

VARIATIONAL ALGORITHMS FOR APPROXIMATE BAYESIAN INFERENCE

by

Matthew J. Beal

M.A., M.Sci., Physics, University of Cambridge, UK (1998)



**The Gatsby Computational Neuroscience Unit
University College London
17 Queen Square
London WC1N 3AR**

**A Thesis submitted for the degree of
Doctor of Philosophy of the University of London**

May 2003

Abstract

The Bayesian framework for machine learning allows for the incorporation of prior knowledge in a coherent way, avoids overfitting problems, and provides a principled basis for selecting between alternative models. Unfortunately the computations required are usually intractable. This thesis presents a unified variational Bayesian (VB) framework which approximates these computations in models with latent variables using a lower bound on the marginal likelihood.

Chapter 1 presents background material on Bayesian inference, graphical models, and propagation algorithms. Chapter 2 forms the theoretical core of the thesis, generalising the expectation-maximisation (EM) algorithm for learning maximum likelihood parameters to the VB EM algorithm which integrates over model parameters. The algorithm is then specialised to the large family of conjugate-exponential (CE) graphical models, and several theorems are presented to pave the road for automated VB derivation procedures in both directed and undirected graphs (Bayesian and Markov networks, respectively).

Chapters 3-5 derive and apply the VB EM algorithm to three commonly-used and important models: mixtures of factor analysers, linear dynamical systems, and hidden Markov models. It is shown how model selection tasks such as determining the dimensionality, cardinality, or number of variables are possible using VB approximations. Also explored are methods for combining sampling procedures with variational approximations, to estimate the tightness of VB bounds and to obtain more effective sampling algorithms. Chapter 6 applies VB learning to a long-standing problem of scoring discrete-variable directed acyclic graphs, and compares the performance to annealed importance sampling amongst other methods. Throughout, the VB approximation is compared to other methods including sampling, Cheeseman-Stutz, and asymptotic approximations such as BIC. The thesis concludes with a discussion of evolving directions for model selection including infinite models and alternative approximations to the marginal likelihood.

Acknowledgements

I am very grateful to my advisor Zoubin Ghahramani for his guidance in this work, bringing energy and thoughtful insight into every one of our discussions. I would also like to thank other senior Gatsby Unit members including Hagai Attias, Phil Dawid, Peter Dayan, Geoff Hinton, Carl Rasmussen and Sam Roweis, for numerous discussions and inspirational comments.

My research has been punctuated by two internships at Microsoft Research in Cambridge and in Redmond. Whilst this thesis does not contain research carried out in these labs, I would like to thank colleagues there for interesting and often seductive discussion, including Christopher Bishop, Andrew Blake, David Heckerman, Nebojsa Jojic and Neil Lawrence.

Amongst many others I would like to thank especially the following people for their support and useful comments: Andrew Brown, Nando de Freitas, Oliver Downs, Alex Gray, Yoel Haitovsky, Sham Kakade, Alex Korenberg, David MacKay, James Miskin, Quaid Morris, Iain Murray, Radford Neal, Simon Osindero, Lawrence Saul, Matthias Seeger, Amos Storkey, Yee-Whye Teh, Eric Tuttle, Naonori Ueda, John Winn, Chris Williams, and Angela Yu.

I should thank my friends, in particular Paola Atkinson, Tania Lillywhite, Amanda Parmar, James Tinworth and Mark West for providing me with various combinations of shelter, companionship and retreat during my time in London. Last, but by no means least I would like to thank my family for their love and nurture in all my years, and especially my dear fiancée Cassandre Creswell for her love, encouragement and endless patience with me.

The work in this thesis was carried out at the Gatsby Computational Neuroscience Unit which is funded by the Gatsby Charitable Foundation. I am grateful to the Institute of Physics, the NIPS foundation, the UCL graduate school and Microsoft Research for generous travel grants.

Contents

Abstract	2
Acknowledgements	3
Contents	4
List of figures	8
List of tables	11
List of algorithms	12
1 Introduction	13
1.1 Probabilistic inference	16
1.1.1 Probabilistic graphical models: directed and undirected networks	17
1.1.2 Propagation algorithms	19
1.2 Bayesian model selection	24
1.2.1 Marginal likelihood and Occam’s razor	25
1.2.2 Choice of priors	27
1.3 Practical Bayesian approaches	32
1.3.1 Maximum a posteriori (MAP) parameter estimates	33
1.3.2 Laplace’s method	34
1.3.3 Identifiability: aliasing and degeneracy	35
1.3.4 BIC and MDL	36
1.3.5 Cheeseman & Stutz’s method	37
1.3.6 Monte Carlo methods	38
1.4 Summary of the remaining chapters	42
2 Variational Bayesian Theory	44
2.1 Introduction	44
2.2 Variational methods for ML / MAP learning	46
2.2.1 The scenario for parameter learning	46
2.2.2 EM for unconstrained (exact) optimisation	48

