



Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

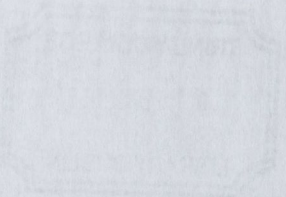
Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The Variational Bayes Approach in Signal Processing

Václav Šmídl



A thesis submitted to the University of Dublin, Trinity College
in fulfillment of the requirements for the degree of
Doctor of Philosophy

April 2004

TRINITY COLLEGE

07 FEB 2005

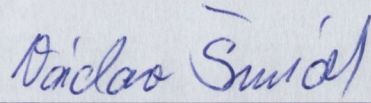
LIBRARY DUBLIN

THOS

7494

Declaration

I, the undersigned, declare that this work has not previously been submitted to this or any other University, and that unless otherwise stated, it is entirely my own work.



Václav Šmíd

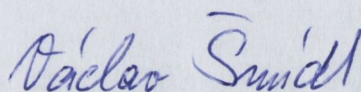
Dated: 1st November 2004

Declaration

Handwritten signature or text, possibly "Handwritten".

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.



Václav Šmíd

Dated: 1st November 2004

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Michael Smith

Author

Dated: 1st November 2004

Summary

This thesis is concerned with Bayesian identification of parameters of linear models. Linear models are used in many important problems of Digital Signal Processing (DSP). Computationally efficient methods of parameter inference are available under certain restrictive assumptions, such as known transformation of system output, known number of signal sources, etc. These assumptions, however, limit the applicability of these models. In this thesis, we study four important special cases of the linear model as listed below. When we relax the restrictive assumption, in each case, the Bayesian inference becomes intractable. Tractability is restored using the Variational Bayes (VB) approximation technique. Special attention is paid to computational efficiency and flow of control of the associated inference algorithms.

Chapters 2 and 3 review the relevant state-of-the-art knowledge. In Chapter 2, the basics of Bayesian parameter inference, and the most common approximation techniques, are reviewed. The Variational Bayes (VB) method is chosen as a reasonable trade-off between accuracy and computational requirements. In Chapter 3, the linear model is introduced and existing Bayesian inference methods are reviewed for this context. At the end of Chapter 3, in Section 3.5, four special cases of the linear model are selected for detailed consideration in the rest of the thesis. For each of these models, a computationally efficient Bayesian inference technique is not currently available and the aim of the thesis is to derive one.

The main contributions of the thesis are presented in Chapters 4–7, in the context of each of these four models:

Chapter 4: the AutoRegressive (AR) model with unknown transformation of its output is studied.

The unknown transformation is approximated by a finite mixture of known candidates. What follows is a new Mixture-based Extension of the AR model (MEAR). Computationally efficient inference algorithms for the MEAR model are derived. The model is successfully applied to the reconstruction of an AR process corrupted by outliers, burst noise, and in a speech reconstruction problem respectively.

Chapter 5: the AR model with *non-stationary* parameters is studied. We relax the assumption of known forgetting factor underlying an established Bayesian approach to this problem. The resulting recursive Bayesian identification algorithm has better tracking ability with respect to the non-stationary parameters. Improvements over the standard fixed-forgetting approach are demonstrated in a simulation study involving an AR process with abrupt change-points.

Chapter 6: the problem of Bayesian Principal Component Analysis (PCA) is studied. The traditional Maximum Likelihood (ML) estimation of model parameters is numerically efficient. However,

it does not provide an estimate of the number of relevant principal components, nor any associated uncertainty bounds. The known Bayesian solutions do not take into account the rotational ambiguity inherent in the model and are, therefore, computationally inefficient. We show that an approximate Bayesian solution can be found with a computational cost comparable to that of the ML solution. This VB solution is potentially attractive in the many scientific areas where PCA is used, but where, currently, inference of rank, and measures of uncertainty are unavailable.

Chapter 7: the problem of functional analysis of medical image data is studied. The standard mathematical model used for this task is reformulated. The complexity of the model demands that the standard approach to parameter estimation is achieved in three separate steps: (i) pre-processing, (ii) orthogonal analysis, and (iii) oblique analysis. We show that the VB-approximate inference unifies all these steps. Moreover, the resulting Bayesian inference solves tasks that were not addressed before, such as selection of the number of relevant physiological sources.

Conclusions and suggestions for further work, are presented in Chapter 8.

Acknowledgement

I am grateful to Anthony Quinn for being my supervisor for this thesis and for his support and inspiration.

I am also thankful to Miroslav Kárný for his insightful thoughts, and to Petr Nedoma for the use of his excellent software MixTools.

I thank Martin Šámal for introducing me to medical imaging, for his support, and for the use of his data.

Finally, I would like to thank my family, for their love and support over the years.

The financial support of the EU IST Project No. 1999-12058, ProDaCTool, is gratefully acknowledged.

it does not provide an estimate of the number of false and principal components, but an estimated uncertainty bounds. The known Bayesian solutions do not take into account the inferential ambiguity inherent in the model and are, therefore, computationally inefficient. We show that an approximate Bayesian solution can be found with a computational cost comparable to that of the ML solution. This VB solution is potentially attractive for large-scale data sets, but where, currently, inference of rank and sparsity of covariance matrices are unavoidable.

Chapter 7: the problem of functional analysis of medical data is studied. The standard functional model used for this task is reformulated. The complexity of the model is reduced and a new model is proposed. The proposed model is simpler and more efficient than the standard model. The proposed model is based on (i) a new approach to the functional analysis, (ii) orthogonal analysis, and (iii) sparse analysis. We show that the proposed model is more efficient than the standard model. Moreover, the proposed Bayesian solution is more efficient than the standard solution. The proposed model is also useful for the functional analysis of medical data. The proposed model is also useful for the functional analysis of medical data.

Conclusions and perspectives are given in Chapter 8. A special thanks to my family for their love and support, and for the financial support of the EU FP7 Project No. 1999-12058 (ProDaTool), is gratefully acknowledged.

Finally, I would like to thank my family for their love and support over the years.

The financial support of the EU FP7 Project No. 1999-12058 (ProDaTool), is gratefully acknowledged.

Contents

Notational Conventions	xvii
List of Acronyms	xix
1. Introduction	1
2. Distributional Approximations in Bayesian Inference	3
2.1. The Basics of Bayesian Theory	3
2.1.1. Choice of prior pdf	5
2.1.2. Conjugate prior pdfs	5
2.2. Off-line Distributional Approximations	6
2.2.1. Certainty Equivalence Approximation	6
2.2.2. Laplace's Approximation	8
2.2.3. Fixed-form Minimum Distance Approximation	8
2.2.4. Variational Bayes (VB) Approximation	9
2.2.5. Quasi-Bayes (QB) Approximation	13
2.2.6. Markov Chain Monte Carlo (MCMC) Approximation	14
2.2.7. Summary	15
2.3. Distributional Approximations for Recursive Identification	15
2.3.1. Bayes-closed Approximation	15
2.3.2. One-step Approximation	16
2.3.3. On-line Variational Bayes	17
3. Linear Models: Classes and Their Inference	19
3.1. Bayesian Methods for Linear Models	19
3.2. The Multivariate AutoRegressive (AR) Model	24
3.2.1. Bayesian Inference of the AR model	25
3.2.1.1. Computational Issues	27
3.2.1.2. Prediction	28
3.2.1.3. Model Structure Determination	28
3.2.2. The Extended AutoRegressive (EAR) Model	29
3.2.3. Modelling Time-Varying Parameters of Non-Stationary AR Models Using Forgetting	31

3.2.3.1.	Explicit Modelling of Parameter Evolution	31
3.2.3.2.	Modelling of Time-varying Parameters via Forgetting	31
3.3.	Probabilistic Principal Component Analysis (PPCA)	33
3.3.1.	Maximum Likelihood Inference	34
3.3.2.	Maximum A Posteriori Inference	35
3.3.3.	Variational Bayes Inference	36
3.4.	Functional Analysis of Medical Image Sequences (FAMIS)	38
3.4.1.	Physiological Model	38
3.4.2.	Data pre-processing	39
3.4.3.	Orthogonal analysis	40
3.4.4.	Oblique analysis	40
3.4.5.	The FAMIS model	41
3.5.	Open problems	41
4.	Mixture-Based Extension of the EAR (MEAR) model	45
4.1.	The MEAR Model	45
4.2.	Bayesian Formulation	47
4.3.	Variational Bayes (VB) Approximation	48
4.3.1.	VB-conjugate Prior	48
4.3.2.	VB-optimized Posterior Distribution	50
4.4.	Quasi-Bayes (QB) Approximation	53
4.4.1.	Fixing the VB-marginal for the Label Field	53
4.4.2.	QB-optimal Posterior Distribution	54
4.5.	Viterbi-Like (VL) approximation	54
4.6.	Inference with the MEAR model	55
4.6.1.	Computational Issues	55
4.6.2.	MEAR-based Prediction	57
4.6.3.	MEAR Model with Non-stationary Parameters	57
4.6.4.	MEAR Model Structure Determination	58
4.7.	Inference of an AR Model Robust to Outliers	58
4.7.1.	Filter-Bank Design	59
4.7.2.	Simulation Study	60
4.8.	Inference of an AR Model Robust to Burst Noise	62
4.8.1.	Filter-Bank Design for Burst Noise	62
4.8.2.	Simulation Study	64
4.9.	Application of the MEAR model in Speech Reconstruction	67
4.10.	Discussion	68
5.	Bayesian Inference of Non-stationary AutoRegressive Models Using Time-variant Forgetting	71
5.1.	Bayesian Formulation	71
5.2.	Variational Bayes (VB) Approximation	72

5.2.1.	VB-conjugate Prior	72
5.2.2.	VB-optimal Posterior Distribution	73
5.3.	Time-variant Forgetting for the MEAR model	75
5.4.	Inference of an AR process with Switching Parameters	76
5.5.	Inference of a Stationary AR Process using Time Variant Forgetting	77
5.6.	Discussion	80
6.	Bayesian Treatment of Principal Component Analysis	81
6.1.	Toy Problem: Scalar Decompositions	81
6.1.1.	Bayesian Formulation	82
6.1.2.	Full Bayesian Solution	83
6.1.3.	Variational Bayes (VB) Approximation	84
6.1.4.	Non-Degenerate Parameterization for the Toy Problem	87
6.1.5.	The Lessons Learnt	89
6.2.	Fast Variational PCA (FVPCA)	90
6.3.	Orthogonal Variational PCA (OVPCA)	93
6.3.1.	Orthogonal Parameterization of the PPCA Model	93
6.3.2.	Bayesian Formulation	94
6.3.3.	Variational Bayes (VB) Approximation	95
6.3.4.	Inference of Rank	100
6.3.5.	Moments of the Model Parameters	101
6.4.	Simulation Studies	102
6.4.1.	VPCA vs. FVPCA	102
6.4.2.	Initial conditions	105
6.4.3.	Comparison of Methods for Inference of Rank	106
6.4.4.	Comparison of Moments	107
6.5.	Discussion	110
7.	Bayesian Modelling for Functional Analysis of Medical Image Data	113
7.1.	Bayesian Formulation	113
7.2.	Variational Bayes (VB) Approximation	114
7.3.	Computational Simplifications	116
7.4.	Experiments	118
7.4.1.	Comparison of Methods for Inference of Rank in Orthogonal Analysis	118
7.4.2.	Variational FAMIS	120
7.5.	Discussion	124
7.5.1.	Model Matching	124
7.5.2.	Consequence for Medical Applications	124
8.	Conclusions	127
8.1.	Discussion of the Work	127
8.2.	Key Contributions of the Thesis	127

8.3. Further work	129
8.3.1. Short Term Extensions	129
8.3.2. Future Research Directions	130
A. Required Probability Distributions	131
A.1. Matrix Normal distribution	131
A.2. Normal-Wishart Distribution	132
A.3. Dirichlet Distribution	133
A.4. Truncated Normal Distribution	134
A.5. Von Mises-Fisher Matrix distribution	134
A.5.1. Definition	134
A.5.2. First Moment	135
A.5.3. Second Moment and Uncertainty Bounds	136
A.6. Gamma Distribution	136
A.7. Truncated Exponential Distribution	137
B. Analytical Solution of Fast VPCA, Using MAPLE	139
B.1. Closed-form Solution	139
B.2. Determination of Acceptable Solutions	140
C. Hypergeometric Functions	145
C.1. Hypergeometric Function of Scalar Argument	145
C.2. Approximation of ${}_0F_1$ of Matrix Argument by ${}_0F_1$ of Scalar Arguments	146
Bibliography	149

List of Figures

3.1. Table of established linear models	22
3.2. Signal flow graphs of the AR and EAR models.	26
4.1. The Recursive summed-dyad computational scheme for Quasi-Bayes identification of the MEAR model.	55
4.2. Reconstruction an AR(2) process corrupted by isolated outliers.	61
4.3. Identification of an AR(2) process corrupted by isolated outliers.	61
4.4. Reconstruction and identification of a non-stationary AR(2) process corrupted by burst noise, using KF variant of the MEAR model.	65
4.5. Reconstruction and identification of a non-stationary AR(2) process corrupted by burst noise, using KF+LPF variant of the MEAR model.	66
4.6. Reconstruction of three sections of the <code>bbcnews.wav</code> speech file.	69
5.1. Results of identification of a non-stationary process using time-variant forgetting.	78
5.2. Results of identification of a stationary process using time-variant forgetting.	79
6.1. Illustration of scaling ambiguity in the toy problem.	83
6.2. Illustration of full Bayesian solution for the toy problem.	84
6.3. Comparison of Laplace and Variational Bayes approximations for the toy problem.	86
6.4. Iterative solution for VB approximation of the toy problem.	87
6.5. Illustration of the VB-posterior for the scalar additive decomposition.	89
6.6. Illustration of properties of von-Mises-Fisher distribution.	102
6.7. Simulated data used for testing of PCA-based inference method.	103
6.8. Monte Carlo study to illustrate convergence of the VPCA algorithm to an orthogonal solution.	104
6.9. Posterior estimates, $\hat{\omega}$, for the data set SIM2 with respect to different initial values $\hat{\omega}^{(0)}$	105
6.10. <i>Ad hoc</i> methods for rank estimation for simulated data with different variance of noise. The methods of visual examination (Remark 3.7) is used for graphs of eigenvalues $\lambda = \text{eig}(DD')$ and cumulative variance (i.e. cumulative sum of eigenvalues λ).	106
7.1. Cumulative percentage of total variation for scintigraphic data.	119
7.2. Expected posterior values of factor images and factor curves for four noise modelling strategies.	121

7.3. Comparison of the posterior expected value of of the precision matrices for different initializations.	122
B.1. Analysis of singular points of roots of the second mode solution of fast VPCA.	142
C.1. Illustration of the hypergeometric function, ${}_0F_1(5, \frac{1}{4}s^2)$	146
A.2. Normal-Wishart Distribution	151
A.3. Dirichlet Distribution	151
A.4. Truncated Normal Distribution	151
A.5. Von Mises-Fisher Matrix	151
A.5.1. Regression	151
A.5.2. Signal flow graphs of the AR and EAR models	151
A.5.3. The Recursive summed-based computational scheme for Least Squares identification of the MEAR model	151
A.5.4. Reconstruction of an AR(2) process corrupted by isolated outliers	151
A.5.5. Identification of a non-stationary AR(2) process corrupted by isolated outliers	151
A.5.6. Reconstruction and identification of a non-stationary AR(2) process corrupted by noise using KF variant of the MEAR model	151
A.5.7. Reconstruction and identification of a non-stationary AR(2) process corrupted by noise using KF+LRF variant of the MEAR model	151
A.5.8. Reconstruction of base sections of the process . way speech signal	151
A.5.9. Identification of base sections of the process . way speech signal	151
A.5.10. Results of identification of base sections of the process using time-varying forgetting	151
A.5.11. Results of identification of base sections of the process using time-varying forgetting	151
A.5.12. Illustration of solving ambiguity in the toy problem	151
A.5.13. Illustration of full Bayesian solution for the toy problem	151
A.5.14. Comparison of Laplace and Variational Bayes approximations for the toy problem	151
A.5.15. Iterative solution for VB approximation of the toy problem	151
A.5.16. Illustration of the VB-posterior for the scalar additive decomposition	151
A.5.17. Illustration of properties of von-Mises-Fisher distribution	151
A.5.18. Simulated data used for testing of PCA-based inference method	151
A.5.19. Monte Carlo study to illustrate convergence of the VPCA algorithm to an orthogonal solution	151
A.5.20. Posterior estimator $\hat{\Sigma}$ for the data set SIM2 with respect to different initial values $\Sigma^{(0)}$	151
A.5.21. ABA method for rank estimation for simulated data with different variance of noise	151
A.5.22. The method of visual examination (Remark B.7) is used for graphs of eigenvalues $\lambda = \text{sig}(\text{DD})$ and cumulative variance (i.e. cumulative sum of eigenvalues λ)	151
A.5.23. Cumulative percentage of total variance for synthetic data	151
A.5.24. Expected posterior values of factor images and factor curves for four noise modeling strategies	151

List of Tables

4.1. Computational complexity of recursive identification algorithms for the MEAR Model.	56
6.1. Comparison of converged values of k_A obtained from VPCA and FVPCA algorithms.	104
6.2. Rank selection for the simulated data using <i>ad hoc</i> methods.	107
6.3. Comparison of formal rank selection methods for simulated data.	107
6.4. Comparison of inference of the diagonal $\mu_{\tilde{M}}$ of the transformed signal \tilde{M} (6.114), using FVPCA and OVPCA.	110
7.1. Comparison of rank selection methods for scintigraphic image data.	119
7.2. Reconstruction of scintigraphic data for different numbers of PCs	120

6.3	Comparison of the posterior expected value of the precision matrices for different initializations	122
6.4	Analysis of singular points of roots of the second mode of fast VPCA	140
6.5	Simulation of the hypergeometric function ${}_2F_1(5, \frac{1}{2}, \frac{3}{2})$	147
7.1	Computational complexity of recursive identification algorithms for the MBR Model	30
6.1	Comparison of converged values of $\hat{\Lambda}_k$ obtained from VPCA and FVPCA algorithms	104
6.2	Rank selection for the simulated data using ad hoc methods	107
6.3	Comparison of formal rank selection methods for simulated data	107
6.4	Comparison of inference of the diagonal μ_{ii} of the transformed signal \tilde{Y} (6.114) using FVPCA and OVPCA	110
7.1	Comparison of rank selection methods for scientific image data	119
7.2	Reconstruction of scientific data for different numbers of PCs	120

Notational Conventions

Linear algebra

\mathfrak{R}	set of real numbers
$A \in \mathfrak{R}^{n \times m}$	matrix of dimensions $n \times m$, generally denoted by a capital letter
$\mathbf{a}_i \in \mathfrak{R}^n$	i th column of matrix A , $i = 1 \dots m$, using bold-face letter
$a_{i,j}, a_{i,j;D}$	(i, j) th element of matrix A , A_D , respectively, $i = 1 \dots n$, $j = 1 \dots m$
$a_i, a_{i;D}$	i th element of vector \mathbf{a} , \mathbf{a}_D , respectively
$A_r, A_{r;D}$	operator selecting the first r columns of matrix A , A_D , respectively
$A_{r,r}, A_{r,r;D}$	operator selecting the $r \times r$ upper-left sub-block of matrix A , A_D , respectively
$\mathbf{a}_r, \mathbf{a}_{r;D}$	operator extracting upper length- r sub-vector of vector \mathbf{a} , \mathbf{a}_D , respectively
$A_{(r)} \in \mathfrak{R}^{n \times m}$	subscript (r) denotes matrix A with restricted rank, $\text{rank}(A) = r \leq \min(n, m)$
A'	transposition of matrix A
$I_r \in \mathfrak{R}^{r \times r}$	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
$\text{diag}(\cdot)$	$A = \text{diag}(\mathbf{a})$, $\mathbf{a} \in \mathfrak{R}^q$, then $a_{i,j} = \begin{cases} a_i & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$, $i, j = 1, \dots, q$
$\text{diag}^{-1}(\cdot)$	pseudo-inverse of $\text{diag}(\cdot)$ operator: $\mathbf{a} = \text{diag}^{-1}(A)$, $A \in \mathfrak{R}^{n \times m}$, then $\mathbf{a} = [a_{1,1}, \dots, a_{q,q}]'$, $q = \min(n, m)$
$\text{tr}(A)$	trace of matrix A
$A = U_A L_A V_A'$	Singular Value Decomposition (SVD) of matrix $A \in \mathfrak{R}^{n \times m}$, where $U_A \in \mathfrak{R}^{n \times q}$, $L_A \in \mathfrak{R}^{q \times q}$, $V_A \in \mathfrak{R}^{m \times q}$, $q = \min(n, m)$. In this thesis, the SVD is expressed in the 'economic' form, i.e. in terms of the only guaranteed non-zero part of L_A , namely the upper-left $q \times q$ diagonal sub-matrix.
$[A \otimes B] \in \mathfrak{R}^{np \times mq}$	Kronecker product of matrices $A \in \mathfrak{R}^{n \times m}$ and $B \in \mathfrak{R}^{p \times q}$, such that
	$A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,m}B \\ \vdots & \ddots & \vdots \\ a_{n,1}B & \cdots & a_{n,m}B \end{bmatrix}$
$[A \circ B] \in \mathfrak{R}^{n \times m}$	Hadamard product of matrices $A \in \mathfrak{R}^{n \times m}$ and $B \in \mathfrak{R}^{n \times m}$, such that

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

Analysis

$\chi(\cdot)$	the indicator (characteristic) function on the argument set
$\operatorname{erf}(x)$	error function: $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$
$\Gamma_r(\frac{1}{2}p)$	Multi-Gamma function, $\Gamma_r(\frac{1}{2}p) = \pi^{\frac{1}{4}r(r-1)} \prod_{j=1}^r \Gamma\{\frac{1}{2}(p-j+1)\}$
${}_0F_1(a, AA')$	hypergeometric function, ${}_pF_q(\cdot)$, with $p = 0$, $q = 1$, scalar parameter a , and symmetric matrix parameter AA'
$\delta(x)$	delta-type function. Exact meaning is determined by the type of the argument, x . If x is a continuous variable, then $\delta(x)$ is the Dirac delta function:

$$\int_x \delta(x - x_0) g(x) dx = g(x_0).$$

If x is a discrete variable, then $\delta(x)$ is the Kronecker function:

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

$\delta_p(i)$	the i th elementary basis vector of length p :
---------------	--

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

Probability calculus

$\operatorname{Pr}(\cdot)$	probability of argument
$f(x \theta)$	probability density function (pdf) of random variable x , conditioned by known θ
$\check{f}(x)$	variable pdf to be optimized ('wildcard' in functional optimization)
$\hat{\theta}$	maximizer of $f(x \theta)$, with latter taken as a function of θ (i.e. the ML estimate)
$E_{f(x)}(\cdot), E_x(\cdot)$	expected value of argument with respect to pdf $f(x)$
$\overline{g(x)}$	simplified notation for $E_x(g(x))$
\bar{x}, \underline{x}	upper bound, lower bound of random variable x , respectively
$\mathcal{N}(\mu, s^2)$	Normal distribution with mean value, μ , and variance, s^2
$t\mathcal{N}(\mu, s^2; (\alpha, \beta))$	truncated Normal of type $\mathcal{N}(\mu, s^2)$, confined to support $(\alpha, \beta]$
$\mathcal{M}(F)$	von-Mises-Fisher distribution with matrix parameter F
$\mathcal{G}(\alpha, \beta)$	Gamma distribution with scalar parameters α and β
$\mathcal{U}(\cdot), \mathcal{U}((\alpha, \beta))$	Uniform distribution on the argument set, on the interval $(\alpha, \beta]$, respectively

List of Acronyms

AR	AutoRegressive (model, process)
ARD	Automatic Relevance Determination (property)
ARMA	AutoRegressive Moving Average (model, process)
ARMMAX	AutoRegressive process with Mixture-of-Moving-Average and eXogenous parts
DSP	Digital Signal Processing
EAR	Extended AutoRegressive (model, process)
FA	Factor Analysis
FVPCA	Fast Variational Principal Component Analysis
GAR	Generalized AutoRegressive (model, process)
GLM	Generalized Linear Model
ICA	Independent Component Analysis
KF	Kalman Filter
KL	Kullback-Leibler (distance)
LPF	Low-Pass Filter
MAP	Maximum <i>A Posteriori</i> Probability
MCMC	Markov Chain Monte Carlo
MEAR	Mixture-based Extension of the AutoRegressive (model)
ML	Maximum Likelihood
OVPCA	Orthogonal Variational Principal Component Analysis
pdf	probability density function
PCA	Principal Component Analysis
PPCA	Probabilistic Principal Component Analysis
QB	Quasi-Bayes
RLS	Recursive Least Squares
RVB	Restricted Variational Bayes
SNR	Signal-to-Noise Ratio
VL	Viterbi-Like
VPCA	Variational PCA

AR	Autoregressive (model process)
ARD	Autoregressive Distributed Lag (model process)
ARMA	Autoregressive Moving Average (model process)
ARMAX	Autoregressive process with Moving-Average and Exogenous input
DSP	Digital Signal Processing
EAR	Extended Autoregressive (model process) or extended factor analysis
FA	Factor Analysis
FVPCA	Fast Variational Principal Component Analysis
GAR	Generalized Autoregressive (model process)
GLM	Generalized Linear Model
ICA	Independent Component Analysis or independent signal processing
KL	Kullback-Leibler Distance
LEP	Linear Expectation Propagation (model process) or linear expectation propagation
MCMC	Markov Chain Monte Carlo
MEAB	Minimum-based Extended Autoregressive (model process)
ML	Maximum Likelihood
OVPCA	Orthogonal Variational Principal Component Analysis
pd	Partial Differential
PCA	Principal Component Analysis
PCCA	Probabilistic PCA
QD	Quasi-Newton
RLS	Recursive Least Squares
RVB	Restricted Variational Bayes
SNR	Signal-to-Noise Ratio
VE	Variational Expectation
VPCA	Variational PCA

Chapter 1.

Introduction

Mathematics looks like a pile of abstract facts, axioms and theorems to most people. It is hard to imagine that in some branches of mathematics, there are unresolved controversies about the meanings of basic notions such as Probability. Statistics is one of these branches, where researchers can be divided into various "schools of thought". This division further propagates into all scientific areas where statistics is applied, notably in (statistical) Digital Signal Processing (DSP).

Traditionally, DSP is dominated by classical methods, such as least squares and maximum likelihood methods, Wiener theory, etc. [1]. The classical interpretation sees each probability as a long-run relative frequency. On the contrary, the Bayesian school sees probability as a quantified degree-of-belief (or plausibility). However, what may sound like a minor philosophical disagreement can lead to very different ways of solving practical problems. In scenarios with plenty of observed data, the differences in results obtained using these philosophies are negligible.

The amount of available data is often limited however. For example, in medical applications, the measurement of data is expensive and may be uncomfortable for the patient. In these cases, classical methods have been found to be unreliable and the Bayesian approach has provided better results [2, 3, 4]. In DSP, these scenarios arise in image processing, analysis of medical data, signal-source separation, and non-stationary processing.

Bayesian inference is analytically tractable only for a limited class of models. The full Bayesian solution for more complicated and realistic models is not tractable and must be approximated. This problem has been studied in many scientific areas and many approximate methods have been proposed. The Markov Chain Monte Carlo (MCMC) approximation is now a popular approximation in DSP and statistics. As an alternative, the method known as Mean Field Theory has been developed in statistical physics [5]. This latter principle was introduced to the machine learning community [6], which developed it as the Variational Bayes (VB) method [7, 8]. Further research into this approach is now an inter-disciplinary activity, ranging over many scientific fields [9]. Its impact in DSP has yet to be felt.

In this thesis, we study the application of the Variational Bayes (VB) method in DSP. The main concern in DSP is with computational efficiency and implementation of inference algorithms. Quite often, implementational restrictions—in terms of memory size and processor speed—must be taken into account when designing a new algorithm. This is very important, for example, in real-time and adaptive signal processing. These implementational restrictions influence the choice of mathematical models, as well as the inference methods employed. The preferred mathematical models are simple—i.e. ones with a small number of parameters and with computationally cheap operations—and primarily of the

1. INTRODUCTION

linear model kind. Computationally cheap inference methods are also preferred, such as least squares methods. However, such methods have limited modelling capabilities and do not perform well in areas such as model order selection, non-stationary processing, and treatment of non-linear distortions. Many extensions of classical linear models have been proposed to address these problems. Inference of parameters of such extended models has proved intractable, and various approximations—often *ad hoc* or heuristic—have had to be made, in order to achieve tractability.

