Signals and Communication Technology

V. Šmídl
A. Quinn

The Variational Bayes Method in Signal Processing

Springer

Springer Series on
Signals and Communication Technology

Signals and Communication Technology

Circuits and Systems
Based on Delta Modulation
Linear, Nonlinear and Mixed Mode Processing
D.G. Zrilic ISBN 3-540-23751-8

Functional Structures in Networks
AMLn – A Language for Model Driven
Development of Telecom Systems
T. Muth ISBN 3-540-22545-5

RadioWave Propagation
for Telecommunication Applications
H. Sizun ISBN 3-540-40758-8

Electronic Noise and Interfering Signals
Principles and Applications
G. Vasilescu ISBN 3-540-40741-3

DVB
The Family of International Standards
for Digital Video Broadcasting, 2nd ed.
U. Reimers ISBN 3-540-43545-X

Digital Interactive TV and Metadata
Future Broadcast Multimedia
A. Lugmayr, S. Niiranen, and S. Kalli
ISBN 3-387-20843-7

Adaptive Antenna Arrays
Trends and Applications
S. Chandran (Ed.) ISBN 3-540-20199-8

Digital Signal Processing
with Field Programmable Gate Arrays
U. Meyer-Baese ISBN 3-540-21119-5

Neuro-Fuzzy and Fuzzy Neural Applications
in Telecommunications
P. Stavroulakis (Ed.) ISBN 3-540-40759-6

SDMA for Multipath Wireless Channels
Limiting Characteristics
and Stochastic Models
I.P. Kovalyov ISBN 3-540-40225-X

Digital Television
A Practical Guide for Engineers
W. Fischer ISBN 3-540-01155-2

Multimedia Communication Technology
Representation, Transmission
and Identification of Multimedia Signals
J.R. Ohm ISBN 3-540-01249-4

Information Measures
Information and its Description in Science
and Engineering
C. Arndt ISBN 3-540-40855-X

Processing of SAR Data
Fundamentals, Signal Processing,
Interferometry
A. Hein ISBN 3-540-05043-4

Chaos-Based Digital
Communication Systems
Operating Principles, Analysis Methods,
and Performance Evalutation
F.C.M. Lau and C.K. Tse ISBN 3-540-00602-8

Adaptive Signal Processing
Application to Real-World Problems
J. Benesty and Y. Huang (Eds.)
ISBN 3-540-00051-8

Multimedia Information Retrieval
and Management
Technological Fundamentals and Applications
D. Feng, W.C. Siu, and H.J. Zhang (Eds.)
ISBN 3-540-00244-8

Structured Cable Systems
A.B. Semenov, S.K. Strizhakov,
and I.R. Suncheley ISBN 3-540-43000-8

UMTS
The Physical Layer of the Universal Mobile
Telecommunications System
A. Springer and R. Weigel
ISBN 3-540-42162-9

Advanced Theory of Signal Detection
Weak Signal Detection in
Generalized Obeservations
I. Song, J. Bae, and S.Y. Kim
ISBN 3-540-43064-4

Wireless Internet Access over GSM and UMTS
M. Taferner and E. Bonek
ISBN 3-540-42551-9

The Variational Bayes Method
in Signal Processing
V. Šmídl and A. Quinn
ISBN 3-540-28819-8

Václav Šmídl
Anthony Quinn

The Variational Bayes Method in Signal Processing

With 65 Figures

Springer

Dr. Václav Šmídl
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic, Department of Adaptive Systems
PO Box 18, 18208 Praha 8, Czech Republic
E-mail: smidl@utia.cas.cz

Dr. Anthony Quinn
Department of Electronic and Electrical Engineering
University of Dublin, Trinity College
Dublin 2, Ireland
E-mail: aquinn@tcd.ie

ISBN-10 3-540-28819-8 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-28819-0 Springer Berlin Heidelberg New York

Library of Congress Control Number: 2005934475

Typesetting and production: SPI Publisher Services
Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 11370918 62/3100/SPI - 5 4 3 2 1 0

Do mo Thuismitheoirí

A.Q.

