to converge, it often works well in practice, and has become the standard approach to decoding state-of-the-art error-correcting codes. Furthermore, convergent algorithms for minimising the Bethe free energy have recently been derived (Yuille, 2001; Welling and Teh, 2001). There are other related methods, such as expectation propagation (EP, Minka, 2001a), approximations which observe higher order correlations in the variables (Leisink and Kappen, 2001), and other more elaborate variational schemes for upper bounds on partition functions (Wainwright et al., 2002).

The question remains open as to whether these methods can be readily applied to Bayesian learning problems. One can view Bayesian learning as simply treating the parameters as hidden variables, and so every method that has been shown to be successful for inference over hidden variables should also do well for integrating over parameters. However, there have been few satisfactory examples of Bayesian learning using any of the other methods described above, and this is an important direction for future research.

Inferring causal relationships

Most research in statistics has focused on inferring probabilistic dependencies between model variables, but more recently people have begun to investigate the more challenging and controversial problem of inferring *causality*. Causality can be understood statistically as a relationship $s \to t$ which is stable regardless of whether s was set through intervention / experimental manipulation or it occurred randomly. An example of this is smoking (s) causing yellowing of the teeth (t). Painting the teeth yellow does not change the probability of smoking, but forcing someone to smoke does change the probability of the teeth becoming yellow. Note that both the models $s \rightarrow t$ and $s \leftarrow t$ have the same conditional independence structure, yet they have very different causal interpretations. Unfortunately this has lead many researchers to believe that such causal relationships cannot be inferred from observational data alone, since these models are *likelihood equivalent* (Heckerman et al., 1995). Likelihood equivalent models are those for which an arc reversal can be accompanied by a change in parameters to yield the same likelihood. As a result these researchers then propose that causation can only be obtained by assessing the impact of active manipulation of one variable on another. However, this neglects the fact that the *prior* over parameters may cause the marginal likelihoods to be different even for likelihood equivalent models (D. MacKay, personal communication). In this context, it would be very interesting to explore the reliability with which variational Bayesian methods can be used to infer such causal relationships in general graphical models. In chapter 6 we showed that variational Bayes could determine the presence or absence of arcs from hidden variables to observed variables in a simple graphical model class. Envisaged then is a similar investigation for examining the directionality of arcs in a perhaps more expressive structure class.

Automated algorithm derivation

One of the problems with the variational Bayesian framework is that, despite the relative simplicity of the theory, the effort required to derive the update rules for the VBE and VBM steps is usually considerable and a hindrance to any implementation. Both the derivation and implementation have to be repeated for each new model, and both steps are prone to error. The variational linear dynamical system discussed in chapter 5 is a good example of a simple model for which the implementation is nevertheless cumbersome.

Our contribution of generalising the procedure for conjugate-exponential (CE) family models (section 2.4) is a step in the right direction for automated algorithm derivation. For CE models, we now know that the complexity of inference for variational Bayesian inference is the same as for point-parameter inference, and that for simple models such as HMMs existing propagation algorithms can be used unaltered with *variational Bayes point* parameters (see theorem 2.2).

There are a number of software implementations available or in development for inference and general automated algorithm derivation. The BUGS software package (Thomas et al., 1992) for automated Bayesian inference using Gibbs sampling is the most widely used at present; the graphical model and functional forms of the conditional probabilities involving both discrete and continuous variables can be specified by hand and then the sampling is left to obtain posterior distributions and marginal probabilities. For more generic algorithm derivation, the *AutoBayes* project (Gray et al., 2003) uses symbolic techniques to automatically derive the equations required for learning and inference in the model and explicitly produces the software to perform the task.

A similar piece of software is being developed in the *VIBES* project (Bishop et al., 2003). This package explicitly uses precisely the CE variational Bayesian results presented in chapter 2 of this thesis to automate the variational inference and learning processes, for (almost) arbitrary models expressed in graphical form. To be fully useful, this package should be able to cope with user-specified further approximation to the posterior, on top of just the parameter / hidden variable factorisation. Furthermore it should be relatively straightforward to allow the user to specify models which have non-CE components, such as logistic sigmoid functions. This would allow for discrete children of continuous parents, and could be made possible by including quadratic lower bounds on the sigmoid function (due to Jaakkola, 1997) to ensure that there is still a valid overall lower bound on the marginal likelihood. Looking further in to the future, these software applications may even be able to suggest 'good' factorisations, or work with a variety of these approximations together or even hierarchically. Also an alternative for coping with non-CE components of the model might be to employ sampling-based inferences in small regions of the graph that are affected.