In this thesis, we are concerned with extensions of linear models which allow for computationally efficient Bayesian inference. In order to achieve this, we start with a basic linear model for which a computationally efficient inference is known. Applicability of this model is, however, limited by the restrictive assumptions it depends on. When we relax the assumption, the Bayesian solution becomes intractable. The Variational Bayes (VB) approximation is used to overcome these difficulties. Computational issues in the resulting inference algorithm are addressed and, where possible, computationally efficient solutions are proposed.

This experience encourages us to study a range of important DSP problems where, currently, severe restrictions are needed to achieve tractability. Approximative Bayesian analysis using the VB approach allows us to achieve effective and numerically efficient results in the much broader contexts where these restrictions are relaxed.

Chapter 2.

Distributional Approximations in Bayesian Inference

The Bayesian methodology is a well established approach to statistical inference [10]. It is appreciated as a consistent framework for dealing with uncertainty [11]. While this is also important for DSP, practical benefits are the primary concern in this field. The Bayesian framework has advantages over other approaches, notably in model order selection [12, 13] and decision making [14, 15]. In this Chapter, we briefly review the parts of the theory we need for further development in DSP.

In Section 2.1, we review the basics of Bayesian theory. Two basic scenarios are considered: (i) the off-line scenario, where all data are available for the inference procedure, and (ii) the on-line scenario, where the data are acquired incrementally and the inference is re-evaluated for each new data record. Numerically efficient inference can be achieved only for a limited class of problems. For more complicated models, full Bayesian treatment of a problem may be computationally prohibitive. Therefore, various approximating techniques are employed to lower the computational load.

The approximations used in the off-line scenario are briefly reviewed in Section 2.2, with emphasis on accuracy and computational cost. The Variational Bayes (VB) method (Section 2.2.4) is chosen for further development as a promising compromise between computational requirements and accuracy. Therefore, this method is studied in detail, with proofs of basic theorems. The approximations used in the on-line scenario are reviewed in Section 2.3. Once again, our concern is with the Variational Bayes approximation.

2.1. The Basics of Bayesian Theory

Let the measured data be denoted by D . A parametric probabilistic model of the data is then usually given by the probability density function (pdf), $f(D|\theta)$, conditioned by knowledge of the parameters, θ . In this thesis, we will use notation $f(\cdot)$ for both continuous and discrete parameters. In this way a significant simplification and unification of all formulas can be achieved. One only has to keep in mind that the integration has to be replaced by regular summation wherever the argument is discrete¹.

The basic concept underlying Bayesian theory is the treatment of the unknown parameter θ as a random variable. In this thesis, we suppose that the data D are composed of n p -dimensional data

¹This can also be achieved by employment of measure theory, operating in a consistent way with probability densities generalized in the Radon-Nikodym sense [16]. The practical effect is the same and therefore is neither necessary nor helpful for our purposes.

records, $\mathbf{d}_i \in \mathbb{R}^p$, $i = 1, \dots, n$. These are aggregated together as follows: $D = [\mathbf{d}_1, \dots, \mathbf{d}_n]$. We recognize two scenarios of parameter inference: (i) *off-line* scenario, when the number of observations, n , is fixed and all data are available before the inference procedure, and (ii) *on-line* scenario, when the data are acquired incrementally. In the latter case, the number of data available at each moment has the role of time index t , and the available data at time t are called the data history, $D_t = [\mathbf{d}_1, \dots, \mathbf{d}_t]$.

Bayesian inference of the model parameters, θ , is based on application of Bayes' rule:

$$f(\theta|D) = \frac{f(\theta, D)}{f(D)} = \frac{f(D|\theta) f(\theta)}{\int_{\theta} f(D|\theta) f(\theta) d\theta}. \quad (2.1)$$

Here, $f(\theta|D)$ will be known as the *posterior* distribution, $f(D|\theta)$ as the *observation model*, and $f(\theta)$ as the *prior* distribution of the parameter θ (i.e. initial belief of the parameter distribution without any information from the measured data). $f(D)$ will be referred to as the *normalizing constant*, ζ ,

$$\zeta = f(D) = \int_{\theta} f(D, \theta) d\theta = \int_{\theta} f(D|\theta) f(\theta) d\theta. \quad (2.2)$$

Bayes' rule (2.1) can be re-written as

$$f(\theta|D) = \frac{1}{\zeta} f(D|\theta) f(\theta) \propto f(D|\theta) f(\theta), \quad (2.3)$$

where \propto means equal up to the normalization constant, ζ . The posterior is fully determined by the product $f(D|\theta) f(\theta)$, since the normalization constant follows from the requirement that $f(\theta|D)$ be a pdf; i.e. $\int_{\theta} f(\theta|D) = 1$. Evaluation of the normalizing constant can be computationally expensive, or even intractable. If the normalizing constant (2.2) is not finite, the distribution is called *improper* [17]. The posterior pdf with explicitly known normalization will be called the *full pdf*.

The task of evaluation of the full posterior pdf will be called *parameter identification* in this thesis. We favor this phrase over the alternative—density estimation—used in some decision theory texts [15]. The full posterior distribution is a complete description of uncertainty in parameters of the assumed model. For many practical tasks, we need to derive conditional and marginal distributions of model parameters, and their moments. Consider the model parameters to be partitioned into two subsets $\theta = [\theta_1, \theta_2]$. The marginal distribution of θ_1 is defined as

$$f(\theta_1|D) = \int_{\theta_2} f(\theta_1, \theta_2|D) d\theta_2. \quad (2.4)$$

The moments of the pdf, i.e. expected values of functions of parameters, $g(\theta)$, will be denoted

$$\mathbb{E}_{f(\theta|D)}(g(\theta)) = \int_{\theta} g(\theta) f(\theta|D) d\theta, \quad (2.5)$$

In context, where it is clear which pdf $f(\theta|D)$ is associated with the parameter θ , the notation can be simplified further to $\mathbb{E}_{f(\theta|D)}(g(\theta)) \equiv \mathbb{E}_{\theta|D}(g(\theta)) \equiv \widehat{g(\theta)}$.

2.1.1. Choice of prior pdf

The required prior distribution (2.1) is a function which must be elicited by the designer of the model. It is an important part of the inference problem. Its use has been widely discussed from a philosophical point-of-view. See [17], for example. In this thesis, we are concerned with practical aspects of priors. The prior distribution is used for:

1. supplementing the data in order to reach an estimate, in cases where there is insufficient data and/or a poorly defined model; This will be called *regularization* (via the prior);
2. imposing various restrictions on the parameter θ reflecting physical constraints. For example, if a prior distribution on a subset of parameter support is zero, then the posterior distribution will also be zero on this subset;
3. appropriately acknowledging our prior ignorance about θ . If the data are assumed to be informative enough, we prefer to choose a *non-informative* prior (i.e. prior with minimal impact on the posterior pdf). The choice of *non-informative* priors was studied, for example, in [18].

In this thesis, we use priors in all three roles described above. However, in case 3., we do not perform full analysis of *non-informative* priors. Instead, we choose the form of the prior using other principles (such as conjugacy, to be explained in the next Section) and we achieve non-informativeness by choosing its statistics such that the prior is flat [16]. This choice of prior will be called a *non-committal* prior.

2.1.2. Conjugate prior pdfs

In the on-line scenario, our concern is with the inference of unknown model parameters at all observation times, $t = 1, 2, 3, \dots$. The Bayesian perspective therefore requires evaluation of a probability distribution on these unknowns at all t . Tractability is assured when the form of the posterior distribution is identical for all t . Such a distribution is known as *self-replicating* [16], or *conjugate* [10]. This principle is briefly reviewed in this Section.

The observation model of the data observed at time t is $f(\mathbf{d}_t|\theta, D_{t-1})$, where $D_0 = \{\}$ by assignment. From Bayes' rule:

$$f(\theta|D_t) \propto f(\mathbf{d}_t|\theta, D_{t-1}) f(\theta|D_{t-1}). \quad (2.6)$$

Since (2.6) is recursive, analytical tractability of the update is assured if distributions $f(\theta|D_t)$ and $f(\theta|D_{t-1})$ are of the same form. This is achieved if there exists a mapping, $\mathbf{s}_t = \mathbf{s}(D_t)$, $\mathbf{s} \in \mathbb{R}^q$, satisfying the condition:

$$f(\theta|D_t) = f(\theta|\mathbf{s}_t), \quad (2.7)$$

with $\mathbf{s}(\cdot)$ time-invariant and finite-dimensional, i.e. $q < \infty$. \mathbf{s}_t are known as *the sufficient statistics* at time t [10]. Substituting (2.7) into (2.6), it follows that

$$f(\theta|\mathbf{s}_t) \propto f(\mathbf{d}_t|\theta, D_{t-1}) f(\theta|\mathbf{s}_{t-1}). \quad (2.8)$$

The distribution is then uniquely determined by s_t and the functional recursion (2.6) can be replaced by an algebraic recursion on s_t :

$$s_t = s(s_{t-1}, D_t), \quad t = 1, 2, 3, \dots \quad (2.9)$$

with initializer s_0 being the parameter of the prior $f(\theta|s_0)$, from (2.7). Note that evaluation of (2.9) may be difficult as $s(\cdot)$ is a function of the whole history D_t . A numerically efficient procedure is assured if $s(\cdot)$ is a function of the last observation only:

$$s_t = s(s_{t-1}, d_t), \quad t = 1, 2, 3, \dots \quad (2.10)$$

Then, $f(\theta|\cdot)$ is said to be conjugate to the observation model, $f(d_t|\theta, D_{t-1})$. One consequence of this is seen when $t = 1$ in (2.8). Then, the prior, $f(\theta|s_0)$, must also be conjugate to the observation model.

It has been proven that a conjugate distribution exists for every observation model belonging to the exponential family [19]. In this case, algebraic recursion (2.10) achieves Bayesian identification of θ , $\forall t$, guaranteeing a numerically tractable procedure. If the observation model does not have a conjugate distribution on parameters, the computational complexity of full Bayesian inference is condemned to grow with time t . To restore computational tractability, we seek approximate inference techniques, as discussed in Section 2.3.

2.2. Off-line Distributional Approximations

Tractability of the full Bayesian analysis—i.e. application of Bayes' rule (2.1), normalization (2.2), marginalization (2.4), and evaluation of moments of posterior distributions (2.5)—is assured only for a limited class of models. Numerical integration can be used, but it is often computationally expensive, especially in higher dimensions.

The problem can be avoided by approximating the true posterior distribution by a distribution that is computationally tractable:

$$f(\theta|D) \approx \tilde{f}(\theta|D). \quad (2.11)$$

Then, all subsequent operations, such as normalization, marginalization and evaluation of moments, are performed on the approximating pdf, $\tilde{f}(\theta|D)$. Various approximation strategies have been developed. In this Section, we review the most common approximation techniques.

2.2.1. Certainty Equivalence Approximation

In many engineering problems, dealing with full pdfs is avoided. A point estimate, i.e. one value of parameter θ , is considered as the summarizing result of parameter inference. This approach will be called *parameter estimation* in this thesis.

The point estimate, $\hat{\theta} = \hat{\theta}(D)$, can be interpreted as an extreme approximation of the posterior pdf by the function $\delta(\cdot)$:

$$f(\theta|D) \approx \tilde{f}(\theta|D) = \delta\left(\left(\theta - \hat{\theta}\right) | D\right), \quad (2.12)$$

where $\hat{\theta}$ is the chosen point estimate of parameter θ , and $\delta(x)$ is the Dirac delta function

$$\int_x \delta(x - \hat{x}) g(x) dx = g(\hat{x}),$$

if x is a continuous variable, and the Kronecker function

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

if x is a discrete variable.

This approximation is known as the *certainty equivalence* principle [20]. It remains to determine an optimal value of the point estimate. This value should be optimal with respect to some criterion, popular choices are:

- Maximum *A Posteriori* (MAP) estimate:

$$\hat{\theta} = \arg \max_{\theta} f(\theta|D). \quad (2.13)$$

This approach may be computationally attractive, as we do not need to evaluate the normalizing constant (2.2).

- Mean value:

$$\hat{\theta} \equiv \hat{\theta} = \int_{\theta} \theta f(\theta|D) d\theta.$$

Evaluation of the mean value may be computationally expensive, owing to the required integration. Therefore, further approximation are usually necessary for this approach.

Remark 2.1 (Maximum likelihood (ML) estimation) is a classical method of parameter point estimation [1]. From a Bayesian perspective, ML estimation corresponds to MAP estimation with uniform prior distributions [17, 18]. The philosophical difference between those two methodologies has been discussed in [17].

Remark 2.2 (Approximations of Marginals by Conditionals) In the point-based context, the true marginal distribution (2.4) can be approximated via

$$f(\theta_1|D) \approx f(\theta_1|D, \hat{\theta}_2), \quad (2.14)$$

i.e. the conditional distribution of θ_1 given a fixed estimate of $\hat{\theta}_2$. The choice of point estimate $\hat{\theta}_2$ is, again, subject to the chosen criterion of optimality.

Algorithm 2.1 (Expectation Maximization (EM) algorithm) is a well known algorithm for ML estimation—and by extension for MAP estimation—of model parameters $\theta = [\theta_1, \theta_2]$ [21]. Here, we follow an alternative derivation of EM via distributional approximations [22]. The task is to estimate parameter θ_1 , of the (intractable) marginal distribution (2.4). Using Jensen's inequality, it is possible to obtain a lower bound on (2.4) which is numerically tractable [22]. The resulting inference algorithm is then a cyclic iteration of two basic steps:

E-step: compute approximate distribution of parameter θ_2 , of type (2.14), at iteration i :

$$\tilde{f}^{(i)}(\theta_2|D) \approx f(\theta_2|D, \hat{\theta}_1^{(i-1)}). \quad (2.15)$$

M-step: using approximate distribution from the E-step, find new estimate $\hat{\theta}_1^{(i)}$:

$$\hat{\theta}_1^{(i)} = \arg \max_{\theta_1} \int_{\theta_2} \tilde{f}^{(i)}(\theta_2|D) \ln f(\theta_1, \theta_2, D) d\theta_2. \quad (2.16)$$

It was proven that this algorithm monotonically increases the marginal likelihood, $f(D|\theta_1)$, thus converging to a local maximum [23].

Note that the posterior, $f(\theta|D)$ and the joint distribution $f(\theta, D)$ differ only in the normalization constant (2.3), which is independent of θ . Hence, (2.16) can also be written as a function of $\ln f(\theta_1, \theta_2|D)$ in place of $\ln f(\theta_1, \theta_2, D)$. We prefer to use the form of (2.16), as it is clear that the normalization constant does not have to be known.

2.2.2. Laplace's Approximation

This method is based on local approximation by a Gaussian distribution at the MAP estimate $\hat{\theta}$, of the posterior pdf $f(\theta|D)$ [24], $\theta \in \mathbb{R}^p$.

Formally, Laplace's method approximates the posterior (2.1) as follows

$$f(\theta|D) \approx \mathcal{N}(\hat{\theta}, H^{-1}) \quad (2.17)$$

where $\hat{\theta}$ is the MAP estimate (2.13), and $H \in \mathbb{R}^{p \times p}$ is the (negative) Hessian matrix of the logarithm of the joint pdf $f(\theta, D)$ with respect to θ , evaluated at $\theta = \hat{\theta}$,

$$H = - \left[\frac{\partial^2 \log f(\theta, D)}{\partial \theta_i \partial \theta_j} \right]_{\theta = \hat{\theta}}, \quad i, j = 1, \dots, p, \quad (2.18)$$

The asymptotic error of approximation was studied in [24].

2.2.3. Fixed-form Minimum Distance Approximation

The approximating distribution $\tilde{f}(\theta|\eta)$ is chosen as a tractable distribution with parameter η . The optimal approximation $\tilde{f}(\theta|\hat{\eta})$ —given the fixed-form function $\tilde{f}(\cdot)$ —is then determined as

$$\hat{\eta} = \arg \min_{\eta} \Delta(\tilde{f}(\theta|\eta), f(\theta|D)), \quad (2.19)$$

where $\Delta(\tilde{f}(\cdot), f(\cdot))$ is an appropriate measure of distance between two pdfs. Various measures are used for specific problems, such as Kullback-Leibler, Levy, chi-squared, L_2 -norm, etc. These are

reviewed in [25]. Specifically, the Kullback-Leibler (KL) distance [26] from $f(\theta|D)$ to $\tilde{f}(\theta|\eta)$,

$$KL\left(f(\theta|D) \parallel \tilde{f}(\theta|\eta)\right) = \int_{\theta} f(\theta|D) \ln \frac{f(\theta|D)}{\tilde{f}(\theta|\eta)} d\theta = E_{f(\theta|D)} \left(\ln \frac{f(\theta|D)}{\tilde{f}(\theta|\eta)} \right), \quad (2.20)$$

is important for two reasons:

1. statistical inference via KL distance was shown to be optimal in statistical utility sense [27].
2. minimization (2.19) with respect the KL distance (2.20) has a unique—and therefore global—solution [28].

Moreover, the KL distance is also used in many practical applications [29, 30, 31]. It has the following properties:

1. $KL\left(f(\theta|D) \parallel \tilde{f}(\theta|\eta)\right) \geq 0$;
2. $KL\left(f(\theta|D) \parallel \tilde{f}(\theta|\eta)\right) = 0$ iff $f(\theta|D) = \tilde{f}(\theta|\eta)$ almost everywhere;
3. $KL\left(f(\theta|D) \parallel \tilde{f}(\theta|\eta)\right) = \infty$ iff on a set of a positive measure $f(\theta|D) > 0$ and $\tilde{f}(\theta|\eta) = 0$;
4. $KL\left(f(\theta|D) \parallel \tilde{f}(\theta|\eta)\right) \neq KL\left(\tilde{f}(\theta|\eta) \parallel f(\theta|D)\right)$ and KL distance does not obey the triangle inequality.

Given 4., care is needed in the syntax describing $KL(\cdot)$. We say that (2.20) is *from* $f(\theta|D)$ *to* $\tilde{f}(\theta|\eta)$.

2.2.4. Variational Bayes (VB) Approximation

Variational Bayes (VB) approximation [7, 4, 8] is also known as Ensemble Learning [32, 33], or naïve Mean Field Theory [5, 34]. Here, we prefer its interpretation as *functional* minimization of the Kullback-Leibler (KL) distance. Compared to *fixed-form* minimum distance approximation (2.19) there are two key differences:

1. the approximating distribution is not confined to a given form, but it is restricted functionally, using the assumption of conditional independence:

$$f(\theta|D) \approx \check{f}(\theta|D) = \check{f}(\theta_1|D) \check{f}(\theta_2|D) \dots \check{f}(\theta_q|D), \quad (2.21)$$

where $\theta = [\theta'_1, \theta'_2, \dots, \theta'_q]'$ is the multivariate parameter partitioned into q elements. Notation $\check{f}(\cdot)$ is used to denote an unspecified functional variant ('wildcard' function) used in optimization procedure which yield the approximating distribution.

2. for reasons of tractability, the VB procedure does not minimize the 'original' KL distance from $f(\theta|D)$ to $\check{f}(\theta|\eta)$ (2.20) but the 'reverse' KL distance $KL\left(\check{f}(\theta|D) \parallel f(\theta|D)\right)$, from $\check{f}(\theta|D)$ to $f(\theta|\eta)$.

These have, respectively, the following consequences:

1. conditional independence:

- the VB approximation can be used only for models with more than one parameter,
- cross-correlation between variables θ_1 and θ_2 is not modelled. Intuitively, the correlated multivariate distribution is modelled as a product of approximating marginals.

2. the use of ‘reverse’ KL distance:

- from property 4. of the KL distance (Section 2.2.3), the ‘reverse’ KL distance is not equal to the ‘original’ one and therefore, it is *less* optimal in the statistical utility sense [27].
- minimum distance approximation via $KL\left(\check{f}(\cdot) || f(\cdot)\right)$ is not guaranteed to have a unique minimum [28].

These disadvantages are, however, outweighed by computational advantages: (i) functional (i.e. free form) optimization has an analytical solution, and (ii) parameters of the optimal approximating posteriors can be evaluated using an alternating algorithm of the EM kind (Algorithm 2.1). These advantages are now described in detail.

Theorem 2.1 (Variational Bayes) *Let $f(\theta|D)$ be the posterior pdf of multivariate parameter θ . The parameter θ is partitioned into $\theta = [\theta'_1, \theta'_2, \dots, \theta'_q]'$. Let $\check{f}(\theta|D)$ be an approximate pdf restricted to the set of conditionally independent distributions on $\theta_1, \theta_2, \dots, \theta_q$:*

$$\check{f}(\theta|D) = \check{f}(\theta_1, \theta_2, \dots, \theta_q|D) = \prod_{i=1}^q \check{f}_i(\theta_i|D). \quad (2.22)$$

Then, the minimum of the KL distance,

$$\tilde{f}(\theta|D) = \arg \min_{\check{f}(\cdot)} KL\left(\check{f}(\theta|D) || f(\theta|D)\right), \quad (2.23)$$

is reached for

$$\tilde{f}_i(\theta_i|D) \propto \exp\left(\mathbb{E}_{\tilde{f}_{/i}(\theta_{/i}|D)}(\ln(f(\theta, D)))\right), \quad i = 1, \dots, q, \quad (2.24)$$

where $\theta_{/i}$ denotes the complement of θ_i in θ , and $\tilde{f}_{/i}(\theta_{/i}|D) = \prod_{j=1, j \neq i}^q \tilde{f}_j(\theta_j|D)$. We will refer to $\tilde{f}(\theta|D)$ (non-unique, see 2. above) as the Variational Extreme. Conditionally independent elements of (2.24) will be called VB-marginals. The parameters of the posterior distributions (2.24) will be called VB-statistics. At the extreme (2.24), the KL distance from the approximant, $\tilde{f}(\cdot)$, to the true posterior, $f(\cdot)$, is

$$KL\left(\tilde{f}(\theta|D) || f(\theta|D)\right) = \ln f(D) - \ln(\zeta_i) + \mathbb{E}_{\tilde{f}_{/i}(\cdot)}\left(\ln\left(\tilde{f}_{/i}(\cdot)\right)\right). \quad (2.25)$$

for any $i \in \{1, \dots, q\}$, where $\zeta_i = \int_{\theta_i} \exp \mathbb{E}_{\tilde{f}_{/i}(\theta_{/i}|D)}(\ln f(\theta, D)) d\theta_i$.

Proof: The KL distance from (2.23) can be rewritten as follows:

$$\begin{aligned}
 KL(\check{f}(\theta|D) || f(\theta|D)) &= \\
 &= \int_{\theta} \check{f}_i(\theta_i|D) \check{f}_{/i}(\theta_{/i}|D) \ln \frac{\check{f}_i(\theta_i|D) \check{f}_{/i}(\theta_{/i}|D)}{f(\theta|D)} \frac{f(D)}{f(D)} d\theta \\
 &= \int_{\theta} \check{f}_i(\theta_i|D) \check{f}_{/i}(\theta_{/i}|D) \ln \check{f}_i(\theta_i|D) d\theta \\
 &\quad - \int_{\theta} \check{f}_i(\theta_i|D) \check{f}_{/i}(\theta_{/i}|D) \ln f(\theta, D) d\theta + \\
 &\quad + \int_{\theta} \check{f}_i(\theta_i|D) \check{f}_{/i}(\theta_{/i}|D) \left[\ln \check{f}_{/i}(\theta_{/i}|D) + \ln f(D) \right] d\theta \\
 &= \int_{\theta_i} \check{f}_i(\theta_i|D) \ln \check{f}_i(\theta_i|D) d\theta_i + \ln f(D) + \eta \\
 &\quad - \int_{\theta_i} \check{f}_i(\theta_i|D) \left[\int_{\theta_{/i}} \check{f}_{/i}(\theta_{/i}|D) \ln f(\theta, D) d\theta_{/i} \right] d\theta_i \quad . \quad (2.26)
 \end{aligned}$$

Here, $\eta = E_{\check{f}_{/i}(\cdot)} \left(\ln \left(\check{f}_{/i}(\cdot) \right) \right)$. For any non-zero scalar ζ_i it holds:

$$\begin{aligned}
 KL(\check{f}(\theta|D) || f(\theta|D)) &= \\
 &= \int_{\theta_i} \check{f}_i(\theta_i|D) \ln \check{f}_i(\theta_i|D) d\theta_i + \ln f(D) + \eta \\
 &\quad - \int_{\theta_i} \check{f}_i(\theta_i|D) \left\{ \ln \left[\frac{\zeta_i}{\check{\zeta}_i} \exp E_{\check{f}_{/i}(\theta_{/i}|D)} (\ln f(\theta, D)) \right] \right\} d\theta_i \\
 &= \int_{\theta_i} \check{f}_i(\theta_i|D) \ln \frac{\check{f}_i(\theta_i|D)}{\frac{1}{\check{\zeta}_i} \exp E_{\check{f}_{/i}(\theta_{/i}|D)} (\ln f(\theta, D))} d\theta_i + \ln f(D) - \ln(\zeta_i) + \eta, \quad (2.27)
 \end{aligned}$$

If ζ_i is chosen as the following normalizing coefficient of $\exp E(\cdot)$,

$$\zeta_i = \int_{\theta_i} \exp E_{\check{f}_{/i}(\theta_{/i}|D)} (\ln f(\theta, D)) d\theta_i,$$

the last equality in (2.27) can be rewritten in terms of a KL distance:

$$\begin{aligned}
 KL(\check{f}(\theta|D) || f(\theta|D)) &= KL \left(\check{f}_i(\theta_i|D) || \frac{1}{\check{\zeta}_i} \exp E_{\check{f}_{/i}(\theta_{/i}|D)} (\ln f(\theta, D)) \right) \\
 &\quad + \ln f(D) - \ln(\zeta_i) + \eta. \quad (2.28)
 \end{aligned}$$

The only term on the right hand side of (2.28) dependent on $\check{f}_i(\cdot)$ is the KL distance. Hence, minimization of (2.28) with respect to $\check{f}_i(\theta_i|D)$, keeping $\check{f}_{/i}(\theta_{/i}|D)$ fixed, is achieved by minimization of the first term. Invoking non-negativity (Property 1) of the KL distance (Section 2.2.3), the minimum of the first term is zero. The minimizer is almost surely $\check{f}_i(\theta_i|D) = \tilde{f}_i(\theta_i|D) \propto \exp \left(E_{\check{f}_{/i}(\theta_{/i}|D)} (\ln f(\theta, D)) \right)$, (i.e. (2.24)), using the second property of the KL distance (Section 2.2.3). ■

This proof is—our own—simpler alternative to the proofs in the literature, which use Lagrange multipliers [33].

The extreme (2.24) is dependent on the data via the joint distribution $f(\theta, D)$, not $f(\theta|D)$. This is important, as the expensive normalization (2.2) is thereby avoided.

The main computational problem of the VB approximation is that the Variational Extreme (2.24) is not given in closed-form. For example, with $q = 2$, the moments of $\tilde{f}_1(\cdot)$, are needed for evaluation of $\tilde{f}_2(\cdot)$, and vice-versa. The solution of (2.24) is usually found via an iterative algorithm that is suggestive of the EM algorithm (Remark 2.1), but where all steps involve expectations of the kind in (2.16), as follows.

Algorithm 2.2 (Variational EM (VEM)) Consider the case where $q = 2$, i.e. $\theta = [\theta'_1, \theta'_2]'$, then cyclic iteration of the following steps, $i = 1, 2, \dots$, converge to a VB extreme (2.24).

E-step: compute approximate distribution of parameter θ_2 at iteration i :

$$\tilde{f}_2^{(i)}(\theta_2|D) \propto \exp \int_{\theta_1} \tilde{f}_2^{(i-1)}(\theta_1|D) \ln f(\theta_1, \theta_2, D) d\theta_1. \quad (2.29)$$

M-step: using approximate distribution from the i th E-step compute approximate distribution of parameter θ_1 at iteration i :

$$\tilde{f}_1^{(i)}(\theta_1|D) \propto \exp \int_{\theta_2} \tilde{f}_1^{(i)}(\theta_2|D) \ln f(\theta_1, \theta_2, D) d\theta_2. \quad (2.30)$$

Where the initializers, i.e. VB-statistics of $\tilde{f}_1^{(0)}(\cdot)$ and $\tilde{f}_2^{(0)}(\cdot)$, may be chosen randomly. Convergence of the algorithm to fixed VB-marginals, $\tilde{f}_i^{(i)}(\theta_i|D)$, $\forall i$, was proven in [35] via natural gradient technique [36].