Preface

Gaussian linear modelling cannot address current signal processing demands. In modern contexts, such as Independent Component Analysis (ICA), progress has been made specifically by imposing non-Gaussian and/or non-linear assumptions. Hence, standard Wiener and Kalman theories no longer enjoy their traditional hegemony in the field, revealing the standard computational engines for these problems. In their place, diverse principles have been explored, leading to a consequent diversity in the implied computational algorithms. The traditional on-line and data-intensive preoccupations of signal processing continue to demand that these algorithms be tractable.

Increasingly, full probability modelling (the so-called *Bayesian* approach)—or partial probability modelling using the likelihood function—is the pathway for design of these algorithms. However, the results are often intractable, and so the area of *distributional approximation* is of increasing relevance in signal processing. The Expectation-Maximization (EM) algorithm and Laplace approximation, for example, are standard approaches to handling difficult models, but these approximations (certainty equivalence, and Gaussian, respectively) are often too drastic to handle the high-dimensional, multi-modal and/or strongly correlated problems that are encountered. Since the 1990s, stochastic simulation methods have come to dominate Bayesian signal processing. Markov Chain Monte Carlo (MCMC) sampling, and related methods, are appreciated for their ability to simulate possibly high-dimensional distributions to arbitrary levels of accuracy. More recently, the particle filtering approach has addressed on-line stochastic simulation. Nevertheless, the wider acceptability of these methods—and, to some extent, Bayesian signal processing itself—has been undermined by the large computational demands they typically make.

The Variational Bayes (VB) method of distributional approximation originates—as does the MCMC method—in statistical physics, in the area known as Mean Field Theory. Its method of approximation is easy to understand: *conditional independence is enforced as a functional constraint in the approximating distribution, and the best such approximation is found by minimization of a Kullback-Leibler divergence (KLD)*. The exact—but intractable—multivariate distribution is therefore factorized into a product of tractable marginal distributions, the so-called *VB-marginals*. This straightforward proposal for approximating a distribution enjoys certain opti-

mality properties. What is of more pragmatic concern to the signal processing community, however, is that the VB-approximation conveniently addresses the following key tasks:

1. *The inference is focused (or, more formally, marginalized) onto selected subsets of parameters of interest in the model:* this one-shot (*i.e.* off-line) use of the VB method can replace numerically intensive marginalization strategies based, for example, on stochastic sampling.
2. *Parameter inferences can be arranged to have an invariant functional form when updated in the light of incoming data:* this leads to feasible on-line tracking algorithms involving the update of fixed- and finite-dimensional statistics. In the language of the Bayesian, *conjugacy* can be achieved under the VB-approximation. There is no reliance on propagating certainty equivalents, stochastically-generated particles, *etc.*

Unusually for a modern Bayesian approach, then, no stochastic sampling is required for the VB method. In its place, the shaping parameters of the VB-marginals are found by iterating a set of implicit equations to convergence. This *Iterative Variational Bayes (IVB) algorithm* enjoys a decisive advantage over the EM algorithm whose computational flow is similar: by design, the VB method yields distributions in place of the point estimates emerging from the EM algorithm. Hence, in common with all Bayesian approaches, the VB method provides, for example, measures of uncertainty for any point estimates of interest, inferences of model order/rank, *etc.*

The machine learning community has led the way in exploiting the VB method in model-based inference, notably in inference for graphical models. It is timely, however, to examine the VB method in the context of signal processing where, to date, little work has been reported. In this book, at all times, we are concerned with the way in which the VB method can lead to the design of *tractable* computational schemes for tasks such as (i) dimensionality reduction, (ii) factor analysis for medical imagery, (iii) on-line filtering of outliers and other non-Gaussian noise processes, (iv) tracking of non-stationary processes, *etc.* Our aim in presenting these VB algorithms is not just to reveal new flows-of-control for these problems, but—perhaps more significantly—to understand the strengths and weaknesses of the VB-approximation in model-based signal processing. In this way, we hope to dismantle the current psychology of dependence in the Bayesian signal processing community on stochastic sampling methods. Without doubt, the ability to model complex problems to arbitrary levels of accuracy will ensure that stochastic sampling methods—such as MCMC—will remain the golden standard for distributional approximation. Notwithstanding this, our purpose here is to show that the VB method of approximation can yield highly effective Bayesian inference algorithms at low computational cost. In showing this, we hope that Bayesian methods might become accessible to a much broader constituency than has been achieved to date.

