

# Variational Message Passing and its Applications

John M. Winn  
St John's College  
Cambridge

A dissertation submitted in candidature for the degree of Doctor of Philosophy,  
University of Cambridge

Inference Group  
Cavendish Laboratory  
University of Cambridge



Submitted October 2003, revised January 2004

# DECLARATION

I hereby declare that my dissertation entitled “Variational Message Passing and its Applications” is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university.

I further state that no part of my dissertation has already been or is being concurrently submitted for any such degree or diploma or other qualification.

Except where explicit reference is made to the work of others, this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration. This dissertation does not exceed sixty thousand words in length.

Date: .....

Signed: .....

John M. Winn  
St John's College  
Cambridge  
January 13th, 2004

# ABSTRACT

This thesis is concerned with the development of Variational Message Passing (VMP), an algorithm for automatically performing variational inference in a probabilistic graphical model. VMP allows learning and reasoning about a system to proceed directly from a given probabilistic model of that system. The utility of VMP has been demonstrated by solving problems in the domains of machine vision and bioinformatics. VMP dramatically simplifies the construction and testing of new variational models and readily allows a range of alternative models to be tested on a given problem.

In chapter 1, a probabilistic approach to automatic learning and reasoning is introduced. Belief propagation, an existing exact inference algorithm that uses message passing in a graphical model, is outlined, along with its limitations. These limitations lead to the need for approximate inference methods, including sampling methods and variational inference. The latter method of variational inference, which provides an analytical approximation to the posterior distribution, is described in detail.

Chapter 2 presents a novel framework for performing automatic variational inference in a wide range of probabilistic models. The core of the framework is the Variational Message Passing algorithm which is an analog of belief propagation that uses message passing within a graphical model to optimise an approximate variational distribution. A software package, called VIBES (Variational Inference in BayESian networks), is presented as an implementation of the VMP framework. A tutorial is included which demonstrates applying VIBES to a small data set.

Chapter 3 sees the framework being applied to the problem of modelling non-linear image manifolds such as those of face images and digits images. In chapter 4, the problems of DNA microarray image analysis and gene expression modelling are addressed, again using the VMP framework.

Chapter 5 extends Variational Message Passing by allowing variational distributions which retain part of the dependency structure of the original model. The resulting Structured VMP algorithm is shown to improve the quality of the approximate inference and hence widen the applicability of the framework. Conclusions and suggestions for future research directions are presented in Chapter 6.

Dedicated to the memory of Carol Melissa Smith and of my mother Barbara Winn.  
They both possessed a joy in life and learning that will always be an inspiration.

# ACKNOWLEDGEMENTS

I am very grateful to David MacKay and Chris Bishop for their guidance, ideas and inspiration. I would also like to thank Matthew Beal, Andrew Blake, Phil Cowans, Bill Fitzgerald, Zoubin Ghahramani, David Kreil, Neil Lawrence, Matthew Orton, Ed Ratzner, David Spiegelhalter, Mike Tipping, Hanna Wallach and Sebastian Wills for their advice, insight and many useful discussions.

Finally, I would like to thank Sara for her love and support and for giving me the strength to complete this doctorate.

This work was supported by Microsoft Research Cambridge, the Engineering Department and the Cavendish Laboratory.

# NOTATION

## Variables and sets of variables

$X, Y, Z \dots$	Variables or corresponding nodes in a graph
$\mathbf{X}, \mathbf{Y}, \mathbf{Z} \dots$	Sets of variables or corresponding sets of nodes in a graph
$X_i$	The $i$ th variable in set $\mathbf{X}$ or the corresponding node
$X_i = x$	Variable $X_i$ is in state $x$
$\{x_i\}_{i=1}^N$	The set $\{x_1, \dots, x_N\}$
$\mathbf{X} \setminus \mathbf{Y}$	The variables in set $\mathbf{X}$ that are not in set $\mathbf{Y}$
$c, d \dots$	Clusters (small subsets of variables)