In general, the algorithm requires q steps—one for each θ_i , $i = 1, \dots, q$ —in each iteration. Following the nomenclature of the ‘EM algorithm’, this algorithm should be called an ‘E^q algorithm’. However, we will use the name Variational EM (VEM) for compatibility with other publications, e.g. [37].

Remark 2.3 (Marginal Lower Bound) An alternative derivation of (2.24) is via the marginal posterior distribution of data [7]. For an arbitrary approximating density, $\check{f}(\theta|D)$, it is true that

$$\begin{aligned} \ln f(D) &= \ln \int_{\theta} f(\theta, D) d\theta = \ln \int_{\theta} \frac{\check{f}(\theta|D)}{\check{f}(\theta|D)} f(\theta, D) d\theta, \\ &\geq \int_{\theta} \check{f}(\theta|D) \ln \frac{f(\theta|D) f(D)}{\check{f}(\theta|D)} d\theta, \end{aligned} \quad (2.31)$$

$$= \ln f(D) - KL(\check{f}(\theta|D) || f(\theta|D)), \quad (2.32)$$

using Jensen’s inequality [7]. Minimizing the KL distance on the right-hand side of (2.31)—e.g. via the VB procedure (Theorem 2.1)—error in the approximation

$$f(D) \approx \exp \int_{\theta} \check{f}(\theta|D) \ln \frac{f(\theta, D)}{\check{f}(\theta|D)} d\theta, \quad (2.33)$$

is minimized.

2.2.5. Quasi-Bayes (QB) Approximation

The iterative evaluation of the Variational Extreme via the VEM algorithm (Algorithm 2.2) may be prohibitive, e.g. in the on-line scenario. Therefore, we seek a modification of the original Variational Bayes approximation that yields a *closed-form* solution.

Corollary 2.1 (of Theorem 2.1, Restricted Variational Bayes (RVB)) Let $f(\theta|D)$ be the posterior pdf of multivariate parameter $\theta = [\theta'_1, \theta'_2, \dots, \theta'_q]'$, and $\bar{f}_{/1}(\theta_{/1}|D)$ be a fixed posterior distribution of $\theta_{/1} = [\theta'_2, \dots, \theta'_q]'$. Let $\check{f}(\theta|D)$ be a conditionally independent approximation of $f(\theta|D)$ of the kind

$$\check{f}(\theta|D) = \check{f}(\theta_1, \theta_2, \dots, \theta_q|D) = \check{f}_1(\theta_1|D) \bar{f}_{/1}(\theta_{/1}|D). \quad (2.34)$$

Then, the minimum of the KL distance, $KL(\check{f}(\theta|D) || f(\theta|D))$, is reached for

$$\check{f}_1(\theta_1|D) \propto \exp\left(\mathbb{E}_{\bar{f}_{/1}(\theta_{/1}|D)}(\ln(f(\theta, D)))\right). \quad (2.35)$$

Proof: Follows directly from (2.24), for choice $i = 1$. ■

Note that Corollary 2.1 is equivalent to the first step of the VEM algorithm (Algorithm 2.2). However, with distribution $\bar{f}_{/1}(\theta_{/1}|D)$ being known, the equation (2.35) is a *closed-form* solution. This greatly reduce the computational load needed for evaluation, since no iterations are required. Since $\bar{f}_{/1}(\theta_{/1})$ is fixed, the KL distance of Variational Extreme (2.24) is less than or equal to the KL distance of the RVB minimum (2.35). These distances can be compared via (2.25).

The quality of the approximation strongly depends on the choice of the fixed approximating distribution $\bar{f}_{/1}(\cdot)$ in (2.34). If $\bar{f}_{/1}(\cdot)$ is chosen close to the VB-optimal posterior (2.24), i.e. $\bar{f}_{/1}(\cdot) \approx \check{f}_{/1}(\cdot)$, then one step of the RVB algorithm avoids many iterations of the original VEM algorithm. Here, we propose one such strategy for choice of $\bar{f}_{/1}(\cdot)$.

Remark 2.4 (Quasi-Bayes (QB)) RVB solution (2.35) holds for any choice of distribution $\bar{f}_{/1}(\theta_{/1}|D)$. We seek a reasonable choice for this function. We choose $q = 2$ for notational clarity, i.e. $\bar{f}_{/1}(\theta_{/1}) = \bar{f}_{/2}(\theta_2)$, however the result is also valid for the general case.

The VB extreme (2.23) is

$$\check{f}_2(\theta_2|D) = \arg \min_{\check{f}_2} \left(\min_{\check{f}_1} KL(\check{f}(\theta|D) || f(\theta|D)) \right). \quad (2.36)$$

As noted (Algorithm 2.2), a solution to (2.36) cannot be found in closed form. Hence we seek a reasonable ‘first guess’. Rewriting the KL distance in (2.36) as

$$\begin{aligned} KL\left(\check{f}(\theta|D) || f(\theta|D)\right) &= \int_{\theta} \check{f}_1(\theta_1|D) \check{f}_2(\theta_2|D) \ln \frac{\check{f}_1(\theta_1|D) \check{f}_2(\theta_2|D)}{f(\theta_1|\theta_2, D) f(\theta_2|D)} d\theta, \\ &= \int_{\theta} \check{f}_1(\theta_1|D) \check{f}_2(\theta_2|D) \ln \frac{\check{f}_1(\theta_1|D)}{f(\theta_1|\theta_2, D)} d\theta \\ &\quad + \int_{\theta_2} \check{f}_2(\theta_2|D) \ln \frac{\check{f}_2(\theta_2|D)}{f(\theta_2|D)} d\theta_2. \end{aligned} \quad (2.37)$$

we note that the second term in (2.37) is $KL\left(\check{f}_2(\theta_2|D) || f(\theta_2|D)\right)$ which is minimized for

$$\bar{f}_2(\theta_2|D) \equiv f(\theta_2|D) = \int_{\theta_1} f(\theta|D) d\theta_1, \quad (2.38)$$

i.e. exact marginal distribution of $\bar{f}_2(\theta_2|D) \equiv f(\theta_2|D) = \int_{\theta_1} f(\theta|D) d\theta_1$, minimum of (2.37) is not reached as the first term in (2.37) is we consider (2.38) as the best analytical choice we can make.

The name Quasi-Bayes (QB) was first used in the context of finite mixture models [25]. There, the choice of (2.14) for the approximation was based on point-estimation arguments (Remark 2.2), choosing $\hat{\theta}_2$ as the expected value of the true posterior marginal:

$$\hat{\theta}_2 = E_{\theta_2|D}(\theta_2). \quad (2.39)$$

Note that, iff $\ln f(\theta_1, \theta_2, D)$ is linear in θ_2 , then, using (2.35), the RVB approximation (Corollary 2.1) yields results equivalent to (2.14) [37]. The choice of $f_2(\theta_2)$ (Remark 2.4) yields (2.39). Therefore, we consider Remark 2.4 to be a generalization of the QB idea expressed in [25].

2.2.6. Markov Chain Monte Carlo (MCMC) Approximation

In this approach, the posterior pdf is approximated by a piece wise constant density on a partitioned support, i.e. via a histogram constructed from a sequence of random samples, $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)}, \dots\}$, of variable θ .

The sequence of random samples is called a Markov chain if the n -th sample $\theta^{(n)}$ is generated from a chosen conditional distribution

$$f\left(\theta^{(n)} | \theta^{(n-1)}\right) \quad (2.40)$$

which depends only upon the previous state of the chain $\theta^{(n-1)}$.

For mild regularity conditions on $f(\cdot|\cdot)$ (2.40), then, as $n \rightarrow \infty$, $\theta^{(n)} \sim f_s(\theta)$, the (time-invariant) stable distribution of the Markov chain defined via the kernel (2.40). Hence i.i.d. samples from $f_s(\theta)$ may be drawn via an appropriate choice of kernel (2.40), if n is chosen sufficiently large. Typically, the associated computational burden is high, especially for high-dimensional parameters.

2.2.7. Summary

The methods described in this Chapter were ordered with increasing complexity and accuracy of approximation. In signal processing, the full distribution must often be collapsed to a point estimate in order to complete a typical task. Therefore, in many applications, point estimates are evaluated without any reference to their full posterior distribution. The Laplace approximation is known in the DSP community in the context of criteria for model order selection, such as the Schwarz criterion or Bayes Information Criterion (BIC), both of which were derived using the Laplace method [24]. Sampling methods—e.g. MCMC—are valued for their ability to provide arbitrarily close approximation. However, for closer and closer approximation, more and more computational power is required. Thus, this approximation is mostly used for low-dimensional problems evaluated off-line (e.g. [38]).

In this thesis, we are concerned mostly with the Variational Bayes approximation. The main advantage of the VB approximation is its ease of use. Note that the form of the approximate posterior distribution is found explicitly. Evaluation of the VB-statistics of these posterior distributions can be achieved by a general iterative algorithm (Algorithm 2.2). Therefore, we see VB as a good starting point in the search for an optimal trade-off between accuracy and computational complexity of the identification procedure. If the accuracy of the VB-posterior distributions is not acceptable, we can use more sophisticated (and thus more computationally expensive) approximations (e.g. mean field theory [34, 39], or sampling methods (MCMC)). In this thesis, we assume that the accuracy of the VB approximation is acceptable, while the computational cost of the iterative VEM algorithm is not acceptable. Therefore, much of our effort will be dedicated to studying further simplifications and approximations of the implied VB-posteriors.

2.3. Distributional Approximations for Recursive Identification

If the observation pdf does not belong to the exponential family—i.e. finite-dimensional sufficient statistics is not available—the full history of data has to be used in each step. Hence, computational complexity grows with each step. To achieve computational tractability, we have to find an approximate representation of the data history at each step. In contrast to the previous Section, we do not seek a single approximation (2.11), but a sequence of approximations:

$$f(\theta|D_t) \approx \tilde{f}(\theta|D_t), \quad t = 1, \dots, \infty \quad (2.41)$$

In this Section, we review the relevant approaches to this problem. The following list is by no means complete. It is provided, by way of introduction to the VB approximation, which is the main subject of the thesis.

2.3.1. Bayes-closed Approximation

The problem of recursive estimation with limited memory was addressed in general in [40]. There, the problem was defined as finding a functional form, $\check{f}(\theta)$, of approximating distributions that is closed

under Bayes' rule, i.e.

$$\check{f}(\theta|D_t) \propto f(\mathbf{d}_t|\theta, D_{t-1}) \check{f}(\theta|D_{t-1}). \quad (2.42)$$

where $\check{f}(\theta|D_{t-1})$ and $\check{f}(\theta|D_t)$ are of the same functional form. Moreover, the form must depend only on a finite-dimensional statistics, $\mathbf{s}_t \in \mathbb{R}^{q \times 1}$, such as

$$\check{f}(\theta|D_t) = \check{f}(\theta|\mathbf{s}_t),$$

where q is assigned, and may be chosen arbitrarily small. Note that \mathbf{s}_t plays the role of sufficient statistics, however, in this case it is *not* sufficient for full description of the posterior.

The approximating family was found in the form of probabilistic mixture of q fixed (known) pdfs $\bar{f}_i(\theta)$, $i = 1, \dots, q$, weighted by elements of \mathbf{s}_t . Non-sufficient statistics \mathbf{s}_t is then updated by a linear functional, $l(\cdot)$,

$$s_{i,t} = s_{i,t-1} + l(\bar{f}_i(\theta), f(\mathbf{d}_t, \theta|D_{t-1})), \quad i = 1, \dots, q. \quad (2.43)$$

Alternatively, the choice of q fixed pdfs, $\bar{f}_i(\theta)$, can be replaced by the choice of q functionals $l_i(\cdot)$, such as

$$s_{i,t} = s_{i,t-1} + l_i(f(\mathbf{d}_t, \theta|D_{t-1})).$$

It was proven then the approximate on-line identification (2.42) is *globally optimal*², [41].

Practical use of the approximation is, however, rather limited. The method requires time- and data-invariant linear operators, $l_i(\cdot)$ to be chosen *a priori*. Design criteria for these operators are available only for special cases. The method was demonstrated to be applicable to low-dimensional problems only.

2.3.2. One-step Approximation

In this case, the requirement for the approximation family to be closed under Bayes' rule is relaxed. The form of the posterior, $\bar{f}(\theta|D_t)$, is given *a priori* and fixed for all t . It is the Bayes' rule what is approximated at each step [25, 42]. If the posterior distribution, $f(\theta|D_t)$, has a form different from the prior, $f(\theta|D_{t-1})$, an approximation of the posterior is found in the family of the prior distribution

$$\bar{f}(\theta|D_t) \approx f(\theta|D_t) \propto f(\mathbf{d}_t|\theta, D_{t-1}) \bar{f}(\theta|D_{t-1}) \quad (2.44)$$

The approximation (2.44) is used as prior in the next step. As the efficient recursive estimation can be achieved only for the exponential family, $\bar{f}(\cdot)$ is chosen from this family.

There are two basic approaches to the choice of the approximation in (2.44):

1. probability fitting: the approximation (2.44) is optimized with respect to a chosen distance (Section 2.2.3).
2. moment fitting (also known as the probabilistic editor [25]): parameters of the approximating distribution are chosen so that moments of the approximating distribution match moments of the true posterior.

²with respect to orthogonal projection on the true posterior distribution.

Note that one-step approximation is only *locally* optimal (i.e. optimal only for one step, not for the whole trajectory), and so the error of approximation may grow with time. Typically, the quality of the approximation is studied asymptotically, i.e. for $t \rightarrow \infty$. Furthermore, the approximation is not closed under Bayes' rule. In practice, this means that on-line identification given a set of i.i.d. observations yields different results depending on the order in which the data are processed [25].

2.3.3. On-line Variational Bayes

The general VB approximation (Section 2.2.4) was extended to the on-line scenario in [35]. It is found that the on-line VB method is a special case of one-step approximation, namely distribution fitting, with Theorem 2.1 used to satisfy (2.44). Convergence of the method was also proven in [35], by showing on-line VB to be a special case of stochastic approximation, which is known to converge [43].

Off-line VB approximation (Section 2.1) is a functional optimization of the KL distance. This functional optimization can be extended to the on-line scenario (2.6) as follows:

$$\check{f}(\theta|D_t) = f(\mathbf{d}_t|\theta, D_{t-1}) \check{f}(\theta|D_{t-1}). \quad (2.45)$$

We seek an optimal approximation of the true posterior under the conditional independence constraint (assume $q = 2$ for algebraic simplicity):

$$\check{f}(\theta|D_t) = \check{f}(\theta_1|D_t) \check{f}(\theta_2|D_t), \quad (2.46)$$

$$\check{f}(\theta|D_{t-1}) = \check{f}(\theta_1|D_{t-1}) \check{f}(\theta_2|D_{t-1}). \quad (2.47)$$

Then, using (2.45) and (2.46) in Theorem 2.1, the VB-optimal form of (2.46) is found in the following form:

$$\begin{aligned} \tilde{f}(\theta_i|D_t) &\propto \exp\left(\mathbb{E}_{\tilde{f}(\theta_i|D_t)}(\ln f(\mathbf{d}_t|\Theta, D_{t-1})) + \ln \check{f}(\theta_i|D_{t-1})\right), \\ &\propto \exp\left(\mathbb{E}_{\tilde{f}(\theta_i|D_t)}(\ln f(\mathbf{d}_t|\Theta, D_{t-1}))\right) \check{f}(\theta_i|D_{t-1}). \end{aligned} \quad (2.48)$$

Equation (2.48) can be rewritten as:

$$\tilde{f}(\theta_i|D_t) = f_{VB}(\mathbf{d}_t|\Theta, D_{t-1}) \check{f}(\theta_i|D_{t-1}), \quad i = 1, 2, \quad (2.49)$$

$$f_{VB}(\mathbf{d}_t|\Theta, D_{t-1}) \propto \exp\left(\mathbb{E}_{\tilde{f}(\theta_i|D_t)}(\ln f(\mathbf{d}_t|\Theta, D_{t-1}))\right). \quad (2.50)$$

Then, (2.49) is the VB-approximate update of parameter distribution, where $f_{VB}(\mathbf{d}_t|\Theta, D_{t-1})$ plays the role of VB-approximate observation model. Hence, the choice of $\check{f}(\theta_i|\cdot)$ conjugate with the VB observation model (2.50) yields a numerically tractable recursive identification algorithm. This *VB-conjugate* distribution can be found if the VB observation model (2.50) is from the exponential family.

Note that (2.50) is, in fact, in the form of the Bayes-closed approximation (2.43) with $\mathbb{E}_{\tilde{f}(\theta_i|D_t)}(\cdot)$ playing the role of linear operator $l_i(\cdot)$. However, the expected value, $\mathbb{E}_{\tilde{f}(\theta_i|D_t)}(\cdot)$, is conditioned by D_t and is, therefore, time-variant. This is not allowed for the linear operators used in the Bayes-closed approximation. Therefore, the on-line VB approximation (2.49) is not closed under Bayes' rule.

Chapter 3.

Linear Models: Classes and Their Inference

Classification of an observation model into linear and non-linear classes is of primary importance. Non-linear models are intrinsically more flexible but imply bigger intellectual challenges for their identification. Linear models—which have been studied for a very long time—have many attractions, including (i) analytical tractability, and (ii) computationally efficient evaluation (resulting from (i)). The main drawback of linear models are their modelling limitations. Various non-linear methodologies were proposed to match real-life problems, but these are often analytical intractable, and computationally expensive. Thus, linear models continue to be used, almost exclusively, in areas like real-time data processing or processing of high-dimensional data, where computational tractability is essential.

In this Chapter, we review the published Bayesian solution to special cases of the linear model. We focus on some special cases for which an efficient parameter inference procedure is available, namely, the AutoRegressive (AR) model (Section 3.2) and Principal Component Analysis (PCA) (Section 3.3). The use of the PCA model in the area of medical image processing is described in Section 3.4. In Section 3.5, we introduce possible extensions of these models and formulate the principal challenges of the thesis.

3.1. Bayesian Methods for Linear Models

We define a linear model as one to satisfying two conditions. First, it is assumed that the observed data are additively decomposed into an underlying signal and additive noise

$$D = M + E, \tag{3.1}$$

where $D \in \mathfrak{R}^{p \times n}$ are the observed data, $M \in \mathfrak{R}^{p \times n}$ is the signal, and $E \in \mathfrak{R}^{p \times n}$ is the noise. Second, the signal, M , is a linear combination,

$$M = AX, \tag{3.2}$$

of underlying parameters $A \in \mathfrak{R}^{p \times r}$ and $X \in \mathfrak{R}^{r \times n}$. Naming conventions for A and X differ in different application contexts. For the purpose of this thesis, we call A the matrix parameter and X the regressor. The rich linear model class described above, (3.1) and (3.2), has been studied with many different restrictive assumptions, yielding other rich classes constituting distinct research directions. We will review the most important models in this Chapter.

We start the classification of (3.1) and (3.2) by consideration of the dimensionality of the matrices A and X . Note that the role of p and n is interchangeable, as transposition of (3.1) yields the same model (with roles of A and X swapped). However, for the sake of clarity, we adopt a common convention, that the observed data matrix D is composed of n observations of the p -dimensional variable, \mathbf{d}_i , $i = 1, \dots, n$, i.e. $D = [\mathbf{d}_1, \dots, \mathbf{d}_n]$. We recognize two basic cases of model parameter inference based on the nature of the observation process:

off-line: the data has been acquired, and n data records have been saved. The task is to infer unknown A and X given all data, D , at once.

on-line: data are acquired incrementally with n growing possibly up to infinity. The task is to infer unknown A and X given all available data at given time t , as discussed in Section 2.1.2. It is useful to rewrite the model (3.1), (3.2) in terms of the time-indexed observation, \mathbf{d}_t :

$$\mathbf{d}_t = A\mathbf{x}_t + \mathbf{e}_t, \quad (3.3)$$

The accumulated data available at time t will be denoted $D_t = [\mathbf{d}_1, \dots, \mathbf{d}_t]$. Hence, t replaces n in (3.1), (3.2).

Further classification of linear models is related to *a priori* knowledge available about A and X . We recognize two families:

Regression models: either A or X is assumed to be known. The task is to infer the other. In this scenario, dimension r typically satisfies $r > p$.

Signal separation models: where both A and X are assumed to be unknown. The task is to infer both of them. In this scenario, the dimension r typically satisfies $r \leq p$.

Further sub-classed can be defined with respect to an assumed distribution for the noise, $f(E)$. In this thesis, we focus our attention to models with Normal distributed noise. The most general case of the Normal distribution of E can be written as

$$f(\text{vec}(E)) = \mathcal{N}(\text{vec}(\mu_E), \Sigma_E),$$

with: mean value, $\mu_E \in \mathbb{R}^{p \times n}$, transformed into a vector, $\text{vec}(\mu_E) \in \mathbb{R}^{pn \times 1}$, and symmetric, positive definite covariance matrix $\Sigma_E \in \mathbb{R}^{pn \times pn}$. Note that Σ_E is typically a large matrix, with $\frac{1}{2}(pn + 1)pn$ distinct elements. It is therefore much larger than the number pn of available data D . Therefore, a restricted covariance structure must be considered. One such restriction comes from confining $f(E)$ to the following matrix Normal distribution (Appendix A.1):

$$f(E) = \mathcal{N}(\mu_E, \Sigma_p \otimes \Sigma_n), \quad (3.4)$$

where $\mu_E \in \mathbb{R}^{p \times n}$, while $\Sigma_p \in \mathbb{R}^{p \times p}$ and $\Sigma_n \in \mathbb{R}^{n \times n}$ are symmetric, positive definite matrices. The covariance matrix $\Sigma_E = \Sigma_p \otimes \Sigma_n$ has now $\frac{1}{2}(p + 1)p + \frac{1}{2}(n + 1)n$ distinct elements, which is, once again, more than the number pn of available data D . A typical further restriction in regression models

is independent identically distributed (i.i.d.) assumption defined via a time-invariant distribution for the noise vector, e_t :

$$f(e_t | \mu_e, \Sigma_e) = \mathcal{N}(\mu_e, \Sigma_e), \quad (3.5)$$

with $\mu_e \in \mathbb{R}^{p \times 1}$, $\Sigma_e \in \mathbb{R}^{p \times p}$ constant for all $t = 1, \dots$, and e_t, e_s independent for $t \neq s$. This can be written in the matrix Normal distribution form as:

$$f(E | \mu_e, \Sigma_e) = \prod_{t=1}^n f(e_t | \mu_e, \Sigma_e) = \mathcal{N}(\mu_e \mathbf{1}_{1,n}, \Sigma_e \otimes I_n).$$

Thusfar, we have defined the Normal distribution in terms of its covariance matrix Σ . However, for some problems, it is more convenient to work with the precision matrix, Ω , instead of the covariance matrix, in which case, we denote $\Sigma = \Omega^{-1}$.

The above mentioned classes and scenarios can be mutually combined to yield a wide class of identification problems. Further restriction and assumptions often constitute research directions (e.g. Factor Analysis, General Linear Models, etc.). The most important special cases of the linear model, (3.1) and (3.2), are listed in Table 3.1. The name of each model often comes from the associated inference technique (e.g. PCA or FA) since the inference technique was developed before its associated assumptions were recognized as constituting a special case of the linear model. In these cases, by convention, we will use the name of the method as part of the name of the model, such as *Factor Analysis (FA) Model*. We list references to both point-based inference (ML or MAP) and Bayesian inference methods in Table 3.1.

General Linear Models (GLM) are traditional statistical models for modelling time-series and for forecasting [17, 54]. Typically, the on-line scenario is considered. Assumptions underlying the model often vary, the most common being a known matrix of parameters A , unknown x_t , and Gaussian distribution of noise

$$f(E) = \mathcal{N}(\mathbf{0}_{p,n}, \omega^{-1} I_p \otimes I_n),$$

with scalar precision parameter $\omega > 0$. A full Bayesian solution is available in [17].

Various extensions of the model has been studied, such as Dynamic Generalized Linear Models [54]. The extensions impose extra parameterization that should be known *a priori*. The model is then identified using Kalman-filter theory (which can be interpreted as approximate Bayesian identification [55]) or using an MCMC approach [56]. These models are traditionally applied in analysis of econometric data, where computational cost is not critical. However, recently, it was successfully used in real-time processing [57].

Autoregressive Models (AR) can be considered as a special case of GLMs. However, they were developed independently, and for different application contexts, such as control theory [16, 58]. Typical assumptions are that the parameter A is unknown, and regression vector x_t is known, being a function of previously observed data. Noise is considered to be Normal,

$$f(E) = \mathcal{N}(\mathbf{0}_{p,t}, \Omega^{-1} \otimes I_t),$$

model name	A	X	E	ML/MAP	full Bayesian
General Linear Model (GLM)	K	U	$\mathcal{N}(\mathbf{0}_{p,n}, \omega^{-1} I_p \otimes I_n)$	Analytical [17, 44]	Analytical [17, 44]
AutoRegressive Model (AR)	U	K	$\mathcal{N}(\mathbf{0}_{p,n}, \Omega^{-1} \otimes I_n)$, Ω positive definite	Analytical [45]	Analytical [16]
Generalized AR Model (GAR)	U	K	$f(E) = \prod_{t=1}^n f(e_t)$, $f(e_t) = \sum_{i=1}^c \alpha_i \mathcal{N}(\mu, \omega^{-1} I_p)$ with unknown weights $\alpha_i, i = 1 \dots c$		VB [46]
(Probabilistic) Principal Component Analysis Model (PPCA)	U	U	$\mathcal{N}(\boldsymbol{\mu} \mathbf{1}_{1,n}, \omega^{-1} I_p \otimes I_n)$	Full [47], EM[48]	VB [7]
Factor Analysis Model (FA)	U*	U	$\mathcal{N}(\boldsymbol{\mu} \mathbf{1}_{1,n}, \Sigma \otimes I_n)$, Σ diagonal	Iterative [47], EM [49]	Partial [50], VB [4]
Independent Component Analysis Model (ICA)	U	U	$E = \mathbf{0}_{p,n}$	EM[51]	VB[33]
Independent Factor Analysis Model	U*	U	$\mathcal{N}(\boldsymbol{\mu} \mathbf{1}_{1,n}, \Sigma \otimes I_n)$, Σ diagonal	EM [52]	VB[53]
Legend:					
U - unknown, i.e. to be inferred.					
K - known <i>a priori</i> .					
* methods differ in prior distributions imposed on A .					

Figure 3.1.: Table of established linear models

with symmetric, positive definite precision matrix $\Omega \in \mathbb{R}^{p \times p}$. Recursive Bayesian identification is available in [16], using the principle of conjugacy. The model can be extended while preserving conjugacy to cases of a known transformation of the system output, and non-stationary parameters, as will be discussed in Section 3.2.

A popular extension of the AR model is via state-space modelling. In the state-space model, \mathbf{x}_t is considered unknown but modelled by another linear model [59]. Hence, we consider the state-space model to be bi-linear. In general, full analytical recursive Bayesian identification of parameters of the state-space model cannot be achieved. An approximate Bayesian inference of the parameters of the state-space model, using the VB approximation, was presented in [60].

Generalized AR Models (GAR) are extensions of the AR model to allow a mixture type noise [46],

$$f(e_t | \Omega, \boldsymbol{\mu}_e, \boldsymbol{\alpha}) = \sum_{i=1}^c \alpha_i \mathcal{N}(\boldsymbol{\mu}_i, \Omega_i^{-1}),$$

with $\Omega = \{\Omega_1, \dots, \Omega_c\}$, $\boldsymbol{\mu}_e = [\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_c]$, and $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_c]'$ being the mixture weights. An approximate Bayesian identification was presented in [46], using the VB approximation. However, recursive identification was not achieved.

AutoRegressive Moving Average Models (ARMA) are AR models with correlated noise. The noise distribution is then

$$f(E | \Omega, C) = \mathcal{N}(\mathbf{0}_{p,n}, \Omega^{-1} \otimes [\Phi(C)]^{-1}),$$

with parameterization extended via a symmetric positive definite matrix, $\Phi(C) \in \mathbb{R}^{m \times n}$, where $C \in \mathbb{R}^{m \times 1}$ is formed by the coefficients of the order- m Moving Average (MA) part of the model [61]. Full Bayesian identification is achieved if C is known [61]. Bayesian identification for unknown C was addressed in [62], using mixture-based methods and Quasi-Bayes (QB) approximation [25].