Praha, Dublin — *Václav Šmídl*
October 2005 — *Anthony Quinn*

Contents

Notational Conventions

Linear Algebra	
$\mathbb{R}$, $\mathbb{X}$, Θ^*	Set of real numbers, set of elements x and set of elements θ, respectively.
x	$x \in \mathbb{R}$, a real scalar.
$A \in \mathbb{R}^{n \times m}$	Matrix of dimensions $n \times m$, generally denoted by a capital letter.
$\boldsymbol{a}_i$, $\boldsymbol{a}_{i,D}$	ith column of matrix A, A_D, respectively.
$a_{i,j}$, $a_{i,j,D}$	(i,j)th element of matrix A, A_D, respectively, $i = 1 \ldots n$, $j = 1 \ldots m$.
b_i, $b_{i,D}$	ith element of vector b, b_D, respectively.
$\operatorname{diag}(\cdot)$	$A = \operatorname{diag}(a)$, $a \in \mathbb{R}^q$, then $a_{i,j} = \begin{cases} a_i & \text{if i=j} \\ 0 & \text{if i}\neq\text{j} \end{cases}$, $i,j = 1,\ldots,q$.
a	Diagonal vector of given matrix A (the context will distinguish this from a scalar, a (see 2nd entry, above)).
$\operatorname{diag}^{-1}(\cdot)$	$a = \operatorname{diag}^{-1}(A)$, $A \in \mathbb{R}^{n \times m}$, then $a = [a_{1,1}, \ldots, a_{q,q}]'$, $q = \min(n,m)$.
$A_{;r}$, $A_{D;r}$	Operator selecting the first r columns of matrix A, A_D, respectively.
$A_{;r,r}$, $A_{D;r,r}$	Operator selecting the $r \times r$ upper-left sub-block of matrix A, A_D, respectively.
$a_{;r}$, $a_{D;r}$	Operator extracting upper length-r sub-vector of vector a, a_D, respectively.
$A_{(r)} \in \mathbb{R}^{n \times m}$	Subscript $_{(r)}$ denotes matrix A with restricted rank, $\operatorname{rank}(A) = r \leq \min(n,m)$.
A'	Transpose of matrix A.
$I_r \in \mathbb{R}^{r \times r}$	Square identity matrix.
$\mathbf{1}_{p,q}$, $\mathbf{0}_{p,q}$	Matrix of size $p \times q$ with all elements equal to one, zero, respectively.
$\operatorname{tr}(A)$	Trace of matrix A.

$\vec{a} = \text{vec}\,(A)$ Operator restructuring elements of $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$ into a vector $\vec{a} = [\boldsymbol{a}_1', \ldots, \boldsymbol{a}_n']'$.

$A = \text{vect}\,(a, p)$ Operator restructuring elements of vector $a \in \mathbb{R}^{pn}$ into matrix $A \in \mathbb{R}^{p \times n}$, as follows:

$$A = \begin{bmatrix} a_1 & a_{p+1} & \cdots & a_{p(n-1)+1} \\ \vdots & \vdots & & \vdots \\ a_p & a_{2p} & \cdots & a_{pn} \end{bmatrix}.$$

$A = U_A L_A V_A'$ Singular Value Decomposition (SVD) of matrix $A \in \mathbb{R}^{n \times m}$. In this monograph, the SVD is expressed in the 'economic' form, where $U_A \in \mathbb{R}^{n \times q}$, $L_A \in \mathbb{R}^{q \times q}$, $V_A \in \mathbb{R}^{m \times q}$, $q = \min(n, m)$.