Combining the variational Bayesian theory with a user-friendly interface in the form of VIBES or similar software could lead to the mass use of variational Bayesian methods in a wide variety of application fields. This would allow the ready comparison of a host of different models, and greatly improve the efficiency of current research on variational Bayes. However there is the caveat, which perhaps has not been emphasised enough in this thesis, that blind applications of variational Bayes may lead to the wrong conclusions, and that any inferences should be considered in the context of the approximations that have been made. This reasoning may not come easily to an automated piece of software, and the only sure answer to the query of whether the variational lower bound is reliable is to compare it to the exact marginal likelihood. It should not be difficult to overlay onto VIBES or similar software a set of sampling components to do exactly this task of estimating the marginal likelihood very accurately for diagnostic purposes; one such candidate for this task could be annealed importance sampling.

7.2 Summary of contributions

The aim of this thesis has been to investigate the variational Bayesian method for approximating Bayesian inference and learning in a variety of statistical models used in machine learning applications. Chapter 1 reviewed some of the basics of probabilistic inference in graphical models, such as the junction tree and belief propagation algorithms for exact inference in both undirected and directed graphs. These algorithms are used for inferring the distribution over hidden variables given observed data, for a *particular setting* of the model parameters. We showed that in situations where the parameters of the model are unknown the correct Bayesian procedure is to integrate over this uncertainty to form the marginal likelihood of the model. We explained how the marginal likelihood is the key quantity for choosing between models in a model selection task, but also explained that it is intractable to compute for almost all interesting models.

We reviewed a number of current methods for approximating the marginal likelihood, such as Laplace's method, the Bayesian information criterion (BIC), and the Cheeseman-Stutz criterion (CS). We discussed how each of these have significant drawbacks in their approximations. Perhaps the most salient deficiency is that they are based on maximum a posteriori parameter (MAP) estimates of the model parameters, which are arrived at by maximising the posterior density of the parameters, and so the MAP estimate may not be representative of the posterior mass at all. In addition we noted that the MAP optimisation is basis dependent, which means that two different experimenters with the same model and priors, but with different parameterisations, do not produce the same predictions using their MAP estimates. We also discussed a variety of sampling methods, and noted that these are guaranteed to give an exact answer for the marginal likelihood only in the limit of an infinite number of samples, and one often requires infeasibly long sampling runs to obtain accurate and reliable estimates.

In chapter 2 we presented the variational Bayesian method for approximating the marginal likelihood. We first showed how the standard expectation-maximisation (EM) algorithm for learning ML and MAP parameters can be interpreted as a variational optimisation of a lower bound on the likelihood of the data. In this optimisation, the E step can either be exact, in which case the lower bound is tight after each E step, or it can be restricted to a particular family of distributions in which case the bound is loose. The amount by which the bound is loose is exactly the Kullback-Leibler divergence between the variational hidden variable posterior and the exact posterior. We then generalised this methodology to the variational Bayesian EM algorithm which integrates over the parameters. The algorithm alternates between a VBE step which obtains a variational posterior distribution over the hidden variables given a distribution over the parameters, and a VBM step which infers the variational distribution over the parameters given the result of the VBE step. The lower bound gap is then given by the KL divergence between the variational joint posterior over hidden variables and parameters, and the corresponding exact posterior.

Significant progress in understanding the VB EM optimisation was made by considering the form of the update equations in the case of conjugate-exponential (CE) models. We showed that if the complete-data likelihood for the model is in the exponential family and the prior over parameters is conjugate to this likelihood, then the VB update equations take on analytically tractable forms and have attractive intuitive interpretations. We showed that, in theory, it is possible to use existing propagation algorithms for performing the VBE step, even though we have at all times a distribution over the parameters. This is made possible by passing the propagation algorithm the *variational Bayes point* parameter, $\boldsymbol{\theta}_{\mathrm{BP}} \equiv \boldsymbol{\phi}^{-1}(\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})})$, which is the result of inverting the exponential family's natural parameter mapping after averaging the natural parameters under the variational posterior. This is a very powerful result as it means that variational Bayesian inference (the VBE step) is possible in the same time complexity as the standard E step for the point-parameter case (with the only overhead being that of inverting the mapping). We also presented corollaries of this result applied to directed (Bayesian) and undirected (Markov) networks — see corollaries 2.2 and 2.4.

In chapter 3 we presented a straightforward example of this important result applied to Bayesian learning in a hidden Markov model. Here the variational Bayes point parameters are subnormalised transition and emission probabilities for the HMM, and the well-known forward-backward algorithm can be used unchanged with these modified parameters. We carried out experiments (some of which are suggested in MacKay, 1997) which showed that the VB algorithm was capable of determining the number of hidden states used to generate a synthetic data set, and outperforms ML and MAP learning on a task of discriminating between forwards and backwards English sentences. This shows that integrating over the uncertainty in parameters is important, especially for small data set sizes. The linear dynamical system of chapter 5 has the same structure as the HMM, so we might expect it to be equally suitable for the propagation corollary. However for this model it was not found to be possible to invert the natural parameter mapping, but nevertheless a variational Bayesian inference algorithm was derived with the same time complexity as the well-known Rauch-Tung-Striebel smoother. It was then shown that the VB LDS system could use automatic relevance determination methods to successfully determine the dimensionality of the hidden state space in a variety of synthetic data sets, and that the model was able to discard irrelevant driving inputs to the hidden state dynamics and output processes. Some preliminary results on elucidating gene-expression mechanisms were reported, and we expect this to be an active area of future research.