Probabilistic PCA (PPCA) is a probabilistic formulation of the Principal Component Analysis (PCA) method. The model assumes a Normal distribution of noise:

$$f(E) = \mathcal{N}(\mathbf{0}_{p,n}, \omega^{-1} I_p \otimes I_n).$$

Its MAP inference was published in [63], and extended to full Bayesian inference in [7] using a VB approximation.

Factor Analysis (FA) is a classical model for addressing the signal separation problem [47]. Both A and X are assumed to be unknown, and $r < \min(p, n)$ is typically assumed to be known. There may be various restrictions on A or X . The noise is normally distributed,

$$f(E) = \mathcal{N}(\mathbf{0}_{p,n}, \Omega^{-1} \otimes I_n), \quad (3.6)$$

with positive definite matrix Ω assumed *diagonal*. The inference is traditionally done via the ML approach [47]. The Bayesian solution was studied in [50] and evaluated using an MCMC approximation [64]. The Variational Bayes approximation was published in [4].

Independent Component Analysis (ICA) is a new, popular model for addressing the signal separation problem. The noise is typically considered to be zero, hence D is fully modelled by AX with $r = p$. The signal-noise separation is achieved with respect to columns of A and rows of X respectively. Separation of A and X is achieved by imposing priors (typically non-Normal) on columns a_i of parameter A . An iterative MAP estimation was published in [52]. Bayesian identification was considered in [33] using a VB approximation.

Independent Factor Analysis is an extension of the ICA idea to include the noise model of Factor Analysis (3.6). The product AX is modelled in the same way as for ICA. An iterative MAP estimation was published in [53].

The above list of methods is by no means exhaustive. Further extensions of model assumptions can be (and indeed are being) made to extend modelling capability of the basic linear model. The resulting inference schemes naturally involve more parameters, thus increasing the number of samples which must be generated by MCMC, or the number of iterations in prospective EM or VEM algorithms.

In this thesis, our concern is with computational tractability of the VB distributional approximation. For better understanding of the problem, we start with simple models such as (i) the AutoRegressive (AR) model, and (ii) Principal Component Analysis (PCA). Both models enjoy, under certain modelling restrictions, analytically tractable inference. Relaxation of these restrictions leads to a loss of analytical tractability, which has to be restored via further approximations. Successful application of VB approximation has been reported for (i) mixture-based extension of an AR process [46], and (ii) Bayesian identification of the PPCA model with unknown rank r [7]. Both methods use the VEM algorithm (Algorithm 2.2) for evaluation of the parameter inference. We seek a simplification of the inference method (or re-parameterization of the model), yielding results comparable to these VEM-based solutions but at significantly lower computational cost.

3.2. The Multivariate AutoRegressive (AR) Model

Linear AR processes are widely applied in filtering [65], speech analysis [66], spectrum analysis [67], control [68], etc. The main advantage of the model is analytical tractability which results in computationally efficient and stable estimation algorithms. However, its underlying assumptions (i.e. linear combination of measured values, and Gaussian distribution for the residue) are rarely met in practice. Physical models, typically requiring complex non-linear modelling, may be used to fit the observed data. Attempts to extend the AR model itself have also been made [46, 69]. However, these solutions are computationally expensive and thus unsuitable for processing of large amounts of data or for on-line (real-time) evaluation. Typically, therefore, AR models continue to be used even in these cases.

In this Section, we study an extension to the AR model that preserves its analytical tractability, allowing fast on-line estimation of the model. Of particular concern is the study of numerically tractable

recursive algorithms. Recursive estimation algorithms are widely used for on-line control applications [58], and for adaptive filtering [70]. Computational simplicity is a key requirement for real-time adaptive estimation. In off-line cases, the emphasis on computational issues and recursive methods can also pay off, for example in the off-line processing of massive datasets [31].

3.2.1. Bayesian Inference of the AR model

As outlined in Section 3.2, the AR model is a special case of the recursive linear model (3.3) with the noise distributed as

$$f(e_t|\Omega) = \mathcal{N}(\mathbf{0}_p, \Omega^{-1}). \quad (3.7)$$

The noise vectors e_t, e_τ are independent for $t \neq \tau$. $\Omega \in \mathbb{R}^{p \times p}$ is an unknown positive definite matrix. Parameters A, Ω are considered time-invariant, and so the observation process (3.3) is stationary. This assumption will be relaxed in Section 3.2.3.

Regressor \mathbf{x}_t is assumed to be known, i.e. it may contain any observed variables or their *known* transformations. Formally,

$$\mathbf{x}_t = \mathbf{g}(D_{t-1}, W_t), \quad (3.8)$$

where auxiliary variable, W_t , may contain any known variables, such as a measured external (exogenous) signal, time variable, t , for time-variant systems, etc.

The model described above is rather general. For better intuition we list a few special cases:

1. Univariate autoregressive (AR) model:

$$d_t = \sum_{k=1}^r a_k d_{t-k} + e_t, \quad (3.9)$$

where $\mathbf{x}_t = \mathbf{g}(D_{t-1}) = [d_{t-1}, \dots, d_{t-r}]$. This is illustrated on Figure 3.2 (left) in standard signal flow graph form.

2. The ARX model, i.e. AR with exogenous observed input w_t . In this case, the model becomes

$$d_t = \sum_{k=1}^m a_k d_{t-k} + \sum_{k=m+1}^r a_k w_{t-k+m+1} + e_t. \quad (3.10)$$

The external input, w_t , can be seen as a time-variant auxiliary variable, W_t , and the transformation set, $\mathbf{g} = [g_1, \dots, g_r]'$, is defined as:

$$x_{i;t} = g_i(D_{t-1}, W_t) = \begin{cases} d_{t-i} & i = 1, \dots, m \\ w_{t-i+m+1} & i = m+1, \dots, r \end{cases}.$$

Remark 3.1 (Order of an AR model) *The choice of $\mathbf{g}(D_t)$ does not imply that the regressor must contain all historical values. Such a formulation would not be tractable. Typically, only a finite length*

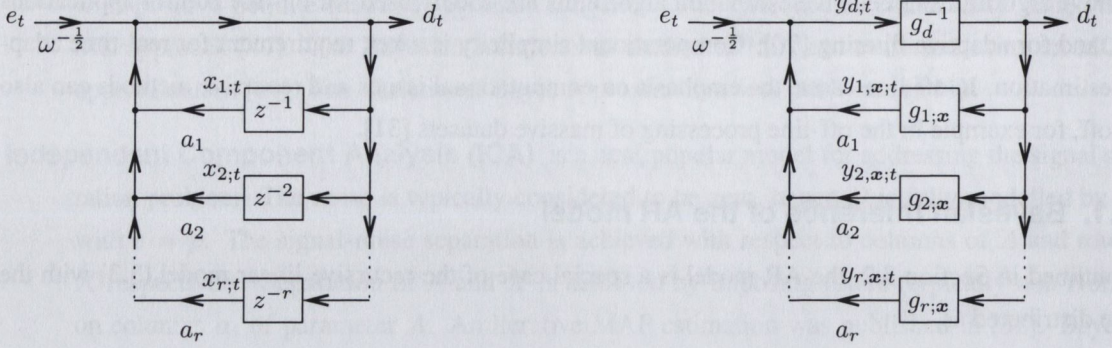


Figure 3.2.: Signal flow graphs of the AR (left) and EAR (right) models.

of historical data is used: $\mathbf{x}_t = \mathbf{g}(\mathbf{d}_{t-1}, \mathbf{d}_{t-2}, \dots, \mathbf{d}_{t-q})$. The maximum delay, q , of an observation used in the regression will be called the order. It follows that (3.3) is valid only for $t > q$. Specifically, for the univariate AR model (3.9) $q = r$.

Remark 3.2 (Classical approach) In the classical literature, a univariate case is usually considered, i.e. $p = 1$. The classical solution to this problem is based on the prediction-error criterion. The predictor is a Wiener filter with unknown coefficients. Parameter estimates are obtained by solution of the normal equations. Two principal approaches to its solution are the covariance and correlation methods respectively [45]. There are many techniques for the numerical solution of these equations, including recursive ones, such as the Recursive Least Squares (RLS) algorithm [58].

The problem of Bayesian inference is to find posterior distributions of the unknown, real parameters, Ω and A , of this model. Combining (3.3), and (3.7) we obtain the conditional distribution of observations, \mathbf{d}_t :

$$f(\mathbf{d}_t | A, \Omega, \mathbf{x}_t) = \mathcal{N}(A\mathbf{x}_t, \Omega^{-1}). \quad (3.11)$$

Inference of the unknown parameters Ω and A follows from Bayes' rule:

$$f(A, \Omega | D_t, X_t) \propto f(\mathbf{d}_t | A, \Omega, X_{t-1}) f(A, \Omega | D_{t-1}, X_t). \quad (3.12)$$

The model (3.11) belongs to the exponential family, and so both a conjugate prior and sufficient statistics are available (Section 2.1.2). The conjugate distribution of parameters for (3.11) is of the Normal-Wishart (\mathcal{NW}) type [10]:

$$\mathcal{NW}_{A,\Omega}(V, \nu) \equiv \frac{|\Omega|^{\frac{1}{2}\nu}}{\zeta_{\mathcal{NW}}(V, \nu)} \exp \left\{ -\frac{1}{2} \Omega [-I_p, A] V [-I_p, A]' \right\}, \quad (3.13)$$

$$\zeta_{\mathcal{NW}}(V, \nu) = \Gamma_p \left(\frac{1}{2} (\nu - r + p + 1) \right) |\Lambda|^{-\frac{1}{2}(\nu - r + p + 1)} \times |V_{aa}|^{-0.5p} 2^{0.5p(\nu + p + 1)} \pi^{\frac{r}{2}}, \quad (3.14)$$

$$V = \begin{bmatrix} V_{dd} & V'_{ad} \\ V_{ad} & V_{aa} \end{bmatrix}, \quad \Lambda = V_{dd} - V'_{ad} V_{aa}^{-1} V_{ad}, \quad (3.15)$$

with V_{dd} being the $p \times p$ upper-left sub-block of matrix V . V, ν are the sufficient statistics of $\mathcal{NW}_{A,\Omega}(\cdot)$. $\Gamma_p(\frac{1}{2}\nu)$ is the Multi-Gamma function

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

with $\Gamma(\cdot)$ being the Gamma function [71].

In this Section, we are concerned with on-line identification, i.e. we evaluate distribution (3.13) at each time $t > q$. We will distinguish statistics at each time moment by subscript t , i.e. V_t, ν_t . If the variable already has a subscript, this time index will be separated from it by a semi-colon, e.g. $V_{dd;t}$.

The statistics of the conjugate prior distribution, V_0, ν_0 , are chosen to reflect our initial knowledge of parameters. If we do not have any preference, we use a very flat (non-committal) \mathcal{NW} distribution. Typically $V_0 = \varepsilon I_{p+r}$, where ε is a small positive scalar. We choose $\nu \geq r - 2$ in order that the normalizing constant (3.14) be finite.

Substituting (3.11) into (3.12) and invoking (3.13) at time $t - 1$, then the posterior distribution at time $t > q$ is

$$f(A, \Omega | D_t, X_t) = \mathcal{NW}_{A,\Omega}(V_t, \nu_t), \quad (3.16)$$

$$V_t = V_{t-1} + \begin{bmatrix} \mathbf{d}_t \\ \mathbf{x}_t \end{bmatrix} \begin{bmatrix} \mathbf{d}'_t, \mathbf{x}'_t \end{bmatrix} = V_{t-1} + \mathbf{y}_t \mathbf{y}'_t = V_0 + \sum_{i=q+1}^t \mathbf{y}_i \mathbf{y}'_i, \quad (3.17)$$

$$\nu_t = \nu_{t-1} + 1 = \nu_0 + (t - q). \quad (3.18)$$

Here,

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{d}_t \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \mathbf{d}_t \\ \mathbf{g}(D_{t-1}, W_t) \end{bmatrix}, \quad (3.19)$$

is the extended regression vector. The outer product, $\mathbf{y}_t \mathbf{y}'_t$ will be called a *dyad* in this thesis. The history of the extended regressor will be denoted by $Y_t = [\mathbf{y}_1, \dots, \mathbf{y}_t]$. Since the recursion begins at $t = q + 1$, V_q and ν_q are chosen to be $V_q = V_0$ and $\nu_q = \nu_0$. This is equivalent to choosing the distribution on parameters to be stationary for $0 \leq t \leq q$, of the form given by prior. Finally, from (3.18), note ν_t acts as a counter of incoming data samples.

Remark 3.3 (Moments of the distribution) *Note, from Appendix A.2, that the mean values of posterior distribution (3.13) are*

$$\hat{A}_t = V'_{ad;t} V_{aa;t}^{-1}, \quad (3.20)$$

$$\hat{\Omega}_t = \frac{1}{\nu_t - r + p + 1} \Lambda_t^{-1}. \quad (3.21)$$

3.2.1.1. Computational Issues

The Bayesian posterior estimates presented above are closely related to approaches available in the signal processing literature. Key properties are now summarized:

1. (3.20)–(3.21)—evaluated via recursions (3.17, 3.18)—are algorithmically identical to the covariance method [45], and are valid, $\forall t > q$, as derived.
2. V_t (3.15), (3.17) is asymptotically Toeplitz, and so (3.20) becomes algorithmically the same as the correlation method iff $t \rightarrow \infty$. Only then do the computational benefits of Toeplitzity accrue to the Bayesian approach, namely ease of updating and storage of r rather than $r(r+1)/2$ numbers. While this simplification is popular in real-time applications such as audio processing [72], it is unreliable for t small and/or for non-stationary data.
3. A numerically efficient solution to (3.17), (3.20) is based on the LD decomposition [73], i.e. $V_t = L_t T_t L_t'$, where L_t is lower triangular and T_t is diagonal. The update of the sufficient statistics (3.17) is replaced by recursions on L_t and T_t [74]. This approach is superior to accumulation of the full matrix V_t for the following reasons:
 - a) *Compactness*: all operations are performed on triangular matrices, i.e. $(r+p)(r+p)/2$ values, compared to $(r+p)^2$ for full V_t .
 - b) *Computational Efficiency*: the estimation update requires $O((r+p)^2)$ operations in each step to re-evaluate L_t, T_t , followed by evaluation of the normalizing coefficient (3.14) with complexity $O(r+p)$ and finally evaluation of (3.20) with complexity $O((r+p)^2)$. In contrast, operations (3.14) and (3.20) are of $O((r+p)^3)$ for full matrix V_t . Implementation of the update with full matrix V_t using the matrix inversion lemma [16, 58] is of the same complexity as using the LD decomposition.
 - c) *Regularity*: elements of T_t are certain to be positive, which guarantees positive-definiteness of V_t . This property is unique to the LD decomposition.

3.2.1.2. Prediction

One of the main benefits of AR modelling of time series is its appropriateness for prediction of future observations. The one-step-ahead predictive distribution is given by the ratio of normalizing coefficients (3.14), a result established in general for the exponential family in [10]. For the AR model:

$$f(\mathbf{d}_{t+1} | Y_t, \mathbf{x}_{t+1}) = (2\pi)^{-\frac{1}{2}} \frac{\zeta_{\mathcal{N}\mathcal{W}}(V_t + \mathbf{y}_{t+1} \mathbf{y}'_{t+1}, \nu_t + 1)}{\zeta_{\mathcal{N}\mathcal{W}}(V_t, \nu_t)}, \quad t \geq q. \quad (3.22)$$

This is the Student- t distribution with $\nu_t - r + p + 1$ degrees of freedom [16]. The mean value of this distribution is readily found to be

$$\mathbf{E}(\mathbf{d}_{t+1} | Y_t) = \hat{A}_t \mathbf{x}_{t+1} = \hat{\mathbf{d}}_{t+1}, \quad (3.23)$$

which is equal to the intuitively appealing result from classical theory [45].

3.2.1.3. Model Structure Determination

The structure of the regression model is determined by choice of the set of transformation functions $g(D_{t-1}, W_t)$. This is the choice of the model designer. The problem of choice of the appropriate

model is called model structure determination problem, treated in detail in [75]. We can assemble a finite set of possibilities $G = \{\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_c\}$, where \mathbf{g}_i denotes the i th possible choice of regressor structure. It is then necessary to calculate the *a posteriori* probabilities of all cases in G . Using Bayes' rule:

$$f(\mathbf{g}_i | Y_t) \propto f(Y_{t \setminus q(i)} | Y_{q(i)}, \mathbf{g}_i) f(\mathbf{g}_i), \quad i = 1, \dots, c, \quad (3.24)$$

where $Y_{t \setminus q(i)} = [\mathbf{y}_{q(i)+1}, \dots, \mathbf{y}_t]$, and $q(i)$ is the longest data memory with respect to D_t across all functions \mathbf{g}_i .

From (3.22):

$$\begin{aligned} f(Y_{t \setminus q(i)} | Y_{q(i)}, \mathbf{g}_i) &= (2\pi)^{-\frac{t-q(i)}{2}} \prod_{j=q(i)+1}^t \frac{\zeta_{\mathcal{N}\mathcal{W}}(V_j(\mathbf{g}_i), \nu_j)}{\zeta_{\mathcal{N}\mathcal{W}}(V_{j-1}(\mathbf{g}_i), \nu_{j-1})} \\ &= (2\pi)^{-\frac{t-q(i)}{2}} \frac{\zeta_{\mathcal{N}\mathcal{W}}(V_t(\mathbf{g}_i), \nu_t)}{\zeta_{\mathcal{N}\mathcal{W}}(V_0(\mathbf{g}_i), \nu_0)}, \quad i = 1 \dots c. \end{aligned} \quad (3.25)$$

Here, we use notation $V_t(\mathbf{g}_i)$ to emphasize the fact that the statistics V_t (3.17) are accumulated differently for each choice of the structure \mathbf{g}_i in G . In situations where it is clear which structure was used to obtain the statistics, we will use the simplified notation V_t .

Note that (3.24), (3.25) engenders Ockham's Razor since the involved determinant term in (3.14) penalizes candidates of greater complexity [13]. (3.24) provides a posterior inference for unknown $\mathbf{g} \in G$, going the way towards relaxing the former restriction on regression models that the transformations be known.

3.2.2. The Extended AutoRegressive (EAR) Model

In this section, we review the widest class of models for which the algorithms in Section 3.2 remain valid. The favourable algorithmic properties for the AR model are based on the elegant recursive form (3.17), (3.18) of the $\mathcal{N}\mathcal{W}$ sufficient statistics (3.13), and so this feature must be conserved under any extension. We note that the posterior distribution remains $\mathcal{N}\mathcal{W}$ if the extended regressor, \mathbf{y}_t , is constructed from

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{y}_{d,t} \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \mathbf{g}_d(D_t, W_t) \\ \mathbf{g}_x(D_{t-1}, W_t) \end{bmatrix} = \mathbf{g}(D_t, W_t), \quad (3.26)$$

as compared to (3.19). Here, $\mathbf{y}_{d,t}$ denotes transformed data, \mathbf{d}_t , corresponding to the model

$$\mathbf{y}_{d,t} = A\mathbf{x}_t + \Omega^{-\frac{1}{2}}\mathbf{e}_t, \quad (3.27)$$

$$\mathbf{d}_t = \mathbf{g}_d^{-1}(\mathbf{y}_t, D_{t-1}, W_t). \quad (3.28)$$

This model structure is illustrated in Figure 3.2 (right), see page 26. The distribution of observations is now obtained by transformation of (3.11):

$$f(\mathbf{d}_t | A, \Omega, Y_{t-1}) = |J_t(\mathbf{d}_t)| \mathcal{N}(-A\mathbf{y}_t, \Omega^{-1}), \quad (3.29)$$

where J_t is the Jacobian of transformation $\mathbf{g}_d(\cdot)$ (3.26); i.e. $J_t(\mathbf{d}_t) = \frac{\partial \mathbf{g}_d}{\partial \mathbf{d}_t} \in \mathbb{R}^{p \times p}$. This creates an additional restriction that $\mathbf{g}_d(\cdot)$ be a differentiable one-to-one (bijective) mapping for each setting of W_t . Moreover, $\mathbf{g}_d(\cdot)$ (3.27) must explicitly be a function of \mathbf{d}_t in order that $J_t \neq 0$. This ensures the necessary uncertainty propagation from e_t to \mathbf{d}_t (Figure 3.2 (right)). In conclusion, Bayesian estimation with this model is, by design, of the same form as for the AR model (3.13)-(3.18).

The EAR model class (3.29) includes the following important cases [16]:

1. An AR process with bijective known non-linear transformation of observations: $d_t = \tau(y_t)$. The transformation \mathbf{g} is then defined as the inverse of this non-linearity: $\mathbf{g}_d(\cdot) = \tau^{-1}(\cdot)$, and $\mathbf{g}_x(\cdot) = \tau^{-1}(\cdot)$.
2. The ARMA model with a *known* MA part, i.e. an AR model driven by coloured noise of known covariance matrix. Transformation \mathbf{g} is then the necessary pre-whitening filter on the colored innovations. This process has a numerically efficient recursive identification [61].
3. The incremental AR process with the regression defined on increments of the measurement process.

Both prediction and model structure identification must be adjusted for the observation model (3.29). The marginal predictive distribution becomes, from (3.22),

$$f(\mathbf{d}_{t+1} | Y_t, \mathbf{x}_{t+1}) = |J_{t+1}(\mathbf{d}_{t+1})| (2\pi)^{-\frac{1}{2}} \frac{\zeta_{\mathcal{N}\mathcal{W}}(V_t + \mathbf{y}_{t+1} \mathbf{y}'_{t+1}, \nu_t + 1)}{\zeta_{\mathcal{N}\mathcal{W}}(V_t, \nu_t)}, \quad (3.30)$$

and model structure identification is adjusted from (3.25) using (3.30), as follows:

$$\begin{aligned} f(D_{t \setminus q(i)} | D_{q(i)}, \mathbf{g}_i) &= (2\pi)^{-\frac{t-q(i)}{2}} \prod_{j=q(i)+1}^t |J_j(\mathbf{d}_j)| \frac{\zeta_{\mathcal{N}\mathcal{W}}(V_j(\mathbf{g}_i), \nu_j)}{\zeta_{\mathcal{N}\mathcal{W}}(V_{j-1}(\mathbf{g}_i), \nu_{j-1})}, \\ &= (2\pi)^{-\frac{t-q(i)}{2}} \frac{\zeta_{\mathcal{N}\mathcal{W}}(V_t(\mathbf{g}_i), \nu_t)}{\zeta_{\mathcal{N}\mathcal{W}}(V_0(\mathbf{g}_i), \nu_0)} \prod_{j=q(i)+1}^t |J_j(\mathbf{d}_j)|, \quad i = 1, \dots, c. \end{aligned} \quad (3.31)$$

Remark 3.4 (Linear transformations) Note that iff Jacobian J_t (3.29) is independent of \mathbf{d}_t (i.e. iff $\mathbf{g}_d(\cdot)$ (3.27) is a linear transformation), then (3.30) implies that the expected prediction is

$$\hat{\mathbf{d}}_{t+1} = \mathbf{g}_d^{-1} \left(-\hat{A}_t \mathbf{x}_{t+1} \right), \quad (3.32)$$

in analogy with (3.23). Moreover, if the transformation $\mathbf{g}_d(\cdot)$ constitutes a simple scaling (i.e. $\mathbf{y}_{d,t} = \alpha \mathbf{d}_t$, so that $J_t = \alpha$), this is further simplified to:

$$\hat{\mathbf{d}}_{t+1} = -\hat{A}_t \mathbf{x}_{t+1} \alpha^{-1}. \quad (3.33)$$

3.2.3. Modelling Time-Varying Parameters of Non-Stationary AR Models Using Forgetting

The assumption of constant parameter values is rarely met in practice. In many applications, however, a complete model of parameter variations is not known. The problem is then under-determined, obviating the full Bayesian solution and leading to many heuristic techniques. The standard batch (off-line) algorithm uses windowing [76]. Alternatively, the concept of forgetting [77] is used in adaptive signal processing [78] and recursive estimation [58].

3.2.3.1. Explicit Modelling of Parameter Evolution

Non-stationarity of the parameters is handled by modelling $\theta_t = [A_t, \Omega_t]$ as a new random variable for each time t . Then, the observation model (3.11), $f(\mathbf{d}_t | \theta_t, Y_{t-1}, \mathbf{x}_t)$, does not update the posterior distribution of parameters at time $t-1$, $f(\theta_{t-1} | Y_{t-1})$. This can be overcome by the explicit modelling of parameter evolution by a pdf $f(\theta_t | \theta_{t-1}, Y_{t-1})$. The joint distribution is then

$$f(\mathbf{d}_t, \theta_t | \theta_{t-1}, Y_{t-1}) = f(\mathbf{d}_t | \theta_t, Y_{t-1}) f(\theta_t | \theta_{t-1}, Y_{t-1}). \quad (3.34)$$

The update of parameter distributions via (3.34),

$$f(\theta_t, \theta_{t-1} | Y_t) \propto f(\mathbf{d}_t, \theta_t | \theta_{t-1}, Y_{t-1}) f(\theta_{t-1} | Y_{t-1}), \quad (3.35)$$

causes proliferation of random variables, in that a new random variable θ_t is introduced at each step. Hence, the parameter distributions at times t and $t-1$ have different functional forms, violating conjugacy. Therefore, computationally efficient on-line identification cannot be achieved.

θ_{t-1} can be eliminated from (3.35) by marginalization:

$$f(\theta_t | Y_t) \propto f(\mathbf{d}_t | \theta_t, Y_{t-1}, \mathbf{x}_t) f(\theta_t | Y_{t-1}), \quad (3.36)$$

$$f(\theta_t | Y_{t-1}) = \int_{\theta_{t-1}} f(\theta_t | \theta_{t-1}, Y_{t-1}) f(\theta_{t-1} | Y_{t-1}) d\theta_{t-1}. \quad (3.37)$$

The choice of the parameter evolution model $f(\theta_t | \theta_{t-1}, Y_{t-1})$ is discussed in [79]. Integration of (3.37) is feasible, for example, for a random-walk process

$$f(\theta_t | \theta_{t-1}, S) = \mathcal{N}(\theta_t - \theta_{t-1}, S), \quad (3.38)$$

where S is a covariance matrix of dimensions $p(2p+r) \times p(2p+r)$, chosen *a priori*. In many practical situations, we may not have any guide about how to choose S , and wrong choice may lead to poor performance of the identification.

3.2.3.2. Modelling of Time-varying Parameters via Forgetting

As an alternative approach, the technique of forgetting was suggested in [79]. There, the explicit model of parameter evolution (3.34), and the subsequent integration (3.37), are replaced via a probabilistic

operator:

$$f(\theta_t|Y_{t-1}, \phi_t) \propto [f(\theta_{t-1}|Y_{t-1})_{\theta_t}]^{\phi_t} \times \bar{f}_t(\theta_t|Y_{t-1})^{1-\phi_t}. \quad (3.39)$$

The notation $f(\cdot)_{\theta_t}$ indicates the replacement of the argument of $f(\cdot)$ by θ_t , where θ_t is the time-variant unknown parameter set at time t . $\bar{f}(\cdot)$ is a chosen alternative distribution, expressing auxiliary knowledge about θ_t at time t . Coefficient ϕ_t , $0 \leq \phi_t \leq 1$ is known as the forgetting factor. From (3.39), the limits are interpreted as follows:

for $\phi_t = 1$: prior information, at time t , about the new variable θ_t is identical to the posterior of θ_{t-1} at $t - 1$:

$$f(\theta_t|Y_{t-1}, \phi_t) = f(\theta_{t-1}|Y_{t-1})_{\theta_t}.$$

This is consistent with the choice $\theta_t = \theta_{t-1}$, i.e. the time-invariant parameter assumption.

for $\phi_t = 0$: prior information, at time t , about the new variable θ_t is chosen as the alternative distribution:

$$f(\theta_t|Y_{t-1}, \phi_t) = \bar{f}(\theta_t|Y_{t-1}).$$

This is consistent with the choice of independence between θ_t and θ_{t-1} , i.e.

$$f(\theta_t, \theta_{t-1}|Y_{t-1}) = f(\theta_t|Y_{t-1}) f(\theta_{t-1}|Y_{t-1}).$$

The forgetting factor is typically considered as fixed and it is chosen by the designer of the model. The choice of ϕ_t close to 1 models slowly varying parameters. The choice of ϕ_t close to 0 models rapidly varying parameters.

