

**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 Evaluation

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 Observations

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 Šmíd

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

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media.

springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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.

Contents

| | | |
|----------|--|----|
| 1 | Introduction | 1 |
| 1.1 | How to be a Bayesian | 1 |
| 1.2 | The Variational Bayes (VB) Method | 2 |
| 1.3 | A First Example of the VB Method: Scalar Additive Decomposition | 3 |
| 1.3.1 | A First Choice of Prior | 3 |
| 1.3.2 | The Prior Choice Revisited | 4 |
| 1.4 | The VB Method in its Context | 6 |
| 1.5 | VB as a Distributional Approximation | 8 |
| 1.6 | Layout of the Work | 10 |
| 1.7 | Acknowledgement | 11 |
| 2 | Bayesian Theory | 13 |
| 2.1 | Bayesian Benefits | 13 |
| 2.1.1 | Off-line vs. On-line Parametric Inference | 14 |
| 2.2 | Bayesian Parametric Inference: the Off-Line Case | 15 |
| 2.2.1 | The Subjective Philosophy | 16 |
| 2.2.2 | Posterior Inferences and Decisions | 16 |
| 2.2.3 | Prior Elicitation | 18 |
| 2.2.3.1 | Conjugate priors | 19 |
| 2.3 | Bayesian Parametric Inference: the On-line Case | 19 |
| 2.3.1 | Time-invariant Parameterization | 20 |
| 2.3.2 | Time-variant Parameterization | 20 |
| 2.3.3 | Prediction | 22 |
| 2.4 | Summary | 22 |
| 3 | Off-line Distributional Approximations and the Variational Bayes Method | 25 |
| 3.1 | Distributional Approximation | 25 |
| 3.2 | How to Choose a Distributional Approximation | 26 |
| 3.2.1 | Distributional Approximation as an Optimization Problem .. | 26 |
| 3.2.2 | The Bayesian Approach to Distributional Approximation ... | 27 |

| | | |
|----------|---|-----------|
| 3.3 | The Variational Bayes (VB) Method of Distributional Approximation | 28 |
| 3.3.1 | The VB Theorem | 28 |
| 3.3.2 | The VB Method of Approximation as an Operator | 32 |
| 3.3.3 | The VB Method | 33 |
| 3.3.4 | The VB Method for Scalar Additive Decomposition | 37 |
| 3.4 | VB-related Distributional Approximations | 39 |
| 3.4.1 | Optimization with Minimum-Risk KL Divergence | 39 |
| 3.4.2 | Fixed-form (FF) Approximation | 40 |
| 3.4.3 | Restricted VB (RVB) Approximation | 40 |
| 3.4.3.1 | Adaptation of the VB method for the RVB Approximation | 41 |
| 3.4.3.2 | The Quasi-Bayes (QB) Approximation | 42 |
| 3.4.4 | The Expectation-Maximization (EM) Algorithm | 44 |
| 3.5 | Other Deterministic Distributional Approximations | 45 |
| 3.5.1 | The Certainty Equivalence Approximation | 45 |
| 3.5.2 | The Laplace Approximation | 45 |
| 3.5.3 | The Maximum Entropy (MaxEnt) Approximation | 45 |
| 3.6 | Stochastic Distributional Approximations | 46 |
| 3.6.1 | Distributional Estimation | 47 |
| 3.7 | Example: Scalar Multiplicative Decomposition | 48 |
| 3.7.1 | Classical Modelling | 48 |
| 3.7.2 | The Bayesian Formulation | 48 |
| 3.7.3 | Full Bayesian Solution | 49 |
| 3.7.4 | The Variational Bayes (VB) Approximation | 51 |
| 3.7.5 | Comparison with Other Techniques | 54 |
| 3.8 | Conclusion | 56 |
| 4 | Principal Component Analysis and Matrix Decompositions | 57 |
| 4.1 | Probabilistic Principal Component Analysis (PPCA) | 58 |
| 4.1.1 | Maximum Likelihood (ML) Estimation for the PPCA Model | 59 |
| 4.1.2 | Marginal Likelihood Inference of A | 61 |
| 4.1.3 | Exact Bayesian Analysis | 61 |
| 4.1.4 | The Laplace Approximation | 62 |
| 4.2 | The Variational Bayes (VB) Method for the PPCA Model | 62 |
| 4.3 | Orthogonal Variational PCA (OVPCA) | 69 |
| 4.3.1 | The Orthogonal PPCA Model | 70 |
| 4.3.2 | The VB Method for the Orthogonal PPCA Model | 70 |
| 4.3.3 | Inference of Rank | 77 |
| 4.3.4 | Moments of the Model Parameters | 78 |
| 4.4 | Simulation Studies | 79 |
| 4.4.1 | Convergence to Orthogonal Solutions: VPCA vs. FVPCA | 79 |
| 4.4.2 | Local Minima in FVPCA and OVPCA | 82 |
| 4.4.3 | Comparison of Methods for Inference of Rank | 83 |
| 4.5 | Application: Inference of Rank in a Medical Image Sequence | 85 |
| 4.6 | Conclusion | 87 |