## Probability theory

$P(X Y)$	The probability distribution over $X$ given $Y$ (also used to describe a conditional probability density)
$P(X = x   Y = y)$ or $P(x y)$	The probability that variable $X$ is in state $x$ , given that $Y$ is in state $y$
$X \sim P(X)$	$X$ is distributed according to $P(X)$
$Q(X Y)$	The probability distribution which is a variational approximation to $P(X Y)$
$\text{KL}(Q  P)$	The Kullback-Leibler divergence between the distributions $Q$ and $P$
$\langle f \rangle_Q$	The expectation of $f$ under the probability distribution $Q$
$\mathbb{H}(P)$	The entropy of the distribution $P$
$\delta(X x_0)$	The continuous probability distribution with the property $\int f(X) \delta(X x_0) dX = f(x_0)$ or the discrete distribution that assigns probability 1 to $X = x_0$ and 0 to $X \neq x_0$
$\Phi(c), \Psi(d)$	Cluster potentials (non-negative functions of cluster variables)

## Graphical models

$\text{pa}_i$	The variables or nodes corresponding to the parents of $X_i$ in a directed graph
$\text{ch}_i$	The variables or nodes corresponding to the children of $X_i$ in a directed graph
$\text{cp}_i^{(j)}$	The set of parents of $X_i$ excluding the parent $X_j$ (the co-parents)
$\text{ne}_i$	The variables or nodes corresponding to the neighbours of $X_i$ in a graph

# CONTENTS

<b>Chapter 1</b>	<b>Inference in Graphical Models</b>	<b>1</b>
1.1	Learning Machines . . . . .	2
1.2	Representing Uncertainty . . . . .	2
1.3	Probabilistic Models . . . . .	3
1.4	Graphical Models . . . . .	5
1.4.1	Bayesian networks . . . . .	5
1.4.2	Factor graphs . . . . .	6
1.4.3	Form of local functions . . . . .	7
1.5	Learning a Probabilistic Model . . . . .	8
1.5.1	Model fitting . . . . .	8
1.5.2	Model selection . . . . .	8
1.6	Bayesian Inference . . . . .	9
1.6.1	Variable elimination . . . . .	10
1.6.2	Belief propagation . . . . .	10
1.6.3	Inference in graphs with cycles . . . . .	14
1.6.4	Tractability of exact probabilistic inference . . . . .	16
1.7	Sampling Methods . . . . .	16
1.8	Variational Inference . . . . .	17
1.8.1	Kullback–Leibler divergence . . . . .	18
1.8.2	Minimising the divergence . . . . .	19
1.8.3	Variational model selection . . . . .	20
1.8.4	Factorised Q distribution . . . . .	21
1.8.5	Example: a univariate Gaussian model . . . . .	23
1.8.6	Comparison to Maximum A Posteriori . . . . .	25
1.8.7	Example: a Gaussian mixture model . . . . .	27
1.9	Overview . . . . .	31
<b>Chapter 2</b>	<b>A Variational Inference Framework</b>	<b>32</b>
2.1	Variational Message Passing . . . . .	33
2.1.1	A factorised Q distribution leads to local computations . . . . .	33
2.1.2	Message passing in conjugate-exponential models . . . . .	34

---

2.1.3	Example: the univariate Gaussian model . . . . .	37
2.1.4	Calculation of the lower bound $\mathcal{L}(\mathbf{Q})$ . . . . .	39
2.2	Allowable Models . . . . .	41
2.2.1	Conjugacy constraints . . . . .	41
2.2.2	Deterministic functions . . . . .	42
2.2.3	Mixture models . . . . .	45
2.2.4	Multivariate distributions . . . . .	47
2.2.5	Summary of allowable models . . . . .	48
2.3	VIBES: A Software Implementation . . . . .	48
2.4	Tutorial: the Gaussian Mixture Model . . . . .	50
2.5	Discussion . . . . .	55
2.5.1	Initialisation of the variational distribution . . . . .	56
2.5.2	Interpretation as a factor graph algorithm . . . . .	56
2.6	Extensions to the Framework . . . . .	58
2.6.1	Finding a Maximum A Posteriori solution . . . . .	58
2.6.2	Non-conjugate priors . . . . .	59
2.7	Summary . . . . .	60
<b>Chapter 3</b>	<b>Application: Non-linear Image Modelling</b>	<b>61</b>
3.1	Modelling Image Subspaces . . . . .	61
3.2	Models for Manifolds . . . . .	62
3.2.1	Maximum likelihood PCA . . . . .	63
3.2.2	Bayesian PCA . . . . .	64
3.2.3	Mixtures of Bayesian PCA models . . . . .	67
3.3	Variational Inference . . . . .	68
3.3.1	Derivation of the variational solution . . . . .	68
3.4	Results . . . . .	70
3.4.1	Synthetic data: noisy sinusoid . . . . .	70
3.4.2	Synthetic data: noisy sphere . . . . .	70
3.4.3	Faces data set . . . . .	72
3.4.4	Handwritten digits data set . . . . .	73
3.4.5	Image compression . . . . .	74
3.5	Discussion . . . . .	76
<b>Chapter 4</b>	<b>Application: Microarray Image Analysis</b>	<b>77</b>
4.1	DNA Microarrays . . . . .	77
4.2	Microarray Images . . . . .	78
4.2.1	Experimental methodology . . . . .	79
4.3	A Probabilistic Model for Microarray Images . . . . .	80
4.3.1	Latent variables and their prior distributions . . . . .	80
4.3.2	The likelihood function . . . . .	81