2.2.3	EM with constrained (approximate) optimisation	49
2.3	Variational methods for Bayesian learning	53
2.3.1	Deriving the learning rules	53
2.3.2	Discussion	58
2.4	Conjugate-Exponential models	64
2.4.1	Definition	64
2.4.2	Variational Bayesian EM for CE models	66
2.4.3	Implications	69
2.5	Directed and undirected graphs	73
2.5.1	Implications for directed networks	73
2.5.2	Implications for undirected networks	74
2.6	Comparisons of VB to other criteria	75
2.6.1	BIC is recovered from VB in the limit of large data	75
2.6.2	Comparison to Cheeseman-Stutz (CS) approximation	76
2.7	Summary	80
3	Variational Bayesian Hidden Markov Models	82
3.1	Introduction	82
3.2	Inference and learning for maximum likelihood HMMs	83
3.3	Bayesian HMMs	88
3.4	Variational Bayesian formulation	91
3.4.1	Derivation of the VBEM optimisation procedure	92
3.4.2	Predictive probability of the VB model	97
3.5	Experiments	98
3.5.1	Synthetic: discovering model structure	98
3.5.2	Forwards-backwards English discrimination	99
3.6	Discussion	104
4	Variational Bayesian Mixtures of Factor Analysers	106
4.1	Introduction	106
4.1.1	Dimensionality reduction using factor analysis	107
4.1.2	Mixture models for manifold learning	109
4.2	Bayesian Mixture of Factor Analysers	110
4.2.1	Parameter priors for MFA	111
4.2.2	Inferring dimensionality using ARD	114
4.2.3	Variational Bayesian derivation	115
4.2.4	Optimising the lower bound	119
4.2.5	Optimising the hyperparameters	122
4.3	Model exploration: birth and death	124
4.3.1	Heuristics for component death	126
4.3.2	Heuristics for component birth	127

4.3.3	Heuristics for the optimisation endgame	130
4.4	Handling the predictive density	130
4.5	Synthetic experiments	132
4.5.1	Determining the number of components	133
4.5.2	Embedded Gaussian clusters	133
4.5.3	Spiral dataset	135
4.6	Digit experiments	138
4.6.1	Fully-unsupervised learning	138
4.6.2	Classification performance of BIC and VB models	141
4.7	Combining VB approximations with Monte Carlo	144
4.7.1	Importance sampling with the variational approximation	144
4.7.2	Example: Tightness of the lower bound for MFAs	148
4.7.3	Extending simple importance sampling	151
4.8	Summary	157
5	Variational Bayesian Linear Dynamical Systems	159
5.1	Introduction	159
5.2	The Linear Dynamical System model	160
5.2.1	Variables and topology	160
5.2.2	Specification of parameter and hidden state priors	163
5.3	The variational treatment	168
5.3.1	VBM step: Parameter distributions	170
5.3.2	VBE step: The Variational Kalman Smoother	173
5.3.3	Filter (forward recursion)	174
5.3.4	Backward recursion: sequential and parallel	177
5.3.5	Computing the single and joint marginals	181
5.3.6	Hyperparameter learning	184
5.3.7	Calculation of \mathcal{F}	185
5.3.8	Modifications when learning from multiple sequences	186
5.3.9	Modifications for a fully hierarchical model	189
5.4	Synthetic Experiments	189
5.4.1	Hidden state space dimensionality determination (no inputs)	189
5.4.2	Hidden state space dimensionality determination (input-driven)	191
5.5	Elucidating gene expression mechanisms	195
5.5.1	Generalisation errors	198
5.5.2	Recovering gene-gene interactions	200
5.6	Possible extensions and future research	201
5.7	Summary	204
6	Learning the structure of discrete-variable graphical models with hidden variables	206

6.1	Introduction	206
6.2	Calculating marginal likelihoods of DAGs	207
6.3	Estimating the marginal likelihood	210
6.3.1	ML and MAP parameter estimation	210
6.3.2	BIC	212
6.3.3	Cheeseman-Stutz	213
6.3.4	The VB lower bound	215
6.3.5	Annealed Importance Sampling (AIS)	218
6.3.6	Upper bounds on the marginal likelihood	222
6.4	Experiments	223
6.4.1	Comparison of scores to AIS	226
6.4.2	Performance averaged over the parameter prior	232
6.5	Open questions and directions	236
6.5.1	AIS analysis, limitations, and extensions	236
6.5.2	Estimating dimensionalities of the incomplete and complete-data models	245
6.6	Summary	247
7	Conclusion	250
7.1	Discussion	250
7.2	Summary of contributions	254
	Appendix A Conjugate Exponential family examples	259
	Appendix B Useful results from matrix theory	262
B.1	Schur complements and inverting partitioned matrices	262
B.2	The matrix inversion lemma	263
	Appendix C Miscellaneous results	265
C.1	Computing the digamma function	265
C.2	Multivariate gamma hyperparameter optimisation	266
C.3	Marginal KL divergence of gamma-Gaussian variables	267
	Bibliography	270