| | | |
|----------|---|-----|
| 5 | Functional Analysis of Medical Image Sequences | 89 |
| 5.1 | A Physical Model for Medical Image Sequences | 90 |
| 5.1.1 | Classical Inference of the Physiological Model | 92 |
| 5.2 | The FAMIS Observation Model | 92 |
| 5.2.1 | Bayesian Inference of FAMIS and Related Models | 94 |
| 5.3 | The VB Method for the FAMIS Model | 94 |
| 5.4 | The VB Method for FAMIS: Alternative Priors | 99 |
| 5.5 | Analysis of Clinical Data Using the FAMIS Model | 102 |
| 5.6 | Conclusion | 107 |
| | | |
| 6 | On-line Inference of Time-Invariant Parameters | 109 |
| 6.1 | Recursive Inference | 110 |
| 6.2 | Bayesian Recursive Inference | 110 |
| 6.2.1 | The Dynamic Exponential Family (DEF) | 112 |
| 6.2.2 | Example: The AutoRegressive (AR) Model | 114 |
| 6.2.3 | Recursive Inference of non-DEF models | 117 |
| 6.3 | The VB Approximation in On-Line Scenarios | 118 |
| 6.3.1 | Scenario I: VB-Marginalization for Conjugate Updates | 118 |
| 6.3.2 | Scenario II: The VB Method in One-Step Approximation | 121 |
| 6.3.3 | Scenario III: Achieving Conjugacy in non-DEF Models via the VB Approximation | 123 |
| 6.3.4 | The VB Method in the On-Line Scenarios | 126 |
| 6.4 | Related Distributional Approximations | 127 |
| 6.4.1 | The Quasi-Bayes (QB) Approximation in On-Line Scenarios | 128 |
| 6.4.2 | Global Approximation via the Geometric Approach | 128 |
| 6.4.3 | One-step Fixed-Form (FF) Approximation | 129 |
| 6.5 | On-line Inference of a Mixture of AutoRegressive (AR) Models | 130 |
| 6.5.1 | The VB Method for AR Mixtures | 130 |
| 6.5.2 | Related Distributional Approximations for AR Mixtures | 133 |
| 6.5.2.1 | The Quasi-Bayes (QB) Approximation | 133 |
| 6.5.2.2 | One-step Fixed-Form (FF) Approximation | 135 |
| 6.5.3 | Simulation Study: On-line Inference of a Static Mixture | 135 |
| 6.5.3.1 | Inference of a Many-Component Mixture | 136 |
| 6.5.3.2 | Inference of a Two-Component Mixture | 136 |
| 6.5.4 | Data-Intensive Applications of Dynamic Mixtures | 139 |
| 6.5.4.1 | Urban Vehicular Traffic Prediction | 141 |
| 6.6 | Conclusion | 143 |
| | | |
| 7 | On-line Inference of Time-Variant Parameters | 145 |
| 7.1 | Exact Bayesian Filtering | 145 |
| 7.2 | The VB-Approximation in Bayesian Filtering | 147 |
| 7.2.1 | The VB method for Bayesian Filtering | 149 |
| 7.3 | Other Approximation Techniques for Bayesian Filtering | 150 |
| 7.3.1 | Restricted VB (RVB) Approximation | 150 |
| 7.3.2 | Particle Filtering | 152 |

| | | |
|----------|---|------------|
| 7.3.3 | Stabilized Forgetting | 153 |
| 7.3.3.1 | The Choice of the Forgetting Factor | 154 |
| 7.4 | The VB-Approximation in Kalman Filtering | 155 |
| 7.4.1 | The VB method | 156 |
| 7.4.2 | Loss of Moment Information in the VB Approximation | 158 |
| 7.5 | VB-Filtering for the Hidden Markov Model (HMM) | 158 |
| 7.5.1 | Exact Bayesian filtering for known T | 159 |
| 7.5.2 | The VB Method for the HMM Model with Known T | 160 |
| 7.5.3 | The VB Method for the HMM Model with Unknown T | 162 |
| 7.5.4 | Other Approximate Inference Techniques | 164 |
| 7.5.4.1 | Particle Filtering | 164 |
| 7.5.4.2 | Certainty Equivalence Approach | 165 |
| 7.5.5 | Simulation Study: Inference of Soft Bits | 166 |
| 7.6 | The VB-Approximation for an Unknown Forgetting Factor | 168 |
| 7.6.1 | Inference of a Univariate AR Model with Time-Variant Parameters | 169 |
| 7.6.2 | Simulation Study: Non-stationary AR Model Inference via Unknown Forgetting | 173 |
| 7.6.2.1 | Inference of an AR Process with Switching Parameters | 173 |
| 7.6.2.2 | Initialization of Inference for a Stationary AR Process | 174 |
| 7.7 | Conclusion | 176 |
| 8 | The Mixture-based Extension of the AR Model (MEAR) | 179 |
| 8.1 | The Extended AR (EAR) Model | 179 |
| 8.1.1 | Bayesian Inference of the EAR Model | 181 |
| 8.1.2 | Computational Issues | 182 |
| 8.2 | The EAR Model with Unknown Transformation: the MEAR Model | 182 |
| 8.3 | The VB Method for the MEAR Model | 183 |
| 8.4 | Related Distributional Approximations for MEAR | 186 |
| 8.4.1 | The Quasi-Bayes (QB) Approximation | 186 |
| 8.4.2 | The Viterbi-Like (VL) Approximation | 187 |
| 8.5 | Computational Issues | 188 |
| 8.6 | The MEAR Model with Time-Variant Parameters | 191 |
| 8.7 | Application: Inference of an AR Model Robust to Outliers | 192 |
| 8.7.1 | Design of the Filter-bank | 192 |
| 8.7.2 | Simulation Study | 193 |
| 8.8 | Application: Inference of an AR Model Robust to Burst Noise | 196 |
| 8.8.1 | Design of the Filter-Bank | 196 |
| 8.8.2 | Simulation Study | 197 |
| 8.8.3 | Application in Speech Reconstruction | 201 |
| 8.9 | Conclusion | 201 |