We require (3.39) to be conjugate to the observation model (3.11), i.e. belong to the \mathcal{NW} family (3.13). The \mathcal{NW} family is closed under the convex combining (i.e. geometric mean) in (3.39) yielding another member of the same family. Therefore, the \mathcal{NW} distribution with parameters $\bar{V}, \bar{\nu}$ is used as the alternative $\bar{f}(\cdot)$. It is typically chosen as a flat distribution, e.g. with the same parameter values as the prior: $\bar{V} = V_0, \bar{\nu} = \nu_0$. Substituting (3.39) and (3.11) into (3.12) yields the following recursive update of the \mathcal{NW} statistics:

$$f(A_t, \Omega_t|Y_t) = \mathcal{NW}_{A, \Omega}(V_t, \nu_t), \quad (3.40)$$

$$V_t = \phi_t V_{t-1} + \mathbf{y}_t \mathbf{y}_t' + (1 - \phi_t) \bar{V}, \quad (3.41)$$

$$\nu_t = \phi_t \nu_{t-1} + 1 + (1 - \phi_t) \bar{\nu}. \quad (3.42)$$

When $\phi_t = 1$, the update is identical to the stationary equations (3.17, 3.18).

For the case $V_0 = 0, \nu_0 = 0$, and $\phi_t = \phi$ constant, the method is known as *exponential forgetting* because (3.41) implies a sum of dyads weighted by a discrete exponential sequence,

$$V_t = \sum_{i=q+1}^t \phi^{t-i} \mathbf{y}_i \mathbf{y}_i' + \bar{V}, \quad (3.43)$$

$$\nu_t = \sum_{i=p+1}^t \phi^{t-i} + \bar{\nu}. \quad (3.44)$$

This interpretation is helpful, since it provides an intuitive choice for ϕ , as follows.

Remark 3.5 (Intuitive choice of forgetting factor) Here, we compare the exponential forgetting technique with the windowing approach. First, we identify a stationary AR model on the observation window of h samples. We assume that the prior was chosen as regular, i.e. $\nu_0 > q$. Then the degrees of freedom of the posterior distribution is from (3.18):

$$\nu_h = h - q + \nu_0. \quad (3.45)$$

Second, we identify a non-stationary AR model using exponential forgetting with $\bar{\nu} = \nu_0, \phi < 1$. This time, we assume that the identification is done on-line, i.e. based on large number of samples. Then, from (3.44):

$$\nu_h = \frac{1 - \phi^{t-q}}{1 - \phi} + \nu_0 \xrightarrow{t \rightarrow \infty} \frac{1}{1 - \phi} + \nu_0. \quad (3.46)$$

Equating (3.45) and (3.46):

$$\phi = \frac{h - q - 1}{h - q} = 1 - \frac{1}{h - q}. \quad (3.47)$$

The interpretation of this choice of ϕ is that it yields Bayesian posterior estimates for A_h and Ω_h which—under both scenarios—have an equal number of degrees of freedom in their uncertainty.

3.3. Probabilistic Principal Component Analysis (PPCA)

Principal Component Analysis (PCA) is one of the classical data analysis tools for dimensionality reduction. It is used in many application areas including data compression, de-noising, pattern recognition, shape analysis and spectral analysis. For an overview of its use, see [80]. A typical example in DSP is spectral analysis [1] or functional analysis of dynamic image data [81].

Probabilistic Principal Component Analysis (PPCA) [63] is a special case of the linear model (3.1), (3.2), with the following assumptions:

$$M_{(r)} = AX', \quad r < \min(p, n), \quad (3.48)$$

$$f(E|\omega) = \mathcal{N}(\mathbf{0}_{p,n}, \omega^{-1}I_p \otimes I_t), \quad (3.49)$$

where scalar $\omega > 0$ denotes precision, and other symbols have their usual meaning (Section 3.1), i.e. $D \in \mathbb{R}^{p \times n}$, $A \in \mathbb{R}^{p \times r}$, $X \in \mathbb{R}^{n \times r}$, $E \in \mathbb{R}^{p \times n}$, and $M_{(r)} \in \mathbb{R}^{p \times n}$. Note that (3.48) is a special case of (3.2) with restriction $r < \min(p, n)$ and using X' , instead of X , for notational simplicity in the sequel. The restriction on rank, r , implies that $\text{rank}(M_{(r)}) = r$, which is explicitly denoted by subscript $M_{(r)}$. The original model of [63] contains an extra parameter μ , modelling a common mean value for the columns, \mathbf{m}_i , of $M_{(r)}$. In this work, we do not impose the restriction of common mean value, i.e. we assume that the common mean value $\mu = \mathbf{0}_{p,1}$. This issue is further discussed in Section 6.5.

Model (3.1), complemented by (3.48), (3.49), yields:

$$f(D|A, X, \omega, r) = \mathcal{N}(AX', \omega^{-1}I_p \otimes I_t). \quad (3.50)$$

Inference of the parameters of the model (3.50) is now reviewed.

3.3.1. Maximum Likelihood Inference

The maximum (3.50), viewed as a function of A , X and ω , but with given r (i.e. the likelihood function) is reached for

$$\hat{M}_{(r)} = U_{r;D} L_{r,r;D} V'_{r;D}, \quad \hat{\omega} = \frac{pn}{\sum_{i=r+1}^p l_{D,i}^2}. \quad (3.51)$$

Here, $U_{r;D}$, and $V_{r;D}$ are the first r columns of the matrices U_D , and V_D respectively, obtained from the SVD [73]

$$D = U_D L_D V'_D. \quad (3.52)$$

$L_{r,r;D}$ is the $r \times r$ upper-left sub-block of matrix L_D .

Remark 3.6 (Rotational ambiguity) *ML estimates of A and X in (3.48), using (3.51), are not unique, because (3.48) exhibits multiplicative degeneracy; i.e.:*

$$\hat{M}_{(r)} = \hat{A} \hat{X}' = (\hat{A} T) (T^{-1} \hat{X}') = \tilde{A} \tilde{X}', \quad (3.53)$$

for any invertible matrix, $T \in \mathbb{R}^{r \times r}$. This is known as rotational ambiguity in the factor analysis literature [47].

The method of Principal Component Analysis (PCA) was originally developed without any explicit noise model [82]. Correspondence of PCA to ML estimation (3.51) of the PPCA model (3.50) was shown later [47, 48]. We briefly review the connection between (3.51) and the classical PCA now.

The classical method of Principal Component Analysis (PCA) is concerned with projections of p -dimensional vectors \mathbf{d}_i , $i = 1, \dots, n$, into an r -dimensional subspace. Optimality of the projection was studied from both a maximum variation [83], and least squares [82] point-of-view. In both cases, the optimal solution leads to eigen-decomposition of the sample covariance matrix

$$S = \frac{1}{n-1} D D' = U \Lambda U', \quad (3.54)$$

where $\Lambda = \text{diag}(\boldsymbol{\lambda})$ is a matrix of eigenvalues of S , and U is the matrix of associated eigenvectors. The columns \mathbf{u}_i , $i = 1, \dots, r$ of U corresponding to the largest eigenvalues λ_i , $\lambda_1 > \lambda_2 > \dots > \lambda_r$, form a basis for the optimal projection sub-space.

Consider the following decomposition of the ML estimate (3.51) of the linear model (3.50)

$$\hat{A} = U_{r;D}, \quad \hat{X} = L_{r,r;D} V'_{r;D}. \quad (3.55)$$

From (3.52), it follows that

$$D D' = U_D L_D V'_D V_D L_D U'_D = U_D L_D L_D U'_D. \quad (3.56)$$

Hence, comparing (3.54) with (3.56), and using (3.55), the following equalities hold:

$$\hat{A} = U_{r;D} = U_{r;}, \quad L_D = (n-1)^{\frac{1}{2}} \Lambda^{\frac{1}{2}}. \quad (3.57)$$

Equalities (3.57) formalize the relation between PCA and ML estimation of the PPCA model.

Remark 3.7 (Ad hoc choice of rank, r) Rank r has been assumed to be known a priori. If this is not the case, many heuristic methods for selection of r exist [80]. One is based on asymptotic properties of the noise. Specifically, from (3.1), (3.49):

$$\begin{aligned} \mathbb{E}_E(DD') &= \mathbb{E}_E(MM') + M\mathbb{E}_E(E') + \mathbb{E}_E(E)M' + \mathbb{E}_E(EE'), \\ &= MM' + n\omega^{-1}I_p. \end{aligned} \quad (3.58)$$

Using the SVD decomposition, $MM' = U_M L_M^2 U_M'$, and noting the equality, $U_M U_M' = I_p$, then, from (3.58), (3.56):

$$\lim_{n \rightarrow \infty} U_D L_D^2 U_D' = U_M L_M^2 U_M' + n\omega^{-1} U_M U_M'. \quad (3.59)$$

It follows that $\lim_{n \rightarrow \infty} U_M = U_D$, and that

$$l_{i;D}^2 = \begin{cases} l_{i;M}^2 + n\omega^{-1} & i \leq r, \\ n\omega^{-1} & i > r, \end{cases} \quad (3.60)$$

Hence, the index, \hat{i} , for which the singular values $l_{i;D}$, $i > \hat{i}$ are constant is considered to be an estimate of rank r . In finite samples, however, (3.60) holds only approximately. The estimate can be chosen by visual examination of the graphed singular values [80], looking for the characteristic ‘knee’ in the graph. In finite samples, it follows from (3.60):

$$\frac{1}{p-r} \sum_{i=r+1}^p l_{i;D}^2 \approx n\omega^{-1}, \quad (3.61)$$

$$\sum_{i=1}^p l_{i;D}^2 \approx \sum_{i=1}^r l_{i;M}^2 + pn\omega^{-1}. \quad (3.62)$$

From ordering of singular values, $l_{1;D} > l_{2;D} > \dots > l_{p;D}$, it follows that $l_{p;D} < \frac{1}{p-r+1} \sum_{i=r}^p l_{i;D}^2$. Hence, using (3.61), we assign an upper bound on ω , namely $l_{p;D}^2 < n\omega^{-1}$. From (3.62), it follows that $\sum_{i=1}^p l_{i;D}^2 > pn\omega^{-1}$, forming a lower bound on ω . This leads to the following choice of interval for $\hat{\omega}$:

$$\frac{pn}{\sum_{i=1}^p l_{i;D}^2} < \hat{\omega} < \frac{n}{l_{p;D}^2}. \quad (3.63)$$

3.3.2. Maximum A Posteriori Inference

An alternative inference of parameters of the PPCA model (3.50) do not maximize directly the likelihood (3.50), but complement (3.50) by a Gaussian prior on X , $f(X) = \mathcal{N}(\mathbf{0}_{r,n}, I_r \otimes I_n)$ and marginalize over X [48, 47]. The resulting maximum of the marginal likelihood, conditioned by r ,

is then reached for $\hat{\omega}$ given by (3.51) and

$$\hat{A}_r = U_{r;D} (L_{r,r;D}^2 - \hat{\omega}^{-1} I_r) R. \quad (3.64)$$

Here, $U_{r;D}$, $L_{r,r;D}$ are given by (3.52), and $R \in \mathfrak{R}^{r \times r}$ is any orthogonal (i.e. rotation) matrix. In this case, indeterminacy of the model is reduced from an arbitrary invertible matrix T (3.53) to an orthogonal matrix R . This reduction is a direct consequence of the restriction imposed on the model via the prior on X .

3.3.3. Variational Bayes Inference

Bayesian inference of the parameters of the PPCA model (3.50) was considered in [7], using a VB approximation (Section 2.2.4). The observation model (3.50) was complemented by the following priors:

$$f(A|\mathbf{v}) = \mathcal{N}(\mathbf{0}_{p,r}, I_p \otimes \Upsilon^{-1}), \quad (3.65)$$

$$f(X) = \mathcal{N}(\mathbf{0}_{r,n}, I_r \otimes I_n), \quad (3.66)$$

$$f(v_i|\alpha_0, \beta_0) = \mathcal{G}(\alpha_0, \beta_0), \quad i = 1, \dots, r, \quad (3.67)$$

$$f(\omega|\vartheta_0, \rho_0) = \mathcal{G}(\vartheta_0, \rho_0), \quad (3.68)$$

where $\Upsilon \in \mathfrak{R}^{r \times r}$ is a diagonal matrix of hyper-parameters, $\Upsilon = \text{diag}(\mathbf{v})$, $\mathbf{v} = [v_1, \dots, v_r]'$, and $\alpha_0, \beta_0, \vartheta_0, \rho_0$ are known scalar parameters. Complementing (3.50) by (3.65)–(3.68) the joint likelihood is:

$$\begin{aligned} f(D, A, X, \Upsilon, \omega|\alpha_0, \beta_0, \vartheta_0, \rho_0, r) &= \mathcal{N}(AX, \omega^{-1} I_p \otimes I_t) \\ &\quad \mathcal{N}(\mathbf{0}_{p,r}, I_p \otimes \Upsilon^{-1}) \mathcal{N}(\mathbf{0}_{r,n}, I_r \otimes I_n) \\ &\quad [\mathcal{G}(\alpha_0, \beta_0)]^r \mathcal{G}(\vartheta_0, \rho_0). \end{aligned} \quad (3.69)$$

The posterior distribution of the model parameters is then obtained using Bayes' rule:

$$f(A, X, \Upsilon, \omega|D, r) = \frac{f(D, A, X, \Upsilon, \omega|r)}{f(D|r)}. \quad (3.70)$$

Here, conditioning by $\alpha_0, \beta_0, \vartheta_0, \rho_0$, was dropped for brevity. Exact posterior inference from (3.70) is not available.

Corollary 3.1 (Corollary 1 of Theorem 2.1) *Consider the following conditionally independent factorization*

$$\check{f}(A, X, \Upsilon, \omega|D, r) = \check{f}(A|D, r) \check{f}(X|D, r) \check{f}(\Upsilon|D, r) \check{f}(\omega|D, r). \quad (3.71)$$

Using (3.70) and (3.71) in Theorem 2.1, the VB-marginals in (3.71) are found as follows:

$$\tilde{f}(A|D, r) = \mathcal{N}(\mu_A, I_p \otimes \Sigma_A), \quad (3.72)$$

$$\tilde{f}(X|D, r) = \mathcal{N}(\mu_X, \Sigma_X \otimes I_t), \quad (3.73)$$

$$\tilde{f}(v_i|D, r) = \mathcal{G}(\alpha_i, \beta_i), \quad i = 1, \dots, r, \quad (3.74)$$

$$\tilde{f}(\omega|D, r) = \mathcal{G}(\vartheta, \rho), \quad (3.75)$$

with the following VB-statistics:

$$\mu_A = \hat{\omega} D \hat{X}' \Sigma_A, \quad (3.76)$$

$$\Sigma_A = \left(\hat{\omega} n \Sigma_X + \hat{\omega} \hat{X} \hat{X}' + \hat{\Upsilon} \right)^{-1}, \quad (3.77)$$

$$\mu_X = \hat{\omega} \Sigma_X \hat{A}' D, \quad (3.78)$$

$$\Sigma_X = \left(\hat{\omega} p \Sigma_A + \hat{\omega} \hat{A}' \hat{A} + I_r \right)^{-1}, \quad (3.79)$$

$$\alpha_i = \alpha_0 + \frac{p}{2}, \quad i = 1, \dots, r, \quad (3.80)$$

$$\beta_i = \beta_0 + \frac{1}{2} (p \Sigma_{A;i,i} + \hat{\mathbf{a}}_i' \hat{\mathbf{a}}_i), \quad i = 1, \dots, r, \quad (3.81)$$

$$\vartheta = \vartheta_0 + \frac{np}{2}, \quad (3.82)$$

$$\begin{aligned} \rho = \rho_0 + \frac{1}{2} \text{tr} \left((D - \hat{A} \hat{X}) (D - \hat{A} \hat{X})' \right), \\ + \frac{1}{2} \hat{\omega}^{-1} \left(p \Sigma_A \hat{X}' \hat{X} + p n \Sigma_A \Sigma_X + n \Sigma_X \hat{A}' \hat{A} \right). \end{aligned} \quad (3.83)$$

In (3.81), notation \mathbf{a}_i denotes the i th column of matrix A , so that $A = [\mathbf{a}_1, \dots, \mathbf{a}_n]$. \hat{A} , \hat{X} , \hat{v}_i and $\hat{\omega}$ denote the expectation with respect to the VB marginals (3.72)–(3.75), so that the associated moments are: $\hat{A} = \mu_A$, $\hat{X} = \mu_X$, $\hat{v}_i = \frac{\alpha_i}{\beta_i}$, and $\hat{\omega} = \frac{\vartheta}{\rho}$. The VB-statistics— μ_A , Σ_A , μ_X , Σ_X , $\alpha = [\alpha_1, \dots, \alpha_r]'$, $\beta = [\beta_1, \dots, \beta_r]'$, ϑ , and ρ —are evaluated via the VEM algorithm (Algorithm 2.2).

Remark 3.8 (Automatic Rank Determination (ARD) Property of VPCA) The VB-statistics α and β can be used for rank selection. It is observed that for some values of the index, i , the posterior expected values $\hat{v}_i = \alpha_i/\beta_i$ converge to the prior value $\hat{v}_i \rightarrow \alpha_0/\beta_0$. This can be explained via prior domination, i.e. the observed data are not informative in those dimensions. Therefore, the rank is determined as the number of \hat{v}_i that are significantly different from the prior value α_0/β_0 . This observation will be called the as Automatic Rank Determination property (ARD)¹.

Remark 3.9 (Laplace approximation) Estimation of the rank of the PPCA model via the Laplace approximation (Section 2.2.2) was published in [84]. There, the parameter A was restricted by orthogonality constraints $A'A = I_r$. The parameter X was treated in the same way as in the VB approximation above, i.e. as a matrix random value with prior (3.66).

¹In the machine learning community, it is known as the Automatic Relevance Determination property. In our case, however, the relevance is with respect to the unknown rank.

3.4. Functional Analysis of Medical Image Sequences (FAMIS)

Functional analysis of dynamic image data (i.e. sequences of images) is an established area in medical imaging. Its aim is to visualize physiological function of biological organs in living creatures. The physiological function is typically measured by volume of a physiological liquid involved in the process. This liquid is marked by a contrast material (e.g. radiotracer) and a sequence of pictures is taken. The key assumption is that there is no relative movement between the camera (e.g. scintigraphic) and the imaged tissues. Under this assumption, the problem can be modelled as a special case of the linear model (3.1), (3.2). The model is closely related to the FA model (3.6). The problem is described in [81] as Factor Analysis of Medical Image Sequences (FAMIS), a nomenclature we will adopt, and review briefly in this Section. Functional analysis is an example of the linear model where parameters have a physical meaning. Therefore, naming conventions used in this area are rather specific. We will follow these conventions, but the general conventions for linear models will be used when the models are discussed in a wider context.

3.4.1. Physiological Model

The task is to analyze the sequence of n images taken at times $t = 1, \dots, n$. Each image stored column wise as a p -dimensional vector of observations \mathbf{d}_t , while the whole sequence forms the matrix $D \in \mathbb{R}^{p \times n}$. It is assumed that each image in the sequence is formed from a linear combination of $r < n$ images of the physiological organs. Formally,

$$\mathbf{d}_t = \sum_{j=1}^r \mathbf{a}_j x_{j;t} + \mathbf{e}_t, \quad (3.84)$$

where \mathbf{a}_j , $j = 1, \dots, r$ are the underlying images of the physiological organs, known as the *factor images*, and $x_{j;t}$ is the weight assigned to the j th factor image at time t . The vector of weights $\mathbf{x}_j = [x_{j;1}, \dots, x_{j;n}]'$ is known as the *factor curve* or the *activity curve* of the j th factor image. The product $\mathbf{a}_j \mathbf{x}_j'$ is known as the *j th factor*. Vector \mathbf{e}_t models the observation noise.

The main application area of functional analysis is in nuclear medicine. In this context, additional restrictions arise. Each pixel of the observation image is acquired as a count of radioactive particles. This has the following consequences:

1. all pixels aggregated in the matrix D , are positive. The factor images, A , are interpreted as observations of isolated physiological organs, hence, the \mathbf{a}_j s are also assumed to be non-negative. The factor curves, X , are interpreted as the variable activity of the associated factor images, which, at each time t , acts to multiply each pixel by the same amount. Therefore, the \mathbf{x}_j s are assumed to be non-negative:

$$\begin{aligned} a_{i,j} &\geq 0, \quad i = 1, \dots, p, \quad j = 1, \dots, r, \\ x_{i,j} &\geq 0, \quad i = 1, \dots, r, \quad j = 1, \dots, n. \end{aligned} \quad (3.85)$$

2. the observed data are known to be Poisson-distributed:

$$f(d_{i;t}) = \mathcal{P}_o \left(\sum_{j=1}^r a_{i;j} x_{j,t} \right), \quad i = 1, \dots, p, j = 1, \dots, n. \quad (3.86)$$

However, inference of model parameters with this distribution is analytically intractable.

Analysis of the sequence is traditionally decomposed in sub-problems which are solved independently. The basic steps are [81]: (i) data pre-processing, (ii) orthogonal analysis, and (iii) oblique analysis.

3.4.2. Data pre-processing

The aim of the pre-processing step was to transform the data model into the PPCA model (3.50). The first task is to approximate the intractable Poisson distribution by a suitable replacement. The problem has been studied theoretically [85, 86, 87, 88], and it was concluded that, asymptotically, the data distribution may be approximated by a Gaussian:

$$f(D|A, X, \Omega_p, \Omega_n) = \mathcal{N}(AX', \Omega_p^{-1} \otimes \Omega_n^{-1}). \quad (3.87)$$

where AX' models the mean value of the signal, and $\Omega_p \in \mathfrak{R}^{p \times p}$, $\Omega_n \in \mathfrak{R}^{n \times n}$ are positive-definite precision matrices. (3.87) implies the following additive decomposition: $D = AX + E$, hence it is another case of the linear class (3.1), (3.2). If the covariance matrices Ω_p^{-1} and Ω_n^{-1} are known *a priori*, the data may be pre-processed as follows:

$$\tilde{D} = \Omega_p^{\frac{1}{2}} D \Omega_n^{\frac{1}{2}}. \quad (3.88)$$

Here, $\Omega_p^{\frac{1}{2}}$ denotes the matrix square-root $\Omega_p^{\frac{1}{2}} \Omega_p^{\frac{1}{2}} = \Omega_p$ [73]. Then, \tilde{D} can be modelled by the probabilistic PCA model (3.50)

$$f(\tilde{D}|A, X, \Omega_p, \Omega_n) = \mathcal{N} \left(\Omega_p^{\frac{1}{2}} AX' \Omega_n^{\frac{1}{2}}, I_p \otimes I_n \right),$$

using elementary properties of the matrix Normal distribution (Appendix A.1, equation (A.3)). The whitening operation (3.88) is known in the factor analysis literature as *scaling*.

The optimal scaling for the Poisson distribution (3.86) is known as *correspondence analysis* [89]:

$$\tilde{d}_{i,j} = \frac{d_{i,j}}{\left(\sum_{k=1}^p d_{k,j} \cdot \sum_{l=1}^n d_{i,l} \right)^{\frac{1}{2}}}. \quad (3.89)$$

This corresponds to choosing diagonal matrices,

$$\begin{aligned} \Omega_p &= \text{diag}(D \mathbf{1}_{n,1})^{-1}, \\ \Omega_n &= \text{diag}(D' \mathbf{1}_{1,p})^{-1}, \end{aligned} \quad (3.90)$$

in the asymptotic model (3.87), where D is the observation matrix.

Appropriateness of the correspondence analysis for the medical sequences was experimentally compared with other *ad hoc* scaling techniques in [90].

3.4.3. Orthogonal analysis

Following the pre-processing, the transformed data, \tilde{D} , may be modelled using the PPCA model (3.88). From now on, it is assured that the data, D , has been pre-processed in this way (i.e. we drop the tilde from notational conventions). Next, the problem is to find a low rank representation of D , i.e. to infer parameters A and X . Traditionally, inference of A and X is addressed using the ML approach, as reviewed in Section 3.3. However, this approach is not sufficient as it does not provide a solution to the following problems:

Number of relevant factors: the ML solution is available only if the number of relevant factors, r , is known *a priori*. Various methods for selection of r —based on both ad-hoc and formal criteria—are available [80]. However, the problem is typically neglected in functional analysis as it is assumed that r may be reasonably guessed from the *biological* knowledge. However, this assumption is valid only for healthy organs. If the organs are damaged, the number of factors in the sequence can increase significantly, and indeed, become a key indicator in the diagnostics of disease states.

Rotation: the probabilistic PCA model does not impose restrictions of positivity on its parameters. The ML solution, \hat{A} and \hat{X} , is confined only to the sub-space spanned by the columns of $\hat{M}_{(r)}$ (3.53) (Remark 3.6). Uniqueness of the solution is assured if the parameters A and X are orthogonal matrices, this will be studied in Section 6.3.1.

It is assumed that the optimal positive-constrained solution is found close to the r -dimensional sub-space inferred by the ML solution (3.53). Therefore, in this step, the orthogonal solution (3.55) is evaluated and rotation towards the physiological factors is addressed in the next step.

3.4.4. Oblique analysis

In this step, the physiologically restricted solution is being searched close to the optimal sub-space (3.53), identified in the previous step. Physiological restrictions of a general nature may be imposed, such as positivity (Section 3.4.1). Alternatively, specific biological knowledge may be used to rotate to valid physiological factors [87]. Extensive discussion on these restrictions can be found in [91].

Uniqueness of decomposition (3.2) under the assumption of positivity of A and X (3.85) was studied in [92]. It was concluded that the decomposition is unique if there exists at least one pixel, $a_{i,j}$, in each factor image, \mathbf{a}_j , for which all corresponding pixels in the remaining factor images are equal to zero:

$$\forall j = 1, \dots, r, \exists i : a_{i;j} > 0 \Rightarrow a_{i;k} = 0, \forall k = 1, \dots, r, k \neq j.$$

This assumption is known as *simple structure*. An algorithm for rotation of the orthogonal estimates towards valid physiological factors was published in [91], by exploiting this uniqueness property.

3.4.5. The FAMIS model

The name “factor analysis” usually denotes the method of inference of parameters, A , X , and ω of the model (3.50). Therefore, we denote (3.50) as the factor analysis model. In the same spirit, we define the FAMIS model now.

The basic model used for functional analysis is the probabilistic PCA (3.50) with additional assumption on the noise (3.87) and positivity of all elements of A and X . These extensions are handled independently as pre-processing (Section 3.4.2) and oblique analysis (Section 3.4.4) respectively. This can be summarized in a unified model as

$$f(D|A, X, \Omega_p, \Omega_n) = \mathcal{N}(AX', \Omega_p^{-1} \otimes \Omega_n^{-1}), \quad (3.91)$$

where the covariance matrices Ω_n , Ω_p are considered known. However, this is rarely true in practice. The presented method—i.e. scaling (3.90)—is optimal for a large number, $n \rightarrow \infty$, of samples. We seek a solution that is optimal in finite number of samples.

Hence, we now consider covariance matrices, Ω_p and Ω_n , as *unknown* parameters with diagonal structure $\Omega_p = \text{diag}(\omega_p)$, $\omega_{i,p} > 0$, $i = 1, \dots, p$, and $\Omega_n = \text{diag}(\omega_n)$, $\omega_{i,n} > 0$, $i = 1, \dots, n$.

This has the following consequences:

- The measured data are corrupted by additional artefacts that are considered as noise from medical point-of-view. The relaxation of known Ω_p , Ω_n allows these artefacts to be captured, and modelled as noise. This should lead to a better signal and noise separation.
- The assumption of diagonality is similar to that of the Factor Analysis (FA) model (3.6). In consequence, inference of the FA model parameters yields a signal and noise separation in such a form that the correlated part of the data is taken as the signal, and the uncorrelated part as the noise² [47].
- Identification of the model with unrestricted precision matrices is not feasible because the number of parameters is then higher than number of available data. The introduced restriction (covariance in the form of Kronecker product and diagonality) keeps the number of estimated parameters well below the number of available data.
- Appropriateness of the method may be compared with the asymptotic result (3.90). If the underlying assumptions are valid, the expected value of Ω_p and Ω_n should be similar to (3.90).

Inference of model parameters in (3.91) is, again, analytically intractable.

3.5. Open problems

In this Chapter, we have presented a review of linear models. Special cases of the linear model—of concern in this thesis—were reviewed in detail. Namely, the AutoRegressive (AR) model (Section 3.2), the

²This may not be appropriate for some applications in medical imaging. In such cases, other restrictions on the precision matrices Ω_p , and Ω_n must be introduced.