4.4	Variational Message Passing with Importance Sampling . . . . .	84
4.5	Inference in the Microarray Image Model . . . . .	86
4.5.1	Handling missing and obscured spots . . . . .	86
4.5.2	Updating the prior parameters of the model . . . . .	87
4.5.3	Determining the spot intensities . . . . .	88
4.5.4	Spot-finding results . . . . .	88
4.6	Automatic Sub-grid Location . . . . .	88
4.6.1	The sub-grid transform and its prior . . . . .	89
4.6.2	Inferring the sub-grid transform . . . . .	90
4.6.3	Searching through transform space . . . . .	91
4.6.4	Finding the MAP solution . . . . .	93
4.6.5	Results for sub-grid finding . . . . .	94
4.6.6	Overall sub-grid location . . . . .	95
4.7	Discussion . . . . .	95
4.8	Gene Expression Data Analysis . . . . .	96
4.8.1	ICA of gene expression data using VMP . . . . .	97
4.8.2	Conclusion . . . . .	100
<b>Chapter 5</b>	<b>Structured Variational Distributions</b>	<b>101</b>
5.1	Inference Using Structured Variational Distributions . . . . .	101
5.1.1	Optimising structured variational distributions . . . . .	102
5.1.2	Using a Bayesian network as a variational distribution . . . . .	104
5.2	Choice of Structured Variational Distribution . . . . .	104
5.3	Variational Junction Trees . . . . .	106
5.3.1	Inference using variational junction trees . . . . .	107
5.4	Reducing Computation for Inference with VJTs . . . . .	108
5.4.1	Case I: $Q$ junction tree with no internal deleted edges . . . . .	109
5.4.2	Case II: $Q$ junction tree with internal deleted edges . . . . .	110
5.5	An Algorithm for Structured Variational Inference . . . . .	112
5.5.1	Allowable models . . . . .	112
5.5.2	Structured Variational Message Passing algorithm . . . . .	113
5.5.3	Extending the algorithm to allow internal deleted edges (Case II) . . . . .	116
5.6	Structured VIBES: A Partial Implementation of SVMP . . . . .	118
5.6.1	Example: Hidden Markov Model . . . . .	118
5.7	Discussion . . . . .	120
<b>Chapter 6</b>	<b>Conclusions and Future Work</b>	<b>121</b>
6.1	Conclusions . . . . .	121
6.2	Suggestions for Future Work . . . . .	123
6.3	Summary . . . . .	124

---

<b>Appendix A</b>	<b>Exponential Family Distributions</b>	<b>125</b>
A.1	Gaussian distribution . . . . .	125
A.2	Rectified Gaussian distribution . . . . .	126
A.3	Gamma distribution . . . . .	126
A.4	Discrete distribution . . . . .	127
A.5	Dirichlet distribution . . . . .	127