9 Concluding Remarks 205

 9.1 The VB Method 205

 9.2 Contributions of the Work 206

 9.3 Current Issues 206

 9.4 Future Prospects for the VB Method 207

Required Probability Distributions 209

 A.1 Multivariate Normal distribution 209

 A.2 Matrix Normal distribution 209

 A.3 Normal-inverse-Wishart ($\mathcal{N}i\mathcal{W}_{A,\Omega}$) Distribution 210

 A.4 Truncated Normal Distribution 211

 A.5 Gamma Distribution 212

 A.6 Von Mises-Fisher Matrix distribution 212

 A.6.1 Definition 213

 A.6.2 First Moment 213

 A.6.3 Second Moment and Uncertainty Bounds 214

 A.7 Multinomial Distribution 215

 A.8 Dirichlet Distribution 215

 A.9 Truncated Exponential Distribution 216

References 217

Index 225

Notational Conventions

| Linear Algebra | |
|---------------------------------------|---|
| $\mathbb{R}, \mathbb{X}, \Theta^*$ | 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. |
| $\mathbf{a}_i, \mathbf{a}_{i,D}$ | i th column of matrix A, A_D , respectively. |
| $a_{i,j}, a_{i,j,D}$ | (i, j) th element of matrix A, A_D , respectively, $i = 1 \dots n, j = 1 \dots m$. |
| $b_i, b_{i,D}$ | i th element of vector b, b_D , respectively. |
| $\text{diag}(\cdot)$ | $A = \text{diag}(a), a \in \mathbb{R}^q$, then $a_{i,j} = \begin{cases} a_i & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}, i, j = 1, \dots, q$. |
| a | Diagonal vector of given matrix A (the context will distinguish this from a scalar, a (see 2nd entry, above)). |
| $\text{diag}^{-1}(\cdot)$ | $a = \text{diag}^{-1}(A), A \in \mathbb{R}^{n \times m}$, then $a = [a_{1,1}, \dots, a_{q,q}]'$, $q = \min(n, m)$. |
| $A_{:,r}, A_{D:,r}$ | 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, $\text{rank}(A) = r \leq \min(n, m)$. |
| 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. |
| $\text{tr}(A)$ | Trace of matrix A . |

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

$A = \text{vect}(a, p)$ 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 .
 $A^{(i)}$ i th element of set $\{A\}_c$, $i = 1, \dots, c$.

Analysis

$\chi_X(\cdot)$ Indicator (characteristic) function of set X .
 $\text{erf}(x)$ Error function: $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$.
 $\ln(A), \exp(A)$ Natural logarithm and exponential of matrix A respectively. Both operations are performed on elements of the matrix (or vector), *e.g.*

$$\ln([a_1, a_2]') = [\ln a_1, \ln a_2]'$$

$\Gamma(x)$ 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:

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

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

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

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

Probability Calculus

| | |
|---|---|
| $\text{Pr}(\cdot)$ | Probability of given argument. |
| $f(x \theta)$ | Distribution of (discrete or continuous) random variable x , conditioned by known θ . |
| $\check{f}(x)$ | 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 . |
| $E_{f(x)}[\cdot]$ | Expected value of argument with respect to distribution, $f(x)$. |
| $\widehat{g(x)}$ | Simplified notation for $E_{f(x)}[g(x)]$. |
| \bar{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 . |

XVIII Notational Conventions

| | |
|--------------------------------------|--|
| $t\mathcal{N}_x(\mu, r; \mathbb{X})$ | 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 <i>A Posteriori</i> |
| 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 |

XX List of Acronyms

| | |
|------|---------------------------------|
| 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) |

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