Probabilistic PCA (PPCA) model (Section 3.3), and the Factor Analysis for Medical Image Sequences (FAMIS) model (Section 3.4). For these models, a numerically efficient inference of parameters is available under the following restrictive assumptions, respectively:

AR model: the full Bayesian inference is available under assumptions of: (i) known transformation of the system output (Section 3.2.2), and (ii) stationary or slowly-varying parameters (Section 3.2.3).

PPCA model: computationally efficient inference is achieved for the ML and MAP inferences under the assumption of known rank (Section 3.3). The approximate Bayesian inference of the PPCA model (Section 3.3.3) is computationally expensive.

FAMIS model: in order to achieve tractability, the parameter inference is done in three steps (Section 3.4): (i) data pre-processing, (ii) orthogonal analysis, and (iii) oblique analysis. The Bayesian solution is available only for the orthogonal analysis, which is identical to the PPCA problem. The remaining two steps, and indeed the overall problem, has not been addressed from the Bayesian perspective yet.

The aim of this thesis is to relax the above mentioned restrictions and derive numerically tractable inference algorithms for parameter identification in each case. We study the following special cases:

Unknown observation transformation (AR): The parameter inference is analytically tractable if the transformation g is known (Section 3.2.2). Note that modelling of the observation transformation g via an additional linear model is known as the state-space approach [58]. In our forthcoming approach, we do not impose any model on g . We seek a numerically efficient inference algorithm for the model with an unknown g .

Non-stationary parameters (AR): can be modelled by means of the forgetting operator (Section 3.39). The technique of forgetting itself is an optimized approximation of the intractable model involving posterior distribution $f(\theta|D_{t-1})$ and alternative $\bar{f}(\theta|D_t)$ [79]. The analytical solution is preserved if the forgetting factor, ϕ_t , is known at each time t , *a priori*. Typically, it is chosen as time-invariant known constant $\phi_t = \phi$. This is appropriate only for processes with slowly varying parameters. We seek a numerically efficient inference algorithm for the model with an unknown ϕ_t . This would greatly extend the tracking abilities of the inference algorithm for non-stationary AR processes with rapid variations of parameters.

Inference of rank (PPCA): (i.e. number of relevant principal components) is not provided by the ML approach. It can be obtained using Variational Bayes approximate inference for the PPCA model (VPCA) (Section 3.3.3). The computational load of the VPCA algorithm is, however, much higher than that associated with the ML or MAP solution. Moreover, VPCA provides only a point estimate of the rank. In this case, we seek a numerically efficient Bayesian inference of all posterior densities, including distribution for the rank.

Unknown scaling and rank (FAMIS): the standard solution is based on assumptions of (i) known scaling for the pre-processing step, and (ii) known rank (number of relevant factors) for the orthogonal analysis step. The second assumption have been already relaxed since orthogonal analysis is achieved using (Probabilistic) PCA. However, the chosen VB approximation allows to develop a joint identification procedure for the whole model.

For each of these problems we will derive: (i) an analytical analysis of the correct Bayesian solution and justification of the VB approximation; (ii) a Variational Bayesian inference; (iii) a numerically efficient inference algorithm (or, at least, a discussion of this topic); and (iv) experiments on simulated or real-life data. In general, progress in all these tasks will be achieved using the ‘gateway’ of the Variational Bayes (VB) approximation.

Chapter 4.

Mixture-Based Extension of the EAR (MEAR) model

The Extended AutoRegressive model was introduced in Section 3.2.2. The model was designed to extend modelling abilities of the AR model (Section 3.2). This is achieved under the assumption of known and stationary transformation g . In this Chapter, we relax this assumption using a probabilistic mixture approach. Our approach is similar to those in [46, 62]. Our aim is to achieve recursive identification of the EAR model parameters, A, Ω , for a wide class of transformations and distortions.

4.1. The MEAR Model

Following the Bayesian methodology, we treat the unknown *time-variant* transformation, g_t , as a probability entity, γ_t , drawn from a space, \mathcal{G} , of candidates. The conditional distribution (3.29) is then replaced by the marginal

$$f(\mathbf{d}_t | A, \Omega, D_{t-1}) = \int_{\mathcal{G}} f(\mathbf{d}_t | A, \Omega, D_{t-1}, \gamma_t) f(\gamma_t | A, \Omega, D_{t-1}) d\gamma_t, \quad (4.1)$$

where, tacitly, a continuous space is assumed. Evaluation of this distribution is usually prohibitive because the space \mathcal{G} may be extremely rich (recall that the EAR model allows for arbitrary, smooth, non-linear functions with dynamics (Section 3.2.2)). The challenge is to restrict the space \mathcal{G} and reach algorithmically affordable complexity. We assume that \mathcal{G} may be partitioned into a finite number, c , of disjoint subsets:

$$\mathcal{G} = \bigcup_{i=1}^c \mathcal{G}_i. \quad (4.2)$$

Moreover, we assume that the partition can be designed to ensure that effects of all filters in any one subset, \mathcal{G}_i , are very similar. This requirement is summarized in the following conditional independence property for (3.29):

$$f(\mathbf{d}_t | A, \Omega, D_{t-1}, \gamma_t \in \mathcal{G}_i) \approx f(\mathbf{d}_t | A, \Omega, D_{t-1}, g_i), \quad (4.3)$$

where $g_i \in \mathcal{G}_i$ is a representative transformation in the subset \mathcal{G}_i . Substituting (4.3) into (4.1), it follows that:

$$f(\mathbf{d}_t|A, \Omega, D_{t-1}) \approx \sum_{i=1}^c \alpha_i(A, \Omega, D_{t-1}) f(\mathbf{d}_t|A, \Omega, D_{t-1}, g_i) \quad (4.4)$$

$$= f(\mathbf{d}_t|A, \Omega, D_{t-1}, G), \quad (4.5)$$

where

$$\alpha_i(A, \Omega, D_{t-1}) = \int_{\mathcal{G}_i} f(\gamma_t|A, \Omega, D_{t-1}) d\gamma_t, \quad (4.6)$$

$$G = \{g_i, i = 1, \dots, c\}. \quad (4.7)$$

(4.4) is a probabilistic mixture with components, $f(\mathbf{d}_t|A, \Omega, D_{t-1}, g_i)$, with respective weights $\alpha_i(A, \Omega, D_{t-1})$ which are data- and, thus, time-dependent. Note, trivially, that

$$f(\mathbf{d}_t|A, \Omega, D_{t-1}) = f(\mathbf{d}_t|A, \Omega, D_{t-1}, \gamma \in \mathcal{G}).$$

Then, from (4.4):

$$f(\mathbf{d}_t|A, \Omega, D_{t-1}, \gamma \in \mathcal{G}) \approx f(\mathbf{d}_t|A, \Omega, D_{t-1}, G).$$

Hence, approximation (4.4) is valid iff the set G is chosen to satisfy certainty equivalence [20]. (4.7) constitutes a *filter-bank* designed in such a way as to meet this certainty equivalence requirement.

The integral (4.6) can be evaluated only if the partition \mathcal{G}_i is available explicitly. In many practical cases, this will prove difficult to achieve. Therefore, we propose—following the Bayesian methodology—to model the uncertain quantity α_i (4.4) by a probabilistic model. We introduce a labelling transformation:

$$l_i(\gamma_t) = \begin{cases} 1 & \gamma_t \in \mathcal{G}_i, \\ 0 & \gamma_t \notin \mathcal{G}_i, \end{cases} \quad i = 1, \dots, c. \quad (4.8)$$

That can be written in a vector form as $\mathbf{l}(\gamma_t) = [l_1(\gamma_t), \dots, l_c(\gamma_t)]'$. From (4.2), it follows that $\mathbf{l}(\gamma_t) \in \{e_1, \dots, e_c\}$, where

$$e_i = \delta_c(i) = [\delta(i-1), \delta(i-2), \dots, \delta(i-c)]', \quad i = 1, \dots, c.$$

Considering \mathcal{G}_i as unknown, we can define a new random variable $\mathbf{l}_t \equiv \mathbf{l}(\gamma_t)$, with pdf:

$$f(\mathbf{l}_t = e_i|\cdot) = \Pr(\gamma_t \in \mathcal{G}_i|\cdot) = \int_{\mathcal{G}_i} f(\gamma_t|\cdot) d\gamma_t. \quad (4.9)$$

Note that the last term in (4.9) is in the form of the mixture weight (4.6). Hence, using (4.9), we can assign

$$\alpha_i(A, \Omega, D_{t-1}) = f(\mathbf{l}_t = e_i|A, \Omega, D_{t-1}, \mathcal{G}_i). \quad (4.10)$$

However, the weight (4.10) requires an explicit model of the space \mathcal{G}_i . This is prohibitive from a computational point of view. Therefore, most of the literature on statistical mixtures (e.g. [25])

approximates

$$f(\mathbf{l}_t|A, \Omega, D_{t-1}) \approx f(\mathbf{l}_t|\boldsymbol{\alpha}) = \text{Mu}_{\mathbf{l}_t}(\boldsymbol{\alpha}),$$

i.e. a multinomial distribution with time-invariant vector parameter $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_c]'$. This yields a mixture model with stationary weights $\alpha_i(A, \Omega, D_{t-1}) = \alpha_i$. This assumption is however, not realistic for non-stationary processing. We address the problem as follows.

Proposition 4.1 (Markov weights) Variable \mathbf{l}_t constitutes a hidden field which we model via a first-order Markov chain with transition matrix $T \in [0, 1]^{c \times c}$:

$$\alpha_i(A, \Omega, D_{t-1}) \approx f(\mathbf{l}_t|T, \mathbf{l}_{t-1}) = \text{Mu}_{\mathbf{l}_t}(T\mathbf{l}_{t-1}) = \prod_{i=1}^c \prod_{j=1}^c t_{i,j}^{l_{i,t}; l_{j,t-1}}; \quad (4.11)$$

i.e. $\Pr(\mathbf{l}_t = \mathbf{e}_i | T, \mathbf{l}_{t-1} = \mathbf{e}_j) = t_{i,j}$, the ij th element of T . $\text{Mu}_{\mathbf{l}_t}(\cdot)$ denotes the multinomial distribution.

Recall, from Section 3.2, that the Extended AR (EAR) model is an AR model on transformed data:

$$\begin{aligned} \mathbf{y}_{d,t} &= A\mathbf{x}_t + \Omega^{-\frac{1}{2}}\mathbf{e}_t, \\ \mathbf{x}_t &= \mathbf{g}_x(D_{t-1}, W_t). \end{aligned}$$

For algebraic simplicity, we have introduced an extended regressor, $\mathbf{y}_t = [\mathbf{y}_{d,t}, \mathbf{x}_t] = \mathbf{g}(D_{t-1}, W_t)$, which is dependent on the transformation $\mathbf{g}(\cdot)$. In this Chapter, the time-invariant transformation $\mathbf{g}(\cdot)$ was replaced by a filter-bank $G = [\mathbf{g}_1, \dots, \mathbf{g}_c]$. The regressor corresponding to the i th transformation will be denoted as follows:

$$\mathbf{y}_{i,t} = \mathbf{g}_i(D_{t-1}, W_t), \quad i = 1, \dots, c. \quad (4.12)$$

The history of the regressor—which is used mostly in conditioning part of the posterior pdfs—is adapted to $Y_t = \left[[\mathbf{y}'_{1,1}, \dots, \mathbf{y}'_{c,1}]', \dots, [\mathbf{y}'_{1,t}, \dots, \mathbf{y}'_{c,t}]' \right]$. Intuitively, it denotes the knowledge of all observed data under all considered transformations.

Substituting (4.11), into (4.4), the observation model is

$$f(\mathbf{d}_t|A, \Omega, T, Y_{t-1}, G, \mathbf{l}_{t-1}) = \sum_{i=1}^c \prod_{j=1}^c t_{i,j}^{l_{j,t}; l_{j,t-1}} f(\mathbf{d}_t|A, \Omega, \mathbf{y}_{i,t}, G), \quad (4.13)$$

where the conditioning set A, Ω, Y_{t-1}, G has been augmented by T, \mathbf{l}_{t-1} .

4.2. Bayesian Formulation

Consider the joint distribution of the observation \mathbf{d}_t and the label \mathbf{l}_t :

$$f(\mathbf{d}_t, \mathbf{l}_t|A, \Omega, T, Y_{t-1}, G, \mathbf{l}_{t-1}) = f(\mathbf{d}_t|A, \Omega, Y_{t-1}, G, \mathbf{l}_t) f(\mathbf{l}_t|T, \mathbf{l}_{t-1}). \quad (4.14)$$

Then, the marginal distribution of (4.14) over \mathbf{l}_t is the observation model (4.13). Next, consider the posterior distribution of model parameters of (4.14) at time $t - 1$, i.e. $f(A, \Omega, T, \mathbf{l}_{t-1}|D_{t-1}, G)$. This

is updated by (4.14) according to Bayes' rule:

$$f(A, \Omega, T, \mathbf{l}_t, \mathbf{l}_{t-1} | Y_t, G) \propto f(A, \Omega, T, \mathbf{l}_{t-1} | Y_{t-1}, G) f(\mathbf{d}_t, \mathbf{l}_t | A, \Omega, T, \mathbf{l}_{t-1}, Y_{t-1}, G). \quad (4.15)$$

The update introduces, at each step, an extra random variable, \mathbf{l}_t . Hence, the parameter distributions at times t and $t - 1$ have different functional forms, violating conjugacy. After t updates, t random variables will have been generated, with c^t possible states. This scenario has been used in the off-line case [46], but it is unsuitable for on-line identification.

The exponential explosion of terms, described above, is overcome via the following conditional independence approximation of the posterior distribution at time t (4.15):

$$\check{f}(A, \Omega, T, \mathbf{l}_t, \mathbf{l}_{t-1} | Y_t, G) = \check{f}(A, \Omega, T | Y_t, G) \check{f}(\mathbf{l}_t | Y_t) \check{f}(\mathbf{l}_{t-1} | Y_t), \quad (4.16)$$

where the $\check{f}(\cdot)$ denote 'wildcard' approximating distribution. Using (4.16) at both t and $t - 1$ (i.e. for the first two terms in (4.15) respectively), we see that $\check{f}(A, \Omega, T | Y_t)$ is updated in the step from $t - 1$ to t independently of the label sequence \mathbf{l}_t , avoiding the exponential explosion.

4.3. Variational Bayes (VB) Approximation

The conditional independence (4.16) is the underlying assumption of the VB approximation method (Section 2.2.4). In order to achieve conjugacy-based recursive identification, we seek a posterior distribution on parameters, A, Ω, T , at time $t - 1$ to be of the same form as at time t . The functional optimization achieved by the VB approximation allows us to choose the posterior distribution conjugate to the VB-optimized observation model (Section 2.3.3).

4.3.1. VB-conjugate Prior

Let the distribution of model parameters at time $t - 1$ to be of the form (4.16). It is updated by the extended observation model (4.14) to yield the posterior distribution (4.15). Taking the logarithm of the joint distribution, then

$$\begin{aligned} \ln f(\mathbf{d}_t, \mathbf{l}_t, A, \Omega, T, \mathbf{l}_{t-1} | Y_{t-1}, \mathbf{x}_t, G) &= \sum_{i=1}^c l_{i,t} \left[\ln f(\mathbf{d}_t | A, \Omega, \mathbf{x}_{i:t}, \mathbf{g}_{i:t}) + \sum_{j=1}^c l_{j,t-1} \ln t_{i,j} \right] \\ &+ \ln \check{f}(A, \Omega, T | Y_{t-1}) + \ln \check{f}(\mathbf{l}_t | Y_{t-1}) + \ln \check{f}(\mathbf{l}_{t-1} | Y_{t-1}). \end{aligned} \quad (4.17)$$

Using Theorem 2.1 for (4.17) with restrictions (4.16), the VB-approximate distribution on A, Ω, T is found to be

$$\begin{aligned} \tilde{f}(A, \Omega, T|Y_t) \propto & \exp \left(\sum_{i=1}^c \hat{l}_{i,t} \left[\ln f(\mathbf{d}_t|A, \Omega, \mathbf{y}_{i,t}, \mathbf{g}_{i,t}) + \sum_{j=1}^c \hat{l}_{j,t-1} \ln t_{i,j} \right] \right. \\ & \left. + \ln \check{f}(A, \Omega, T|Y_{t-1}) \right), \end{aligned} \quad (4.18)$$

$$\propto \check{f}(A, \Omega, T|Y_{t-1}) \prod_{i=1}^c f(\mathbf{d}_t|A, \Omega, \mathbf{y}_{i,t}, \mathbf{g}_{i,t})^{\hat{l}_{i,t}} \prod_{i=1}^c \prod_{j=1}^c t_{i,j}^{\hat{l}_{i,t} \hat{l}_{j,t-1}}. \quad (4.19)$$

The expected values $\hat{l}_{i,t}$ of $\tilde{f}(\mathbf{l}_t|Y_t)$ and will be evaluated shortly. These approximate distributions of \mathbf{l}_t and \mathbf{l}_{t-1} are found in the form

$$\tilde{f}(\mathbf{l}_t|Y_t) = \check{f}(\mathbf{l}_t|Y_{t-1}) \prod_{i=1}^c \eta_{i,t}^{l_{i,t}}, \quad (4.20)$$

$$\tilde{f}(\mathbf{l}_{t-1}|Y_t) = \check{f}(\mathbf{l}_{t-1}|Y_{t-1}) \prod_{i=1}^c \kappa_{i,t}^{l_{i,t-1}}, \quad (4.21)$$

$$\begin{aligned} \eta_{i,t} &= \exp \left[\mathbb{E}_{A, \Omega, T|Y_{t-1}} \left(-\frac{1}{2} \text{tr}(\Omega [-I_p, A] [\mathbf{y}_{i,t} \mathbf{y}'_{i,t}] [-I_p, A]') \right. \right. \\ & \quad \left. \left. + \sum_{j=1}^c l_{j,t-1} \ln t_{i,j} \right) + \ln |J_{j,t}| \right], \\ \kappa_{i,t} &= \exp \left[\mathbb{E}_{T, \mathbf{l}_t} \left(\sum_{j=1}^c l_{j,t} \ln t_{j,i} \right) \right], \end{aligned}$$

where $\eta_{i,t}$ and $\kappa_{i,t}$, $i = 1, \dots, c$ are statistics of distributions (4.20) and (4.21) respectively.

In order to achieve tractable recursive identification we want to choose approximate distributions $\tilde{f}(\cdot)$ to be closed under these VB updates (i.e. VB-conjugacy, Section 2.3.3). We note the following:

- if $\check{f}(A, \Omega, T|Y_{t-1})$ is chosen in the form $\tilde{f}(A, \Omega|Y_{t-1}) \tilde{f}(T|Y_{t-1})$ —i.e. with independence between the AR parameters A, Ω and weights T —the approximate posterior (4.19) is also independent.
- parameters A, Ω are present in the VB-approximate observation model (4.19), only via the conditioning part of the distribution of data \mathbf{d}_t . This distribution is the product of Normal distributions being therefore a Normal distribution. Hence, $\check{f}(A, \Omega|Y_{t-1})$ of the Normal-Wishart type is conjugate to it.
- parameter T is present in (4.19), only via the product of multinomial distributions which is also a multinomial distribution. Hence, $\check{f}(T|Y_t)$ of the Dirichlet type is conjugate to it.
- both $\tilde{f}(\mathbf{l}_t|Y_{t-1})$ and $\tilde{f}(\mathbf{l}_{t-1}|Y_{t-1})$ in (4.20) and (4.21) are self-replicating if they are chosen as Multinomial.

- $\tilde{f}(\mathbf{l}_t|Y_{t-1})$ is, in fact, the prior distribution on \mathbf{l}_t , as \mathbf{l}_t was not considered in previous updates. We choose it to be uniform on $[1, \dots, c]$, i.e. $Mu_{\mathbf{l}_t}(c^{-1}\mathbf{1}_{p,1})$.

This VB-conjugate approximate distribution (4.16) is then of the form:

$$\begin{aligned} \tilde{f}(A, \Omega, T, \mathbf{l}_t, \mathbf{l}_{t-1}|V_{t-1}, \nu_{t-1}, \Phi_{t-1}) &= \mathcal{NW}_{A,\Omega}(V_{t-1}, \nu_{t-1}) \mathcal{D}i_T(\Phi_{t-1}) \\ &Mu_{\mathbf{l}_t}(c^{-1}\mathbf{1}_{p,1}) Mu_{\mathbf{l}_{t-1}}(\mathbf{w}_{t-1}). \end{aligned} \quad (4.22)$$

Here, $\mathcal{NW}_{A,\Omega}(V_{t-1}, \nu_{t-1})$ is the Normal-Wishart distribution with statistics V_{t-1} and ν_{t-1} ; $\mathcal{D}i_T(\Phi_{t-1})$ denotes the Dirichlet distribution with statistics $\Phi_{t-1} \in \mathbb{R}^{c \times c}$, Appendix A.3; and $Mu_{\mathbf{l}_{t-1}}(\mathbf{w}_{t-1})$ denotes the Multinomial distribution with statistics $\mathbf{w}_{t-1} \in \mathbb{R}^{c \times 1}$.

4.3.2. VB-optimized Posterior Distribution

Substituting (4.22) into (4.15) yields the following joint distribution:

$$\begin{aligned} f(\mathbf{d}_t, \mathbf{l}_t, \mathbf{l}_{t-1}, A, \Omega, T|Y_t) &= \left(\prod_{i=1}^c f(\mathbf{d}_t|A, \Omega, \mathbf{y}_{i,t}, \mathbf{g}_{i,t})^{l_{i,t}} \right) \left(\prod_{i=1}^c \prod_{j=1}^c t_{i,j}^{l_{i,t} l_{j,t-1}} \right) \mathcal{D}i_T(\Phi_{t-1}) \\ &\mathcal{NW}_{A,\Omega}(V_{t-1}, \nu_{t-1}) Mu_{\mathbf{l}_t}(c^{-1}\mathbf{1}_{p,1}) Mu_{\mathbf{l}_{t-1}}(\mathbf{w}_{t-1}). \end{aligned} \quad (4.23)$$

Corollary 4.1 (Corollary 1 of Theorem 2.1) *Using (4.16) and (4.23) in Theorem 2.1, the VB-optimal form of (4.16) is found via the following assignments:*

$$\tilde{f}(A, \Omega|Y_t) = \mathcal{NW}_{A,\Omega}(V_t, \nu_t), \quad (4.24)$$

$$\tilde{f}(T|Y_t) = \mathcal{D}i_T(\Phi_t), \quad (4.25)$$

$$\tilde{f}(\mathbf{l}_t|Y_t) = Mu_{\mathbf{l}_t}(\mathbf{w}_t), \quad (4.26)$$

$$\tilde{f}(\mathbf{l}_{t-1}|Y_t) = Mu_{\mathbf{l}_{t-1}}(\mathbf{u}_t), \quad (4.27)$$

with VB-statistics

$$V_t = V_{t-1} + \sum_{j=1}^c \hat{l}_{j,t} \mathbf{y}_{j,t} \mathbf{y}'_{j,t}, \quad (4.28)$$

$$\nu_t = \nu_{t-1} + 1, \quad (4.29)$$

$$\Phi_t = \Phi_{t-1} + \hat{\mathbf{l}}_t \hat{\mathbf{l}}_{t-1}', \quad (4.30)$$

$$\begin{aligned} w_{i,t} &\propto |J_{i,t}| \exp \left[-\frac{1}{2} \mathbf{y}'_{i,t} \left[-I_p, \hat{A} \right]' \hat{\Omega} \left[-I_p, \hat{A} \right] \mathbf{y}_{i,t} \right. \\ &\quad \left. - \frac{1}{2} p \mathbf{y}'_{i,t} V_{\mathbf{a}\mathbf{a};t}^{-1} \mathbf{y}_{i,t} + \sum_{j=1}^c \hat{l}_{j,t-1} \widehat{\ln t_{i,j}} \right] \end{aligned} \quad (4.31)$$

$$u_{i,t} \propto w_{i,t-1} \sum_{j=1}^c \hat{l}_{j,t} \widehat{\ln t_{j,i}} \quad (4.32)$$

The constants of proportionality in (4.31), and (4.32) follow from normalizations, $\sum_{j=1}^c w_{j,t} = 1$, and $\sum_{j=1}^c u_{j,t} = 1$, respectively. Moments of (4.24), i.e. $\hat{A} = E_{A|V_t, \nu_t}(A)$ and $\hat{\Omega} = E_{\Omega|V_t, \nu_t}(\Omega)$,

are given by (3.20) and (3.21) respectively. The moment of (4.25) required for (4.31), (4.32), i.e. $\widehat{\ln t_{i,j}} = \mathbb{E}_{T|\Phi_t}(\ln t_{i,j})$, is given in Appendix A.3, namely (A.22). The first moment of (4.26) and (4.27) are $\widehat{\mathbf{l}}_t = [w_{1;t}, \dots, w_{c;t}]$ and $\widehat{\mathbf{l}}_{t-1} = [u_{1;t}, \dots, u_{c;t}]$ respectively.

Proof: Logarithm of the joint distribution (4.23) is:

$$\ln f(\mathbf{d}_t, \mathbf{l}_t, \mathbf{l}_{t-1}, A, \Omega | \boldsymbol{\alpha}, Y_t) = \quad (4.33)$$

$$\begin{aligned} &= \sum_{i=1}^c l_{i;t} \left[\ln f(\mathbf{d}_t | A, \Omega, \mathbf{y}_{i,t}, \mathbf{g}_{i,t}) + \sum_{j=1}^c l_{j;t-1} \ln t_{i,j} \right] \\ &\quad + \ln \mathcal{N}_{\mathcal{W}_{A,\Omega}}(V_{t-1}, \nu_{t-1}) + \ln \mathcal{D}i_t(\Phi_{t-1}) + \ln M u_t(\mathbf{w}_{t-1}), \\ &= -\frac{r}{2} \ln(2\pi) + \frac{1}{2} \ln |\Omega| + \sum_{i=1}^c l_{i;t} \left[\ln |J_{i,t}| + \sum_{j=1}^c l_{j;t-1} \ln t_{i,j} \right] \\ &\quad + \sum_{i=1}^c l_{i;t} \left[-\frac{1}{2} \text{tr}(\Omega [-I_p, A] [\mathbf{y}_{i,t} \mathbf{y}'_{i,t}] [-I_p, A]') \right] \\ &\quad + \frac{1}{2} \nu_{t-1} \ln |\Omega| - \ln \zeta_{\mathcal{N}\mathcal{W}}(V_{t-1}, \nu_{t-1}) - \frac{1}{2} \Omega [-I_p, A] V_{t-1} [-I_p, A]' \\ &\quad - \ln \zeta_{\mathcal{D}i}(\Phi_{t-1}) + \sum_{i=1}^c \sum_{j=1}^c \phi_{i,j;t-1} \ln t_{i,j} + \sum_{i=1}^c l_{i,t-1} \ln w_{i;t-1}. \end{aligned} \quad (4.34)$$

From Theorem 2.1, distributions (4.24) and (4.25) are obtained as proportional to exponential of the expected value of (4.33) over $\tilde{f}(\mathbf{l}_t | \mathbf{d}_t, Y_t)$. As (4.33) is linear in \mathbf{l}_t and \mathbf{l}_{t-1} , the expectation is just a replacement of \mathbf{l}_t by $\widehat{\mathbf{l}}_t$, and \mathbf{l}_{t-1} by $\widehat{\mathbf{l}}_{t-1}$. Removing all terms independent of A, Ω from (4.33) and normalizing we obtain (4.24) with assignments (4.28) and (4.29). Removing all terms independent of T from (4.33) and normalizing, we obtain (4.25) with assignment (4.30).