$[A \otimes B] \in \mathbb{R}^{np \times mq}$ Kronecker product of matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{p \times q}$, such that

$$A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,m}B \\ \vdots & \ddots & \vdots \\ a_{n,1}B & \cdots & a_{n,m}B \end{bmatrix}.$$

$[A \circ B] \in \mathbb{R}^{n \times m}$ Hadamard product of matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times m}$, such that

$$A \circ B = \begin{bmatrix} a_{1,1}b_{1,1} & \cdots & a_{1,m}b_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1}b_{n,1} & \cdots & a_{n,m}b_{n,m} \end{bmatrix}.$$

Set Algebra	
$\{A\}_c$	Set of objects A with cardinality c.
$A^{(i)}$	ith element of set $\{A\}_c$, $i = 1, \ldots, c$.

Analysis	
$\chi_X(\cdot)$	Indicator (characteristic) function of set X.
$\text{erf}\,(x)$	Error function: $\text{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-t^2\right) dt$.
$\ln(A), \exp(A)$	Natural logarithm and exponential of matrix A respectively. Both operations are performed on elements of the matrix (or vector), *e.g.* $$\ln\left([a_1, a_2]'\right) = [\ln a_1, \ln a_2]'.$$
$\Gamma(x)$	Gamma function, $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$, $x > 0$.
$\psi_\Gamma(x)$	Digamma (psi) function, $\psi_\Gamma(x) = \frac{\partial}{\partial x} \ln \Gamma(x)$.

$\Gamma_r\left(\frac{1}{2}p\right)$ Multi-gamma function:

$$\Gamma_r\left(\frac{1}{2}p\right) = \pi^{\frac{1}{4}r(r-1)} \prod_{j=1}^{r} \Gamma\left\{\frac{1}{2}(p-j+1)\right\}, r \leq p$$

${}_0F_1(a, AA')$ Hypergeometric function, ${}_pF_q(\cdot)$, with $p = 0$, $q = 1$, scalar parameter a, and symmetric matrix parameter, AA'.

$\delta(x)$ δ-type function. The exact meaning is determined by the type of the argument, x. If x is a continuous variable, then $\delta(x)$ is the Dirac δ-function:

$$\int_{\mathbb{X}} \delta(x - x_0)\, g(x)\, dx = g(x_0),$$

where $x, x_0 \in \mathbb{X}$. If x is an integer, then $\delta(x)$ is the Kronecker function:

$$\delta(x) = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{otherwise.} \end{cases}.$$

$\epsilon_p(i)$ ith elementary vector of $\mathbb{R}^p$, $i = 1, \ldots, p$:

$$\epsilon_p(i) = [\delta(i-1), \delta(i-2), \ldots, \delta(i-p)]'.$$

$\mathbb{I}_{(a,b]}$ Interval $(a, b]$ in $\mathbb{R}$.

Probability Calculus	
$\Pr(\cdot)$	Probability of given argument.
$f(x\|\theta)$	Distribution of (discrete or continuous) random variable x, conditioned by known θ.
$\breve{f}(x)$	Variable distribution to be optimized ('wildcard' in functional optimization).
$x^{[i]}, f^{[i]}(x)$	x and $f(x)$ in the i-th iteration of an iterative algorithm.
$\hat{\theta}$	Point estimate of unknown parameter θ.
$\mathsf{E}_{f(x)}[\cdot]$	Expected value of argument with respect to distribution, $f(x)$.
$\widehat{g(x)}$	Simplified notation for $\mathsf{E}_{f(x)}[g(x)]$.
$\overline{x}, \underline{x}$	Upper bound, lower bound, respectively, on range of random variable x.
$\mathcal{N}_x(\mu, r)$	Scalar Normal distribution of x with mean value, μ, and variance, r.
$\mathcal{N}_x(\mu, \Sigma)$	Multivariate Normal distribution of x with mean value, μ, and covariance matrix, Σ.
$\mathcal{N}_X(M, \Sigma_p \otimes \Sigma_n)$	Matrix Normal distribution of X with mean value, M, and covariance matrices, Σ_p and Σ_n.