Chapter 4 focused on a difficult model selection problem, that of determining the numbers of mixture components in a mixture of factor analysers model. Search over model structures for MFAs is computationally intractable if each analyser is allowed to have different intrinsic dimensionalities. We derived and implemented the variational Bayesian EM algorithm for this MFA model, and showed that by wrapping the VB EM optimisation within a birth and death process we were able to navigate through the space of number of components using the lower bound as a surrogate for the marginal likelihood. Since all the parameters are integrated out in a Bayesian implementation, we are at liberty to begin the search either from the simplest model or from a model with very many components. Including an automatic relevance determination prior on the entries of each of the factor loading matrices' columns allowed the optimisation to simultaneously find the number of components and their dimensionalities. We demonstrated this on several synthetic data sets, and showed improved performance on a digit classification task as compared to a BIC-penalised ML MFA model. We noted that for this mixture model the death process was an automatic procedure, and also suggested several ways in which the birth processes could be implemented to increase the efficiency of the structure search.

Also in this chapter we presented a generally applicable importance sampling procedure for obtaining estimates of the marginal likelihood, predictive density, and the KL divergence between the variational and exact posterior distributions. In the sampler, the variational posteriors are used as proposal distributions for drawing importance samples. We found that although the lower bound tends to correlate well with the importance sampling estimate of the marginal likelihood, the KL divergence (the bound gap) increases approximately linearly with the number of components in the MFA model, which would suggest that the VB approximation has an inherent bias towards simpler models. We note also that importance sampling can fail for poor choices of proposal distribution and is not ideal for high dimensional parameter spaces. We attempted to improve the estimates by using heavier tailed and mixture distributions derived from the variational posteriors, but any improvements are not very conclusive. The problems with simple importance sampling have motivated attempts at combining variational methods with more sophisticated MCMC methods, but to date there have been few successful implementations, and this is an area of future work.

We showed in chapter 2 that the variational Bayesian EM algorithm is a generalisation of the EM algorithm for ML/MAP optimisation — the standard EM algorithm is recovered by restricting the form of the variational posterior distribution over parameters to a delta function, or a point-estimate. There is also the interesting observation that the VB approximation reduces to the BIC approximation in the limit of an infinitely large data set, for which we provided a brief proof in the case of CE models. However, we have also found intriguing connections between the VB lower bound and Cheeseman-Stutz approximations to the marginal likelihood. In particular we proved with theorem 2.3 that the CS criterion is a strict lower bound on the marginal likelihood for arbitrary models (not just those in the CE family), which was a previously unrecognised fact (although Minka (2001b) makes this observation in a mixture modelling context). We then built on this theorem to show with corollary 2.5 that there is a construction for obtaining a VB approximation which *always* results in *at least as tight a bound* as the CS criterion. This is a very interesting and useful result because it means that all existing implementations using CS approximations can now be made more faithful to the exact marginal likelihood by overlaying a variational Bayesian approximation. This is only a very recent discovery, and as a result has not yet been exploited to the full.

We saw superior performance of the variational Bayesian lower bound over the Cheeseman-Stutz and BIC criteria in chapter 6, where the task was finding the particular structure (out of a small class of structures) that gave rise to an observed data set, via the marginal likelihood. This was despite not making use of the aforementioned construction derived in corollary 2.5 (which we were not aware of when carrying out the chapter's experiments). In these experiments we found that VB outperformed both BIC and CS approximations, and also tended to provide more reliable results than the sampling gold standard, annealed importance sampling. Not only does the VB approximation provide a bound on the marginal likelihood (which in the experiments often showed AIS estimates to be 'invalid'), but it also arrives at this bound in a fraction (about

1%) of the time of the sampling approach. Moreover the VB approximation does not require the tuning of proposal distributions, annealing schedules, nor does it require extensive knowledge of the model domain to produce a reliable algorithm. We presented a number of extensions to the AIS algorithm, including a more general algorithm for computing marginal likelihoods which uses estimates based on more than one sample at each temperature (see algorithm 6.1). In the near future we hope to prove whether estimates using this algorithm are biased or not (personal communication with R. Neal).

To conclude, I hope that this thesis has provided an accessible and coherent account of the widely applicable variational Bayesian approximation. We have derived variational Bayesian algorithms for a variety of statistical models and provided the tools with which new models can be tackled, especially with a view to building software for automated algorithm derivation. This should throw open the doors to Bayesian learning in a host of models other than those investigated here. There are many directions for this research to be taken in and much work left to be done. The hope is that the experimental findings and insights documented in these chapters will stimulate and guide future research on variational Bayes.