Distribution (4.26) is proportional to exponential of expected value of (4.33) with respect to distributions (4.24) and (4.25). All terms independent of \mathbf{l}_t become part of the normalizing constant, hence (4.26) is obtained, via assignment

$$w_{i;t} \propto \exp \left[\mathbb{E}_{A,\Omega,T,\mathbf{l}_{t-1}} \left(-\frac{1}{2} \text{tr}(\Omega [-I_p, A] [\mathbf{y}_{i,t} \mathbf{y}'_{i,t}] [-I_p, A]') + \sum_{j=1}^c l_{j;t-1} \ln t_{i,j} \right) + \ln |J_{i,t}| \right].$$

Using elementary properties of the $\text{tr}(\cdot)$ operator and the properties of the matrix Normal distribution (Appendix A.1), namely (A.2):

$$\begin{aligned}
 w_{i;t} &\propto |J_{i;t}| \exp \left[E_{A,\Omega} \left(-\frac{1}{2} \mathbf{y}'_{i,t} [-I_p, A]' \Omega [-I_p, A] \mathbf{y}_{i,t} \right) + \sum_{j=1}^c E_{l_{t-1}}(l_{j;t-1}) E_T(\ln t_{i,j}) \right], \\
 &\propto |J_{i;t}| \exp \left[E_{A,\Omega} \left(-\frac{1}{2} \mathbf{y}'_{i,t} \begin{bmatrix} \Omega & -\Omega A \\ -A' \Omega & A' \Omega A \end{bmatrix} \mathbf{y}_{i,t} \right) + \sum_{j=1}^c \widehat{l}_{j;t-1} \widehat{\ln t_{i,j}} \right], \\
 &\propto |J_{i;t}| \exp \left[E_{\Omega} \left(-\frac{1}{2} \mathbf{y}'_{i,t} \begin{bmatrix} \Omega & -\Omega \widehat{A} \\ -\widehat{A}' \Omega & \widehat{A}' \Omega \widehat{A} + p V_{\mathbf{a}\mathbf{a};t}^{-1} \end{bmatrix} \mathbf{y}_{i,t} \right) + \sum_{j=1}^c \widehat{l}_{j;t-1} \widehat{\ln t_{i,j}} \right], \\
 &\propto |J_{i;t}| \exp \left[-\frac{1}{2} \mathbf{y}'_{i,t} [-I_p, \widehat{A}]' \widehat{\Omega} [-I_p, \widehat{A}] \mathbf{y}_{i,t} \right. \\
 &\quad \left. -\frac{1}{2} p \mathbf{y}'_{i,t} V_{\mathbf{a}\mathbf{a};t}^{-1} \mathbf{y}_{i,t} + \sum_{j=1}^c \widehat{l}_{j;t-1} \widehat{\ln t_{i,j}} \right],
 \end{aligned}$$

proving (4.31). The mean value of (4.26) follows trivially from the fact that all possible realizations of l_t are elementary basis functions. \blacksquare

VB-statistics (4.28)–(4.32) can be evaluated via the standard VEM algorithm (Algorithm 2.2). However, as the VB approximation is applied to a single step, we need to iterate the solution for each step of the on-line algorithm, as follows:

For each t :

1. collect data record \mathbf{d}_t
2. assign initial values $V_t^{(0)}, \nu_t^{(0)}, \Phi_t^{(0)}, \mathbf{w}_t^{(0)}, \mathbf{u}_t^{(0)}$, (e.g. $V_t^{(0)} = V_{t-1}$, etc.)
3. iterate (4.28)–(4.32) using the VEM algorithm (Algorithm 2.2) until convergence is reached at, say, the m th iteration.
4. assign the approximate statistics at time t as: $V_t = V_t^{(m)}, \nu_t = \nu_t^{(m)}, \Phi_t = \Phi_t^{(m)}, \mathbf{w}_t = \mathbf{w}_t^{(m)}, \mathbf{u}_t = \mathbf{u}_t^{(m)}$.

end

This, of course, may prove impractical for applications requiring real-time processing, since the convergence of step 3 is not guaranteed in a given number of operations. This problem can be addressed by setting a threshold m_{\max} on the maximum allowed number of iterations of the VEM algorithm, as suggested, for example, in [35] where $m_{\max} = 1$. For time-invariant models, the algorithm asymptotically (i.e. for $t \rightarrow \infty$) converges to the local Variational extreme, but, it does not hold for the time-varying models. Alternatively, we can use the Restricted VB approximation (RVB) as introduced in Section 2.2.5. We consider this next.

4.4. Quasi-Bayes (QB) Approximation

In this Section, we derive an alternative identification algorithm using Restricted Variational Bayes (RVB) (Section 2.2.5). The RVB approximation requires all but one VB-marginal to be known. Moreover, the Quasi-Bayes (QB) approximation (Remark 2.4) use the analytical marginals of the true posterior distribution. Here, we note that l_t is a discrete variable with $c < \infty$ states, hence, marginalization over this label field is analytically tractable.

4.4.1. Fixing the VB-marginal for the Label Field

Label variables in (4.23), namely l_t and l_{t-1} together possess c^2 possible states. Evaluation of c^2 possibilities may be prohibitive for large c . Therefore, we make the following choice of fixed VB-marginals in the RVB approximation (2.34):

$$\tilde{f}(l_{t-1}|Y_t) = \tilde{f}(l_{t-1}|Y_{t-1}), \quad (4.35)$$

$$\tilde{f}(l_t|Y_t) = \int_{\mathbf{d}_t, A, \Omega, T} \sum_{l_{t-1}} f(\mathbf{d}_t, A, \Omega, T, l_t, l_{t-1}|Y_{t-1}, \mathbf{x}_t) d\mathbf{d}_t dA d\Omega dT. \quad (4.36)$$

Hence, (4.35) was chosen as fixed at the previous time-data step, and (4.36) was chosen as suggested by Remark (2.4).

Marginal (4.36) of the joint distribution (4.23) over $A, \Omega, T, \mathbf{d}_t$, is a discrete distribution of the form

$$\tilde{f}(l_t|Y_t) = \prod_{i=1}^c w_{i;t}^{l_{i;t}}, \quad (4.37)$$

Where $w_{i;t}$ can be found via

$$\begin{aligned} w_{i,t} &= \int_{\mathbf{d}_t, A, \Omega, T} \sum_{j=1}^c w_{j;t-1} f(A, \Omega, T, l_t = e_i, \mathbf{d}_t|Y_{t-1}, \mathbf{x}_t, l_{t-1} = e_j) dA d\Omega dT \\ &\propto \int_{\mathbf{d}_t, A, \Omega, T} \sum_{j=1}^c w_{j;t-1} f(A, \Omega, T, l_t = e_i, \mathbf{d}_t|Y_{t-1}, \mathbf{x}_t, l_{t-1} = e_j) dA d\Omega dT \\ &= \int_{\mathbf{d}_t, A, \Omega, T} \left[f(\mathbf{d}_t|A, \Omega, T, \mathbf{y}_{i,t}) \mathcal{N}\mathcal{W}_{A, \Omega}(V_{t-1}, \nu_{t-1}) + \right. \\ &\quad \left. \sum_{j=1}^c w_{j;t} \mathcal{D}^i_T(\Phi_{t-1}) f(l_t = e_i|T, l_{t-1} = e_j) \right] dA d\Omega dT \end{aligned} \quad (4.38)$$

$$\propto \zeta_{\mathcal{N}\mathcal{W}}(V_{t-1} + \mathbf{y}_{i,t} \mathbf{y}'_{i,t}, \nu_{t-1} + 1) \sum_{j=1}^c w_{j;t-1} \zeta_{\mathcal{D}^i}(\phi_{i,j;t-1} + 1) \quad (4.39)$$

where $\zeta_{\mathcal{N}\mathcal{W}}(\cdot)$ is given by (3.14) and $\zeta_{\mathcal{D}^i}(\cdot)$ is given by (A.20) in Appendix A.20. The constant of proportionality for (4.39) is easily determined from normalization of (4.37), i.e. $\sum_{j=1}^c w_{j;t} = 1$.

4.4.2. QB-optimal Posterior Distribution

The full updating algorithm follows from Corollary 4.1, using the RVB approach (Corollary 2.1, Section 2.2.5).

Corollary 4.2 (Quasi-Bayes (QB) estimation of the MEAR model) *Using (4.16) with assignments (4.35), and (4.36) in Corollary 2.1, the RVB-optimal form of (4.16) is found via the following assignments:*

$$\tilde{f}(A, \Omega | Y_t) = \mathcal{NW}(V_t, \nu_t), \quad (4.40)$$

$$\tilde{f}(T | Y_t) = \mathcal{D}_i(\Phi_t), \quad (4.41)$$

with statistics

$$V_t = V_{t-1} + \sum_{i=1}^c w_{i;t} \mathbf{y}_{i,t} \mathbf{y}'_{i,t}, \quad (4.42)$$

$$\nu_t = \nu_{t-1} + 1, \quad (4.43)$$

$$\Phi_t = \Phi_{t-1} + \mathbf{w}_t \mathbf{w}'_{t-1}. \quad (4.44)$$

Proof: (4.40)–(4.44) are of the same form as (4.24) and (4.25). Expected value $\hat{\mathbf{l}}_t$ follows form (4.37). Substituting $\hat{\mathbf{l}}_t = \mathbf{w}_t$ and $\hat{\mathbf{l}}_{t-1} = \mathbf{w}_{t-1}$ from (4.39) into (4.28)–(4.30) proves (4.42)–(4.44). ■

4.5. Viterbi-Like (VL) approximation

Note that the matrix V_t is updated c -times by a dyad weighted by corresponding weight $w_{i;t}$. Dyadic update is a rather expensive operation (Section 3.2.1.1). In situations where one weight $w_{i;t}$ is dominant, it may be unnecessary to perform dyadic updates for the remaining $c - 1$ dyads with low weights. This motivates the following *ad hoc* proposition.

Proposition 4.2 (Viterbi-like Algorithm) *Further simplification of the QB algorithm may be achieved using an even coarser approximation of the label-field distribution, namely certainty equivalence (Section 2.2.1):*

$$\tilde{f}(\mathbf{l}_t | Y_t) = \delta(\mathbf{l}_t - \hat{\mathbf{l}}_t), \quad (4.45)$$

in place of (4.36). Here, $\hat{\mathbf{l}}_t$ is the MAP estimate from (4.37), i.e.

$$\hat{\mathbf{l}}_t = \arg \max_{\mathbf{l}_t} \tilde{f}(\mathbf{l}_t | Y_t). \quad (4.46)$$

This corresponds to the choice of one ‘active’ model with index $\hat{i}_t \in \{1, \dots, c\}$, such as $\hat{\mathbf{l}}_t = \mathbf{e}_{i_t}$. The idea is related to the Viterbi algorithm [93]. Replacing (4.37) in the Corollary 4.2 by (4.45), we obtain

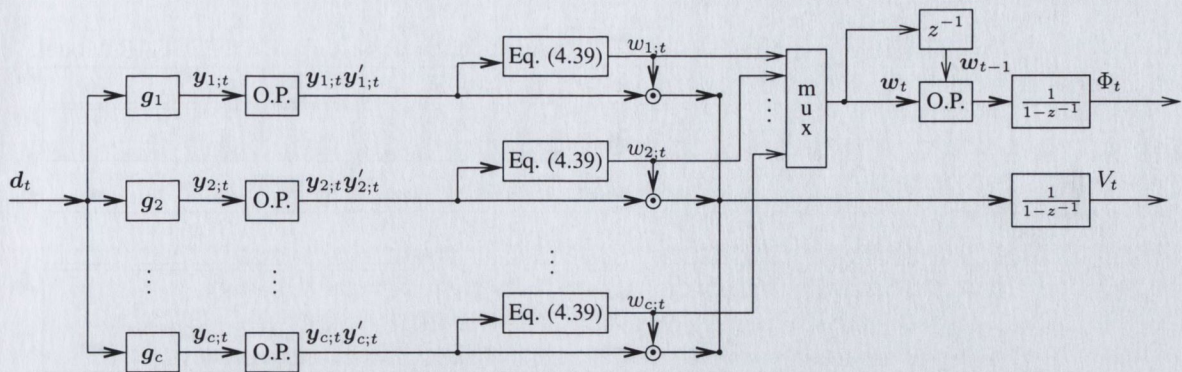


Figure 4.1.: The Recursive summed-dyad computational scheme for Quasi-Bayes identification of the MEAR model. O.P. denotes outer product (dyad). For clarity, dependence of Eq. (4.39) on V_{t-1} and k_{t-1} is not shown.

the approximating distributions in the same form as (4.40)–(4.41), but with statistics

$$V_t = V_{t-1} + \mathbf{y}_{i,t} \mathbf{y}'_{i,t}, \quad (4.47)$$

$$\nu_t = \nu_{t-1} + 1, \quad (4.48)$$

$$\Phi_t = \Phi_{t-1} + \mathbf{w}_t \mathbf{w}'_{t-1}. \quad (4.49)$$

Note that the update of Φ_t (4.49) is computationally cheap. Weights \mathbf{w}_t are already available for evaluation of (4.45), hence, Φ_t is updated as in the QB algorithm (Corollary 4.2, equation (4.44)).

4.6. Inference with the MEAR model

4.6.1. Computational Issues

We have introduced three methods for recursive identification of the MEAR model:

1. Variational Bayes (VB) algorithm (Corollary 4.1)
2. Quasi-Bayes (QB) algorithm (Corollary 4.2)
3. Viterbi-Like (VL) algorithm (Proposition 4.2)

The computational flow is the same for all algorithms involving updates of statistics V_t, ν_t, Φ_t .

The recursive scheme for computation of (4.42)–(4.44) via the QB algorithm is displayed in Figure 4.1. The computational scheme for the VB algorithm is, in principle, the same, but the statistics, $w_{i,t}$, Φ_t and V_t , must be iterated in each time using the VEM algorithm. This is difficult to visualize. The main points of interest are the weight evaluation, i.e. Eq. (4.39), and summation of dyads.

Weights are computed via (4.31) for VB, and by (4.39) for the QB and VL algorithms. The operations required for this step are:

4. MIXTURE-BASED EXTENSION OF THE EAR (MEAR) MODEL

Table 4.1.: Computational complexity of recursive identification algorithms for the MEAR Model.

algorithm	computational complexity of one-step estimation
VB	$m(2c+1) \times O\left((r+p)^2\right)$
QB	$2c \times O\left((r+p)^2\right) + c \times O(r+p)$
Viterbi-like	$(c+1) \times O\left((r+p)^2\right) + c \times O(r+p)$
	m denotes number of iterations of the VEM algorithm (for VB only) c is the number of components in the MEAR model p and r are the dimensions of measured data and regressor, respectively

VB: (i) evaluation of $\hat{A}, \hat{\Omega}, V_{aa;t}$; (ii) evaluation of (4.31) c -times. All operations in (i) and each operation in (ii) are of complexity $O\left((r+p)^2\right)$. (Section 3.2.1.1). Moreover, these evaluations must be repeated for *each* step of the Variational EM algorithm.

QB: (i) update of V_t c -times in the (4.39), and (ii) evaluation of the corresponding normalization constant (4.39). Computational complexity of normalization is $O(r+p)$.

VL: same as QB plus determination of maxima (4.46).

This operation can be done in parallel for each candidate transformation in all cases.

Update of V_t is done via dyadic updates (4.28), (4.42), and (4.47) for VB, QB, and VL respectively.

VB: LD update of V_t , c -times

QB: LD update of V_t , c -times

VL: one LD update of V_t .

This operation must be done sequentially.

The overall computational complexity is summarized in Table 4.1.

The main drawback of the VB algorithm is that the number of iterations, m , of the VEM algorithm at each step, t , is unknown *a priori*. For stationary processes, i.e. $\theta_t = \theta$, it can be expected that with growing number of data, the new data record d_t will cause just a small shift in the expected values of parameters. Hence, the VEM algorithm will converge fast and m may be as low as $m = 2$ or $m = 1$.

Remark 4.1 *The layout of the scheme (Figure 4.1) suggests a multiple model approach [94]. This similarity is not surprising, since the approximation used there is based on the principle of partitioning [95], which is equivalent to the conditional independence assumption (4.3) in this work. The MEAR scheme, with its restrictions (namely a fixed filter bank (4.7)), represents a special case of the interacting multiple model [96]. Specifically, the interactive multiple model updates the covariance matrix (corresponding to V_t) at time t with a vector, this vector being a combination (interaction) of candidate states. Hence, the covariance matrix is updated with a matrix of rank 1. This corresponds to the update of V_t by one dyad, as was the case for the Viterbi-like algorithm for the MEAR model. On the other hand, matrix V_t is updated in the VB and QB algorithms with a weighted sum of dyads. Hence, V_t is updated by a matrix of rank $\min(p+1, c)$.*

It is emphasized that the VB, QB and Viterbi-like algorithms are alternative strategies for collecting the approximate statistics $\bar{V}_t, \nu_t, \bar{\Phi}_t$ in (4.22). All subsequent operations, namely prediction and model structure determination, are determined by the form of the yielded posterior, which is the *same* in all three cases. Hence, these tasks can now be addressed in the following sections without reference to the chosen approximation.

4.6.2. MEAR-based Prediction

The MEAR predictor can be found by marginalization, using (4.13) (replacing t by $t + 1$), (4.22) and the chain rule:

$$f(\mathbf{d}_{t+1}|Y_{t+1}, G) = \int_{A, \Omega, \mathbf{l}_t, \mathbf{l}_{t+1}, T} f(\mathbf{d}_{t+1}, A, \Omega, \mathbf{l}_{t+1}, T, \mathbf{l}_t|Y_{t+1}, G) da d\sigma d\alpha \quad (4.50)$$

which is a task similar to (4.36). The predictor is found as

$$f(\mathbf{d}_{t+1}|Y_{t+1}, G) = \sum_{i=1}^c \alpha_{i,t} f(\mathbf{d}_{t+1}|Y_t, \mathbf{g}_i), \quad (4.51)$$

which is a mixture of EAR predictors (3.30), weighted by the respective component weights $\alpha_{i,t}$:

$$\alpha_{i,t} = \sum_{j=1}^c w_{j;t} \frac{\zeta_{\mathcal{D}_i}(\phi_{i,j;t-1} + 1)}{\zeta_{\mathcal{D}_i}(\phi_{i,j;t-1})}.$$

In typical signal processing applications, only moments of these distributions are of interest:

$$\hat{\mathbf{d}}_{t+1} = \sum_{i=1}^c \alpha_{i,t} \hat{\mathbf{d}}_{i,t+1},$$

where $\hat{\mathbf{d}}_{i,t+1}$ is the prediction of each candidate, in special case it is given by (3.32), (3.33). Note, in general, that all non-central moments of (4.51) can be obtained as this weighted algebraic mean of the respective non-central moments of the candidates. However, this does not hold for the central moments [97].

4.6.3. MEAR Model with Non-stationary Parameters

In Section 3.2.3, we relaxed the assumption of stationarity of parameters A, Ω by means of a forgetting operator (3.39). The same can be done for the MEAR model with parameters A, Ω, α , since A, Ω has the same distribution as in the AR model and T has Dirichlet distribution, which belongs to the exponential family. Distributions on A, Ω and T were chosen conditionally independent (Section 4.3). Hence, we choose distinct forgetting factors, $\phi_{\mathcal{N}\mathcal{W}}$ and $\phi_{\mathcal{D}_i}$, respectively. The prior at time t is then chosen as

$$f(A_t, \Omega_t, \alpha_t|Y_t) = \left[f(A_{t-1}, \Omega_{t-1}|Y_{t-1})_{A_t, \Omega_t} \right]^{\phi_{\mathcal{N}\mathcal{W}}} \left[\bar{f}(A_t, \Omega_t|Y_t)^{1-\phi_{\mathcal{N}\mathcal{W}}} \right] \times \left[f(T_{t-1}|Y_{t-1})_{T_t} \right]^{\phi_{\mathcal{D}_i}} \left[\bar{f}(T_t|Y_t)^{1-\phi_{\mathcal{D}_i}} \right]. \quad (4.52)$$

Where $\bar{f}(A_t, \Omega_t | Y_t)$ and $\bar{f}(\alpha_t | Y_t)$ are alternative distributions chosen by the designer.

Replacing prior (4.22) by (4.52), and choosing $\bar{f}(A_t, \Omega_t | Y_t) = \mathcal{N}\mathcal{W}(\bar{V}, \bar{v})$, $\bar{f}(T_t | Y_t) = \mathcal{D}i(\bar{\Phi})$, stationary for all t , then all foregoing algorithms maintain their structure, with the following modification of the statistics update mechanism:

$$V_t = \phi_{\mathcal{N}\mathcal{W}} V_{t-1} + \sum_{i=1}^c \hat{l}_{i,t} \mathbf{y}_{i,t} \mathbf{y}'_{i,t} + (1 - \phi_{\mathcal{N}\mathcal{W}}) \bar{V}_t, \quad (4.53)$$

$$\nu_t = \phi_{\mathcal{N}\mathcal{W}} \nu_{t-1} + 1 + (1 - \phi_{\mathcal{N}\mathcal{W}}) \bar{v}_t, \quad (4.54)$$

$$\Phi_t = \phi_{\mathcal{D}i} \Phi_{t-1} + \hat{l}_t \hat{l}'_{t-1} + (1 - \phi_{\mathcal{D}i}) \bar{\Phi}_t. \quad (4.55)$$

This is the form for VB and QB variants. The required modification of the Viterbi-Like (VL) variant is of the same kind.

4.6.4. MEAR Model Structure Determination

The key restriction of the MEAR model—namely, common AR parameters A, Ω (4.13)—implies that all filter candidates, $\mathbf{g}_i \in G$, must have the same dimension, $p + r$. The estimation of the MEAR model does not provide inference of the model structure and additional treatment is required.

The likelihood of the whole data set, D_t , can be obtained from the one-step-ahead predictor (4.51), using the chain rule:

$$f(D_{t \setminus q} | D_q, G(q)) \propto \prod_{j=q+1}^t f(\mathbf{d}_j | Y_{j-1}, G(q)). \quad (4.56)$$

From Bayes' rule, we can evaluate the inference on $G(q)$ as:

$$f(G(q) | D_t) \propto f(D_{t \setminus q} | D_q, G(q)) f(G(q)). \quad (4.57)$$

Note that the one-step-ahead predictor (4.50) is based on the expected values of the label, \hat{l}_t . One of the immediate consequences is that the trajectory of \hat{l}_t with respect to t must be recalculated for each setting of $G(q)$.

4.7. Inference of an AR Model Robust to Outliers

One of the main limitations of the AR model is sensitivity of the estimates to outliers in measurements. In this section, we analyse the problem of estimation of a scalar ($p = 1$) AR process (3.9) of order $q = r$, with observations corrupted by isolated outliers. An isolated outlier is not modelled by the AR model because the outlier-affected observed value does not take part in the future regression. Instead the process is autoregressive in *internal* (i.e. not directly measured) variable z_t , i.e.

$$z_t = A \mathbf{x}_t + \omega^{-\frac{1}{2}} e_t, \quad \mathbf{x}_t = [z_{t-1}, \dots, z_{t-q}]', \quad (4.58)$$

which is observed via

$$d_t = z_t + \xi_t. \quad (4.59)$$

Here, ξ_t denotes a possible outlier at time t . Moreover, for an *isolated* outlier it holds that

$$\Pr(\xi_{t \pm i} = 0 | \xi_t \neq 0) = 1, \quad i = 1, \dots, q. \quad (4.60)$$

The AR model is identified via $f(A, \omega | D_t)$ (3.12) (i.e. *not* via $f(A, \omega | Z_t)$) and so the outlier degrades estimation iff it enters the extended regressor $\mathbf{y}_t = [z_t, \dots, z_{t-q}]$ (3.17).

4.7.1. Filter-Bank Design

Since \mathbf{y}_t is of finite length, and since the outliers are isolated, it is easy to define a finite number of mutually exclusive scenarios. Each of these scenarios can be captured via an EAR model and combined together using the MEAR approach, as follows:

1. None of the values in \mathbf{y}_t is affected by an outlier, i.e. $d_{t-i} = z_{t-i}, i = 0, \dots, q$. The set of transformations (3.26), (4.2) is then the singleton set:

$$\mathcal{G}_1 = \{g | \mathbf{y}_t = [d_t, d_{t-1}, \dots, d_{t-q}]'\}. \quad (4.61)$$

2. The observed value, d_t , is affected by an outlier, and so all delayed values are unaffected, given assumption (4.60); i.e. $d_{t-i} = z_{t-i}, i = 1, \dots, q$. For convenience, realization of ξ_t can be expressed as a multiple of realized value e_t , via unknown multiplier $h_t > 0$:

$$\xi_t = h_t \omega^{-\frac{1}{2}} e_t. \quad (4.62)$$

Then (4.58), (4.59) can be rewritten as:

$$d_t = A\mathbf{x}_t + (1 + h_t) \omega^{-\frac{1}{2}} e_t, \quad \mathbf{x}_t = [d_{t-1}, \dots, d_{t-q}]'.$$

This can now be transformed to the form

$$\frac{1}{h_t + 1} d_t = \frac{1}{h_t + 1} A\mathbf{x}_t + \omega^{-\frac{1}{2}} e_t.$$

Therefore, the space of transformations in this case is

$$\mathcal{G}_2 = \left\{ g \mid \mathbf{y}_t = \frac{1}{h_t + 1} [d_t, \mathbf{x}_t']', \quad 1 < h_t < \infty \right\}, \quad (4.63)$$

parameterized by variable h_t . The Jacobian of all $g \in \mathcal{G}_2$ is $J_2 = \frac{1}{h_t + 1}$.

3. The observed value is not affected by an outlier, but a k -steps-delayed observation, $d_{t-k}, k \in \{1, \dots, q\}$, is. In this case, the transformation should replace this value by an appropriate AR

process estimate, \hat{z}_{t-k} . The set of transformations for each $k = 1, \dots, q$ is then:

$$\mathcal{G}_{k+2} = \{g | y_t = [d_t, \dots, d_{t-q}]' + \delta_{q+1}(k+1)(\hat{z}_{t-k} - d_{t-k}), \forall \hat{z}_{t-k} = \rho_k(D_t)\}, \quad (4.64)$$

where $\delta_p(i) = [\delta(1-i), \dots, \delta(p-i)]$. The elements of \mathcal{G}_{k+2} are indexed by every possible function, ρ_k , denoting an estimator of unobserved quantity z_{t-k} from data D_t . The Jacobian of all possible transformations $g \in \mathcal{G}_{k+2}$ is $J_{k+2} = 1$.

We have described $c = q + 2$ different modes (partitions), \mathcal{G}_i , of an ideal filter transforming observation process, d_t , to a process, y_t , for which an AR model is valid. For each of these partitions, a representative candidate, g_i , should be chosen (4.3). The choice of candidate $g_1 = \mathcal{G}_1$ is trivial. Choosing g_2 is equivalent to choosing a known fixed $h_t = h$. Alternatively, if the variance of outliers is known to vary significantly, we can split \mathcal{G}_2 into finer subsets and choose u candidates with fixed values $h_1 < h_2 < \dots < h_u$. The transformation sets, \mathcal{G}_{k+2} , must each be represented by a function ρ_k defined with respect to a known reconstruction (smoothing) filter. We have tested the algorithm for ρ_k chosen as the k -steps delayed values \hat{z}_{t-k} of the expected value \hat{z}_t at time t , given D_t :

$$\hat{z}_t = \sum_{i=1}^c E_{z_t | l_t = e_i}(z_t) f(l_t = e_i | D_t, G). \quad (4.65)$$

Using (3.33), expected value (4.31) of the Multinomial distribution (4.26), and the fact that $z_t = d_t$, for all $i = 1, 3, \dots, c$, the reconstruction filter is

$$\hat{z}_t = d_t \sum_{j=1, j \neq 2}^c w_{j;t} + w_{2;t} \hat{A}_t \mathbf{x}_{t-k}. \quad (4.66)$$

Here, the Bayesian predictor, (3.33) has been used to replace the outlier arising in scenario 2 above.

In our simulations, the interpolation strategy chosen is $\hat{z}_{t-k} = \rho_k(D_t) = \rho_k(D_{t-k})$ (i.e. filtering). This choice requires only one calculation of \hat{z}_{t-k} for each t . Adopting the non-causal (smoothing) choice, $\hat{z}_{t-k} = \rho_k(D_t)$, would require q calculations of \hat{z}_{t-k} for each t , with, presumably, negligible benefit over the causal (filtering) choice.