$t\mathcal{N}_x(\mu, r; \mathbb{X})$	Truncated scalar Normal of x, of type $\mathcal{N}(\mu, r)$, confined to support set $\mathbb{X} \subset \mathbb{R}$.
$\mathcal{M}_X(F)$	Von-Mises-Fisher matrix distribution of X with matrix parameter, F.
$\mathcal{G}_x(\alpha, \beta)$	Scalar Gamma distribution of x with parameters, α and β.
$\mathcal{U}_x(\mathbb{X})$	Scalar Uniform distribution of x on the support set $\mathbb{X} \subset \mathbb{R}$.

List of Acronyms

AR	AutoRegressive (model, process)
ARD	Automatic Rank Determination (property)
CDEF	Conjugate (parameter) distribution to a DEF (observation) model
DEF	Dynamic Exponential Family
DEFS	Dynamic Exponential Family with Separable parameters
DEFH	Dynamic Exponential Family with Hidden variables
EAR	Extended AutoRegressive (model, process)
FA	Factor Analysis
FAMIS	Functional Analysis for Medical Image Sequences (model)
FVPCA	Fast Variational Principal Component Analysis (algorithm)
HMM	Hidden Markov Model
HPD	Highest Posterior Density (region)
ICA	Independent Component Analysis
IVB	Iterative Variational Bayes (algorithm)
KF	Kalman Filter
KLD	Kullback-Leibler Divergence
LPF	Low-Pass Filter
FF	Fixed Form (approximation)
MAP	Maximum *A Posteriori*
MCMC	Markov Chain Monte Carlo
MEAR	Mixture-based Extension of the AutoRegressive model
ML	Maximum Likelihood
OVPCA	Orthogonal Variational Principal Component Analysis
PCA	Principal Component Analysis
PE	Prediction Error
PPCA	Probabilistic Principal Component Analysis
QB	Quasi-Bayes
RLS	Recursive Least Squares
RVB	Restricted Variational Bayes

SNR	Signal-to-Noise Ratio
SVD	Singular Value Decomposition
TI	Time-Invariant
TV	Time-Variant
VB	Variational Bayes
VL	Viterbi-Like (algorithm)
VMF	Von-Mises-Fisher (distribution)
VPCA	Variational PCA (algorithm)

1

Introduction

1.1 How to be a Bayesian

In signal processing, as in all quantitative sciences, we are concerned with data, D, and how we can learn about the system or source which generated D. We will often refer to learning as *inference*. In this book, we will model the data parametrically, so that a set, θ, of unknown parameters describes the data-generating system. In deterministic problems, knowledge of θ determines D under some notional rule, $D = g(\theta)$. This accounts for very few of the data contexts in which we must work. In particular, when D is information-bearing, then we must model the uncertainty (sometimes called the *randomness*) of the process. The defining characteristic of Bayesian methods is that we use *probabilities* to quantify our beliefs amid uncertainty, and the calculus of probability to manipulate these quantitative beliefs [1–3]. Hence, our beliefs about the data are completely expressed via the parametric probabilistic *observation model*, $f(D|\theta)$. In this way, knowledge of θ determines our *beliefs* about D, not D themselves.

In practice, the result of an observational experiment is that we are given D, and our problem is to use them to learn about the system—summarized by the unknown parameters, θ—which generated them. This learning amid uncertainty is known as *inductive inference* [3], and it is solved by constructing the distribution $f(\theta|D)$, namely, the distribution which quantifies our *a posteriori* beliefs about the system, given a specific set of data, D. The simple prescription of Bayes' rule solves the implied *inverse problem* [4], allowing us to reverse the order of the conditioning in the observation model, $f(D|\theta)$:

$$f(\theta|D) \propto f(D|\theta)f(\theta). \tag{1.1}$$

Bayes' rule specifies how our prior beliefs, quantified by the *prior distribution*, $f(\theta)$, are updated in the light of D. Hence, a Bayesian treatment requires prior quantification of our beliefs about the unknown parameters, θ, whether or not θ is by nature fixed or randomly realized. The signal processing community, in particular, has been resistant to the philosophy of *strong Bayesian inference* [3], which assigns