# LIST OF FIGURES

1.1	A Bayesian network for the lie detector probabilistic model . . . . .	5
1.2	Factor graphs for the lie detector model . . . . .	6
1.3	Messages passed during <code>CollectEvidence</code> . . . . .	12
1.4	Messages passed during <code>DistributeEvidence</code> . . . . .	13
1.5	Conversion of a factor graph into a tree by clustering two variables . . .	15
1.6	Illustration of the asymmetry of the KL divergence . . . . .	18
1.7	Relationship between the lower bound, the KL divergence and the marginal log likelihood. . . . .	21
1.8	The Bayesian network for a univariate Gaussian model . . . . .	23
1.9	Variational and true posterior over the parameters of a Gaussian model .	25
1.10	The Bayesian network for a Gaussian mixture model . . . . .	28
1.11	True mixture of Gaussians distribution, along with samples from the variational posterior . . . . .	30
1.12	Mixture of Gaussians model applied to a two-dimensional data set . . . .	31
2.1	The variational update for a factor of the $Q$ distribution depends only the Markov blanket of the corresponding node . . . . .	34
2.2	Variational message passing in a univariate Gaussian model. . . . .	38
2.3	Screenshot of VIBES showing the Bayesian network for a univariate Gaussian model . . . . .	49
2.4	A VIBES model with a single observed node $x$ which has attached data. . . . .	51
2.5	A two-dimensional Gaussian model in VIBES . . . . .	51
2.6	Changing the distribution of $x$ to a mixture of Gaussians. . . . .	52
2.7	The completed Gaussian mixture model showing the discrete indicator node $\lambda$ . . . . .	53
2.8	A Hinton diagram showing the expected value of $\pi$ for each mixture component . . . . .	53
2.9	A Hinton diagram whose columns give the expected two-dimensional value of the mean $\mu$ for each mixture component. . . . .	53
2.10	A graph of the evolution of the lower bound during inference. . . . .	54
2.11	Two modifications to the Mixture of Gaussians model . . . . .	54

2.12	Further modified mixture model where the $\pi$ and $\gamma$ nodes are now common to all data dimensions. . . . .	55
3.1	The Bayesian network for the Principal Component Analysis model . . .	65
3.2	Hinton diagrams showing maximum likelihood vs. Bayesian PCA on a toy data set . . . . .	66
3.3	Effective dimensionality of a Bayesian PCA model for various sizes of data set. . . . .	67
3.4	The Bayesian network for the mixture of PCA models . . . . .	68
3.5	Bayesian PCA mixture model fitted to a highly non-linear one dimensional manifold . . . . .	71
3.6	Bayesian PCA mixture models fitted to a noisy two dimensional manifold: the surface of a sphere . . . . .	71
3.7	Hinton diagram of alpha hyper-parameters showing the manifold dimensionality . . . . .	72
3.8	Synthetic faces obtained by running the learned mixture distribution generatively. . . . .	72
3.9	ROC curves for classifying images as faces versus non-faces . . . . .	73
3.10	Digits synthesised from each of the ten trained Bayesian PCA mixture model by running the models generatively. . . . .	74
3.11	The original image and detail of the reconstructed images for various compression methods . . . . .	75
4.1	Two sub-grids extracted from different microarray images . . . . .	79
4.2	The Bayesian network for a probabilistic model of microarray sub-grid images. . . . .	83
4.3	Results of the microarray image analysis algorithm on two test images .	89
4.4	Diagram showing how area sums can be found for rectangular image regions and an example of approximating spot images with rectangles. . .	91
4.5	Results of sub-grid finding at both whole slide and the individual levels .	94
4.6	The Bayesian network for the Independent Component Analysis model .	97
4.7	Hinton diagram of the mean amplitude matrix for the ovarian tissue sample data set . . . . .	98
4.8	Activity and gene expression levels of the 4th signature . . . . .	99
4.9	Activity and gene expression levels of the 8th signature . . . . .	100
4.10	Activity and gene expression levels of the 15th signature . . . . .	100
5.1	Example showing how a structured variational distribution gives a better approximation to the exact posterior than a fully factorised distribution.	105
5.2	A Bayesian network, a cluster tree and a junction tree. . . . .	106

---

5.3	The Bayesian network for a model, along with the junction trees of $Q$ and graph used when applying Structured VMP to the model . . . . .	113
5.4	Example showing the need for additional moralisation in order to find optimal variational marginals. . . . .	116
5.5	A Case II Bayesian network, $Q$ junction trees and graph used when applying Structured Variational Message Passing. . . . .	116
5.6	VIBES screenshot showing the Bayesian network for a Hidden Markov model. . . . .	119
5.7	VIBES screenshot of the HMM structured variational distribution, showing the junction tree. . . . .	119