List of figures

1.1	The elimination algorithm on a simple Markov network	20
1.2	Forming the junction tree for a simple Markov network	22
1.3	The marginal likelihood embodies Occam's razor	27
2.1	Variational interpretation of EM for ML learning	50
2.2	Variational interpretation of constrained EM for ML learning	51
2.3	Variational Bayesian EM	56
2.4	Hidden-variable / parameter factorisation steps	59
2.5	Hyperparameter learning for VB EM	62
3.1	Graphical model representation of a hidden Markov model	83
3.2	Evolution of the likelihood for ML hidden Markov models, and the subsequent VB lower bound.	100
3.3	Results of ML and VB HMM models trained on synthetic sequences.	101
3.4	Test data log predictive probabilities and discrimination rates for ML, MAP, and VB HMMs	103
4.1	ML Mixtures of Factor Analysers	110
4.2	Bayesian Mixtures of Factor Analysers	114
4.3	Determination of number of components in synthetic data	134
4.4	Factor loading matrices for dimensionality determination	135
4.5	The Spiral data set of Ueda et. al	136
4.6	Birth and death processes with VBMFA on the Spiral data set	137
4.7	Evolution of the lower bound \mathcal{F} for the Spiral data set	137
4.8	Training examples of digits from the CEDAR database	138
4.9	A typical model of the digits learnt by VBMFA	139
4.10	Confusion tables for the training and test digit classifications	140
4.11	Distribution of components to digits in BIC and VB models	143
4.12	Logarithm of the marginal likelihood estimate and the VB lower bound during learning of the digits $\{0, 1, 2\}$	150
4.13	Discrepancies between marginal likelihood and lower bounds during VBMFA model search	152

4.14	Importance sampling estimates of marginal likelihoods for learnt models of data of differently spaced clusters	156
5.1	Graphical model representation of a state-space model	161
5.2	Graphical model for a state-space model with inputs	162
5.3	Graphical model representation of a Bayesian state-space model	164
5.4	Recovered LDS models for increasing data size	190
5.5	Hyperparameter trajectories showing extinction of state-space dimensions	191
5.6	Data for the input-driven LDS synthetic experiment	193
5.7	Evolution of the lower bound and its gradient	194
5.8	Evolution of precision hyperparameters, recovering true model structure	196
5.9	Gene expression data for input-driven experiments on real data	197
5.10	Graphical model of an LDS with feedback of observations into inputs	199
5.11	Reconstruction errors of LDS models trained using MAP and VB algorithms as a function of state-space dimensionality	200
5.12	Gene-gene interaction matrices learnt by MAP and VB algorithms, showing significant entries	202
5.13	Illustration of the gene-gene interactions learnt by the feedback model on expression data	203
6.1	The chosen structure for generating data for the experiments	225
6.2	Illustration of the trends in marginal likelihood estimates as reported by MAP, BIC, BICp, CS, VB and AIS methods, as a function of data set size and number of parameters	228
6.3	Graph of rankings given to the true structure by BIC, BICp, CS, VB and AIS methods	230
6.4	Differences in marginal likelihood estimate of the top-ranked and true structures, by BIC, BICp, CS, VB and AIS	232
6.5	The median ranking given to the true structure over repeated settings of its parameters drawn from the prior, by BIC, BICp, CS and VB methods	233
6.6	Median score difference between the true and top-ranked structures, under BIC, BICp, CS and VB methods.	234
6.7	The best ranking given to the true structure by BIC, BICp, CS and VB methods	235
6.8	The smallest score difference between true and top-ranked structures, by BIC, BICp, CS and VB methods	236
6.9	Overall success rates of BIC, BICp, CS and VB scores, in terms of ranking the true structure top	237
6.10	Example of the variance of the AIS sampler estimates with annealing schedule granularity, using various random initialisations, shown against the BIC and VB estimates for comparison	238

6.11 Acceptance rates of the Metropolis-Hastings proposals as a function of size of data set	240
6.12 Acceptance rates of the Metropolis-Hastings sampler in each of four quarters of the annealing schedule	242
6.13 Non-linear AIS annealing schedules	244

List of tables

2.1	Comparison of EM for ML/MAP estimation against VB EM with CE models	70
4.1	Simultaneous determination of number of components and their dimensionalities	135
4.2	Test classification performance of BIC and VB models	142
4.3	Specifications of six importance sampling distributions	155
6.1	Rankings of the true structure amongst the alternative candidates, by MAP, BIC, BICp, VB and AIS estimates, both corrected and uncorrected for posterior aliasing	230
6.2	Comparison of performance of VB to BIC, BICp and CS methods, as measured by the ranking given to the true model	233
6.3	Improving the AIS estimate by pooling the results of several separate sampling runs	241
6.4	Rate of AIS violations of the VB lower bound, alongside Metropolis-Hastings rejection rates	241
6.5	Number of times the true structure is given the highest ranking by the BIC, BICp, CS, CS [†] , and VB scores	247

List of Algorithms

5.1	Forward recursion for variational Bayesian state-space models	178
5.2	Backward parallel recursion for variational Bayesian state-space models	181
5.3	Pseudocode for variational Bayesian state-space models	187
6.1	AIS algorithm for computing all ratios to estimate the marginal likelihood . . .	221
6.2	Algorithm to estimate the complete- and incomplete-data dimensionalities of a model	246