4.7.2. Simulation Study

A second-order stable AR model with parameters $A = [1.85, -0.95]'$, $\omega = 100$, was simulated with a random outlier on every 30th sample. The total number of samples was $n = 200$. A segment of the simulated data ($t = 55, \dots, 100$) is displayed in Figure 4.2 (dotted line) along with the reconstruction (solid line) (4.66), and corrupted data (dots). Two outliers occurred during the displayed period: a ‘small’ outlier at $t = 60$ and ‘big’ outlier at $t = 90$. The MEAR model with four candidate transformations— \mathcal{G}_1 (4.61), \mathcal{G}_2 (4.63) with $h_t = h = 10$, \mathcal{G}_3 and \mathcal{G}_4 (4.64) as derived in the previous Section—was used for identification of the AR parameters A, ω . The prior distribution was chosen as $\mathcal{NW}(V_0, \nu_0)$ with

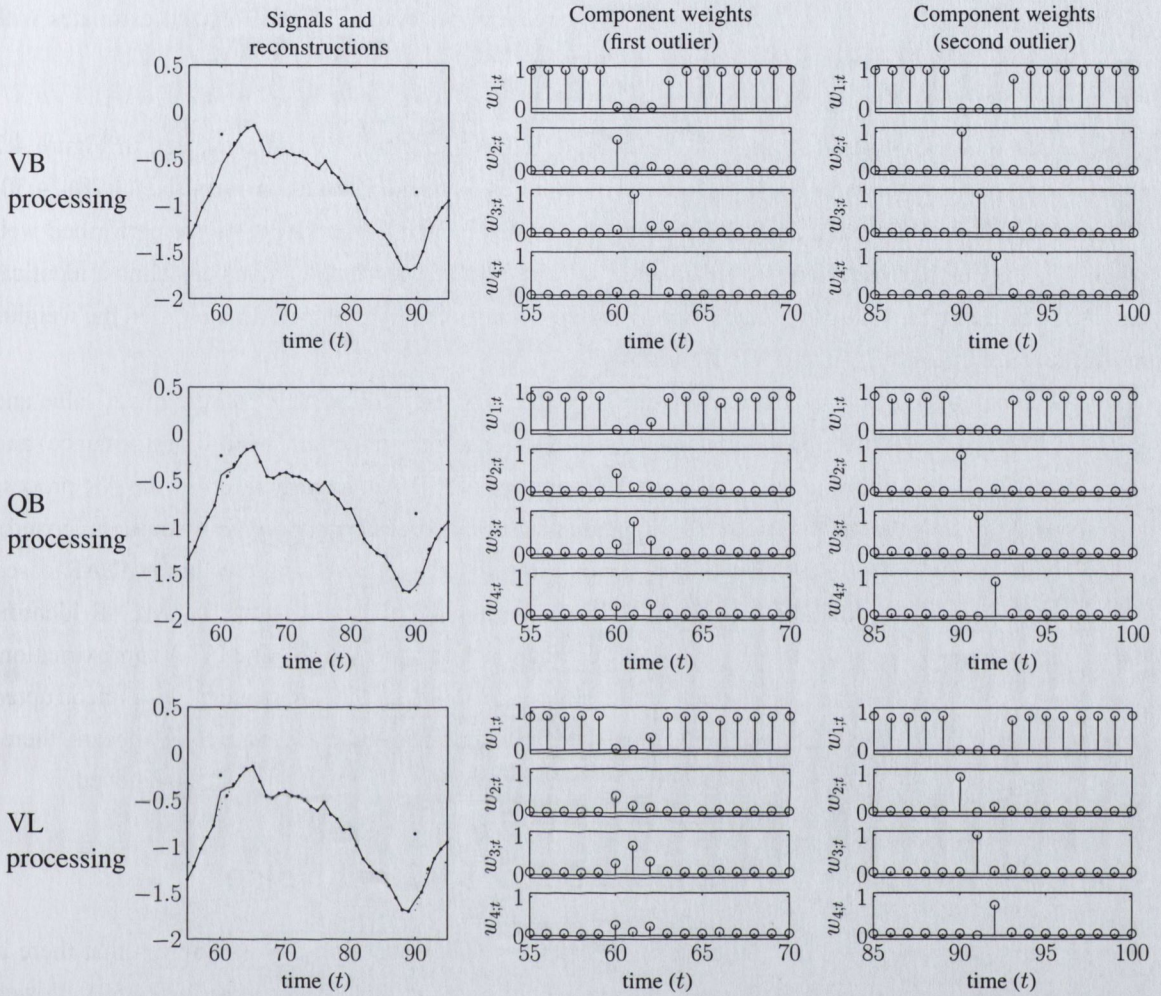


Figure 4.2.: Reconstruction of an AR(2) process corrupted by isolated outliers. Results for VB, QB, and VL algorithms, respectively, are shown. There are outliers at $t = 60$ and $t = 90$.

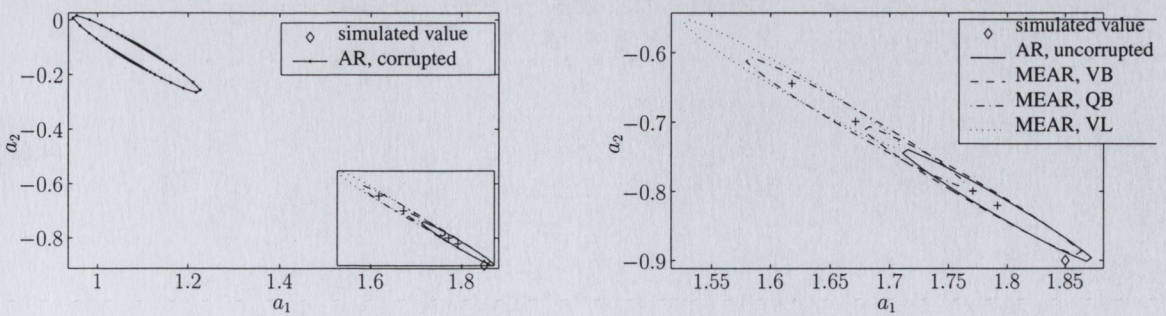


Figure 4.3.: Identification of an AR(2) process corrupted by isolated outliers. **Left**: comparison of the terminal moments of the posterior distribution of A ; **Right**: detail of left, boxed region.

$$V_0 = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.001 & 0 \\ 0 & 0 & 0.001 \end{bmatrix}$$
 and $\nu_0 = 1$. This choice of prior corresponds to point estimates with $\hat{A} = [0, 0]$ (3.20), and $\hat{\omega} = 10$ (3.21).

Note that when an outlier occurs, all candidate filters are sequentially used, as seen in Figure 4.2 (middle and right columns). Thus, the outlier is removed from the estimation formulae (4.28)-(4.30) very effectively. We note that all considered algorithms—i.e. VB, QB, and VL—have performed well when the ‘big’ outlier occurred. The estimated weights and reconstructed values are almost identical across the procedures. However, when ‘small’ outlier occurred, the VB algorithm identified the weights more accurately than the QB and VL algorithms.

The terminal—i.e. $t = n$ —posterior distribution of A , (A.10) is illustrated (via the mean value and 2 standard deviation ellipse) for the various identification methods in the left (overall performance) and right (detail) of Fig 4.3. In the left diagram, the scenarios are (i) AR identification of the AR process *corrupted* by outliers (boxed); (ii) AR identification of the AR process *uncorrupted* by outliers (boxed). In the right, we zoom in on the boxed area surrounding (ii) above, revealing the three MEAR-based identification scenarios: (iii) MEAR identification using the VB approximation; (iv) MEAR identification using the QB approximation; (v) MEAR estimation using the Viterbi-like (VL) approximation. Impressively, the MEAR-based strategies perform almost as well as the AR strategy with uncorrupted data, which is displayed via the full line. The posterior uncertainty in the estimate of A appears, therefore, to be due to the AR process itself, with all deleterious affects of the outlier process removed.

4.8. Inference of an AR Model Robust to Burst Noise

The previous example relied on outliers being isolated (4.60), permitting the assumption that there is only one outlier in the extended regressor y_t . In such a case, additive decomposition (4.59) allowed successful MEAR modelling of d_t via a finite number ($q + 2$) of candidates.

4.8.1. Filter-Bank Design for Burst Noise

A burst noise scenario, in contrast, requires more than one outlier to be considered in the regressor, obviating the filter bank design in the previous example. To address this problem, we need to transform the underlying scalar AR model (4.58) into state-space form [58]:

$$z_{t+1} = Bz_t + k\omega^{-\frac{1}{2}}e_t. \quad (4.67)$$

We choose the model with state variable assignment $\mathbf{z}_t = [d_t, \dots, d_{t-q}]'$. Therefore:

$$B = \begin{bmatrix} -a_1 & -a_2 & -a_3 & \cdots & -a_q \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{k} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (4.68)$$

where $B \in \mathbb{R}^{q \times q}$ and $\mathbf{r} \in \mathbb{R}^{q \times 1}$. The process with burst noise is modelled as

$$d_t = \mathbf{c}'\mathbf{z}_t + h_t\omega^{-\frac{1}{2}}\zeta_t, \quad (4.69)$$

where $\mathbf{c} = [1, 0, \dots, 0]' \in \mathbb{R}^{q \times 1}$, and ζ_t is $\mathcal{N}(0, 1)$, independent of e_t . $h_t\omega^{-\frac{1}{2}}$ denotes the standard deviation of the burst noise which is assumed strictly positive during any burst, and is zero otherwise. Note that the autoregressive part of the model (4.67, 4.68) is identical to the AR model in the previous example. The key difference is in the corruption process (4.69) compared to (4.59), (4.60). Once again, we identify a finite number of mutually exclusive—but now non-exhaustive—scenarios that can be modelled using an EAR process:

1. The current observation d_t and the last q observations, d_{t-1}, \dots, d_{t-q} are all distortion-free; i.e. $h_t = h_{t-1} = \dots = h_{t-q} = 0$. Formally, $\mathcal{G}_1 = \{\mathbf{g}|\mathbf{y}_t = \mathbf{z}_t\}$, a singleton set.
2. The measurements are all affected by burst noise; we set $h_t = h_{t-1} = \dots = h_{t-q} = h$. The state-space model (4.67, 4.69) is now defined by the joint distribution:

$$f(\mathbf{z}_t, d_t | \mathbf{a}, \omega, \mathbf{z}_{t-1}, h) = \mathcal{N} \left(\begin{bmatrix} B\mathbf{z}_{t-1} \\ \mathbf{c}'\mathbf{z}_t \end{bmatrix}, \omega^{-1} \begin{bmatrix} \mathbf{r}\mathbf{r}' & 0 \\ 0 & h^2 \end{bmatrix} \right). \quad (4.70)$$

(4.70) cannot be modelled directly as an EAR process because it contains unobserved state vector \mathbf{z}_t . Using standard Kalman-filter theory [16, 58], we can marginalize (4.70); i.e. we use the chain rule t times and integrate over the unobserved trajectory—namely over $\{\mathbf{z}_1 \dots \mathbf{z}_t\}$ —to obtain the direct observation model:

$$f(d_t | \mathbf{a}, \omega, D_{t-1}, h) = \mathcal{N}(\mathbf{a}\hat{\mathbf{z}}_t, \omega^{-1}\sigma_t) \quad (4.71)$$

with moments defined recursively as follows:

$$\sigma_t = h + \mathbf{c}'S_{t-1}\mathbf{c}, \quad (4.72)$$

$$\tilde{S}_t = S_{t-1} - \sigma_t^{-1}(S_{t-1}\mathbf{c})(S_{t-1}\mathbf{c})', \quad (4.73)$$

$$\hat{\mathbf{z}}_t = B\hat{\mathbf{z}}_{t-1} + h^{-1}\tilde{S}_t\mathbf{c}(d_t - \mathbf{c}'B\hat{\mathbf{z}}_{t-1}), \quad (4.74)$$

$$S_t = \mathbf{r}\mathbf{r}' + B\tilde{S}_tB'. \quad (4.75)$$

(4.71) can be expressed as a valid EAR model (3.29), if \hat{z}_t and σ_t are independent of the unknown AR parameters, A, ω . Unfortunately, both, \hat{z}_t and σ_t are functions of matrix B and thus of A . In order to obtain a valid EAR model, we replace $B = B(A)$ (4.68) in (4.74,4.75) by its expected value, $\hat{B}_t = B(\hat{A}_t)$, via (3.20). Then, (4.71) is a valid EAR model defined by the set of transformations:

$$\mathcal{G}_2 = \left\{ \mathbf{g} \mid \mathbf{y}_t = \frac{1}{\sigma_t} [d_t, \hat{z}_t]', 1 < h < \infty \right\}, \quad (4.76)$$

with time-variant Jacobian, $J_t = \sigma_t^{-1} (D_{t-1})$ evaluated, recursively using (4.72). The space \mathcal{G}_2 is parameterized by the unknown h .

3. Cases 1) and 2) do not consider the situation where h_k is not constant on a regression interval $k \in [t - q, t]$ (i.e. a transitional phase). Complete modelling for all such a cases is too difficult, and so these scenarios are ignored. However, our experiments suggest that this has little impact on performance.

The final step is to define candidates to represent \mathcal{G}_2 ($\mathbf{g}_1 = \mathcal{G}_1$ is trivial). One candidate may be chosen for \mathcal{G}_2 if the variance of burst noise is reasonably well known *a priori*. In other cases, we can partition \mathcal{G}_2 with respect to intervals of h . Candidates are chosen as one element from each interval. The best experimental results were achieved for multiple partitioning of \mathcal{G}_2 , with at least one candidate, $\mathbf{g}_l = \mathbf{g}(h = h_l)$ in (4.76), where h_l is *below* the true value of h , and at least one candidate, $\mathbf{g}_u = \mathbf{g}(h = h_u)$ in (4.76) where h_u is *above* the true value of h . h_l and h_u can conveniently be chosen as the prior lower and upper bounds, respectively, on h . Increasing the number of candidates drawn from \mathcal{G}_2 generates a richer set, G , which better spans the subset \mathcal{G}_2 . This improves the quality of approximation.

4.8.2. Simulation Study

A *non-stationary* AR(2) process was studied, with $a_{1;t}$ in the interval $[-0.98, -1.8]$ (as displayed in Figure 4.4 (top-right)), $a_{2;t} = a_2 = 0.98$, $\omega_t = \omega = 100$, and $n = 200$. Realizations are displayed in Figure 4.4 (top-left, solid line). For $t < 95$, $a_{1;t}$ is increasing, corresponding to faster signal variations. Thereafter, $a_{1;t}$ decreases, yielding slower variations. These variations of $a_{1;t}$ do not influence the absolute value of the complex poles of the system, but only their polar angle. The process was corrupted by two noise bursts (samples 50–80 and 130–180), with parameters $h = 8$ and $h = 6$ respectively (4.69). Realizations of the burst noise process imposed on the simulated signal are displayed in Figure 4.4 (top-left, dotted line).

The process was estimated using $c = 3$ filter candidates: namely the unity transformation, G_1 , along with $G_2(h = 5)$ and $G_2(h = 10)$. Identification results, are displayed in the right column in Figure 4.4 as follows: (i) simulated data, (ii) AR model for *uncorrupted* data, (iii) VB variant of the MEAR model for *corrupted* data, (iv) QB variant of the MEAR model, (v) VL variant of the MEAR model. Specifically, the 95% Highest Posterior Density (HPD) interval, via (A.13) and (A.15), of the marginal Student t -distribution of $a_{1;t}$ and $a_{2;t}$ respectively, is displayed. The process was identified using forgetting factors (4.52) $\phi_{\mathcal{N}\mathcal{W}} = 0.92$, $\phi_{\mathcal{D}i} = 0.9$, and non-committal, stationary, alternative $\mathcal{N}\mathcal{W}$ distribution, $\bar{f}(A, \omega) = \mathcal{N}\mathcal{W}_{A, \Omega}(\bar{V}, \bar{v})$. Furthermore, the matrix parameter, $\bar{\Phi}$, of the stationary, alternative $\mathcal{D}i$

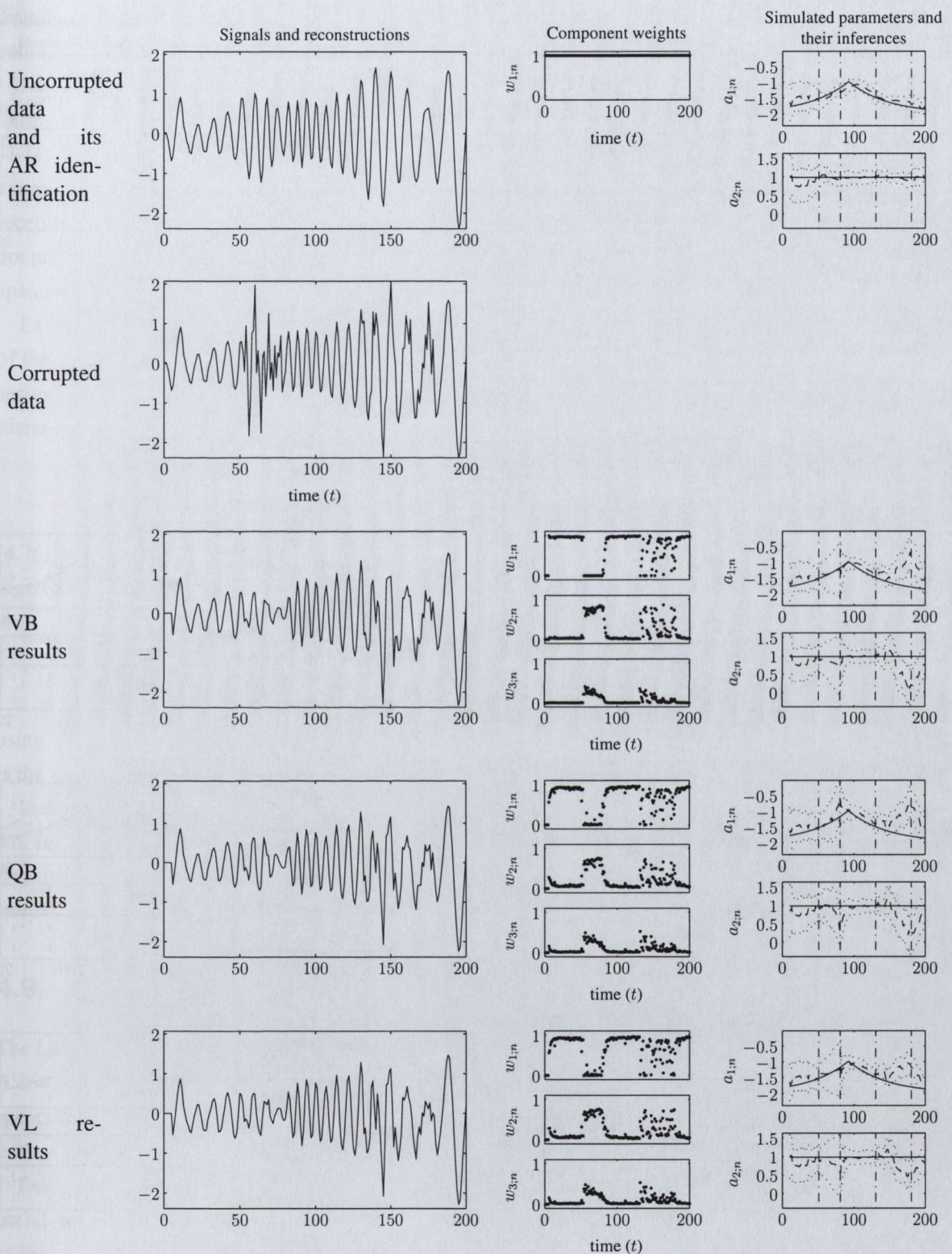


Figure 4.4.: Reconstruction and identification of a non-stationary AR(2) process corrupted by burst noise, using KF variant of the MEAR model. In the final column, full lines denote simulated values of parameters, dashed lines denote posterior expected values, and dotted lines denote uncertainty bounds.

4. MIXTURE-BASED EXTENSION OF THE EAR (MEAR) MODEL

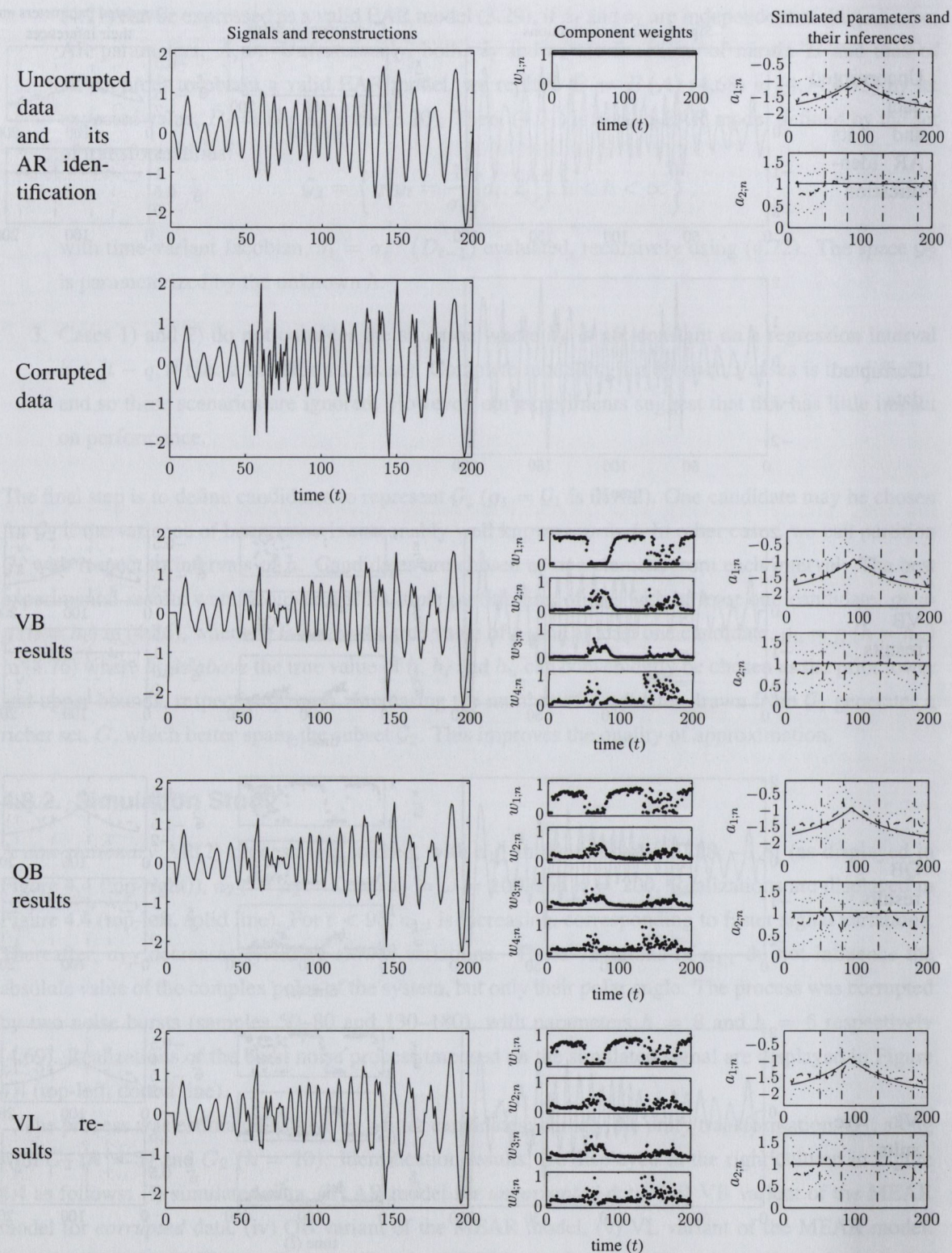


Figure 4.5.: Reconstruction and identification of a non-stationary AR(2) process corrupted by burst noise, using KF+LPF variant of the MEAR model. In the final column, full lines denote simulated values of parameters, dashed lines denote posterior expected values, and dotted lines denote uncertainty bounds.

distribution, $\bar{f}(T)$ (4.52), was chosen to be diagonally dominant with ones on the diagonal. This discourages frequent transitions between filters.

Note that all methods (VB, QB and VL) achieved robust identification of the process parameters during the first burst. As already noted, $\hat{z}_{i,t}$, $i = 2, \dots, c$, (which denotes the reconstructed state vector (4.74) with respect to the i th filter), is correlated with \hat{A}_{t-1} , which may undermine the tracking of time-varying AR parameters, A_t . In this case, each Kalman component predicts observations poorly, and receives low weights, $w_{2,t}$ and $w_{3,t}$ (4.31), in (4.53). This means that the first component—which does not pre-process the data—has a significant weight, $w_{1,t}$. Clearly then, the Kalman components have not spanned the space of necessary pre-processing transformations well, and need to be supplemented.

Extra filters can be ‘plugged in’ in a naïve manner (in the sense that they *may* improve the spanning of the pre-processing space, but should simply be rejected, via (4.31), if poorly designed). During the second burst (Figure 4.4), the process is slowing down. Therefore, we have extended the bank of KF filters by a simple arithmetic mean Low-Pass Filter (LPF) on the observed regressors:

$$G_3 : \mathbf{y}_n = \frac{1}{3} (\mathbf{x}_n + \mathbf{x}_{n-1} + \mathbf{x}_{n-2}). \quad (4.77)$$

(4.76) and (4.77) yield EAR models with the same AR parameterization, and so they can be used together in the MEAR filterbank. Reconstructed values for the KF variant are derived from (4.65):

$$\hat{z}_t = w_{1,t} d_t - \sum_{i=2}^3 w_{i,t} \hat{A}_t \hat{z}_{i,t}, \quad (4.78)$$

using (3.20). For the KF+LPF variant, the term $\frac{w_{4,t}}{3} (d_t + d_{t-1} + d_{t-2})$ is added to (4.78), where $w_{4,t}$ is the estimated weight of the LPF component (4.31), (4.53)–(4.55).

Identification and reconstruction of the process using the KF+LPF filter-bank is displayed in Figure 4.5, in the same layout as Figure 4.4. The distinction is most clearly seen in the final column of each. During the second burst, the added LPF filter received high weights, $w_{4,t}$, Figure 4.5 (middle column). Hence, identification of the parameter A is improved during the second burst.

4.9. Application of the MEAR model in Speech Reconstruction

The MEAR filter-bank for the burst noise case (KF variant) was applied in the reconstruction of speech. A $c = 4$ MEAR model was used, involving $G_1 (\mathbf{y}_t = \mathbf{x}_t)$, $G_2 (h = 3)$, $G_2 (h = 6)$, $G_2 (h = 10)$. The speech was modelled as AR with order $q = 8$ (3.9). The forgetting factors (4.52) were $\phi_{\mathcal{N}\mathcal{W}} = \phi_{D_i} = 0.95$. Once again, a diagonally-dominant $\bar{\Phi}$ was chosen for $\bar{f}(T)$.

During periods of silence in speech, statistics (4.53) are effectively not updated, creating difficulties for adaptive identification. Therefore, we use an *informative* stationary alternative distribution, $\bar{f}(A, \omega)$, of the $\mathcal{N}\mathcal{W}$ type (3.13) for the AR parameters in (4.52). To elicit an appropriate density, we identify the time-invariant alternative statistics, \bar{V} , \bar{v} , using 1800 samples of unvoiced speech. $\bar{f}(A, \omega)$ was then flattened to reduce \bar{v} from 1800 to 2. This choice moderately influences the accumulating statistics at each step, via (4.53). Specifically, after a long period of silence, the influence of data in (4.53) becomes negligible, and V_t is reduced to \bar{V} .

Three sections of the `bbcnews.wav` speech file, sampled at 11kHz, were corrupted by additive noise. Since we are particularly interested in performance in non-stationary epochs, we have considered three transitional cases: (i) voiced-to-unvoiced transition corrupted by zero-mean, white, Gaussian noise, with a realized Signal-to-Noise Ratio (SNR) of -1 dB during the burst; (ii) an unvoiced-to-voiced transition corrupted by zero-mean white uniform noise at -2 dB; and (iii) a silence-to-unvoiced transition corrupted by a click of type $0.25 \cos(3t) \exp(-0.3t)$, superimposed on the silence period.

Reconstructed values using VB, QB and VL methods respectively are displayed in Figure 4.6. All three methods successfully suppressed the burst in the first two cases. In the third case, the click was suppressed by all methods. However, the QB and VL methods also had the deteriorious effect of suppressing the unvoiced speech.

4.10. Discussion

The MEAR model (4.13) proposes a relatively rich extension of the classical AR model. It allows transformations on regressors, which relates it to semi-physical modelling [98]. Being a mixture-based extension, it is also related to the multiple model approach [94], to mixtures of AR processes [99], and to the Generalized AR (i.e. GAR) approach [46]. It must be remembered, though, that the MEAR model is a *single* AR model subject to an unknown transformation of observations. This is formalized as a mixture with *common* AR parameters (4.13). There are two main consequences. Firstly, the MEAR model is appropriate in cases where the transformation/distortion process is independent of the underlying AR process. Secondly, the AR parameter inference (4.24) requires a *single* sufficient statistic matrix, V_n (4.28), updated via a linear combination of c dyads, each calculated from one component in turn.

The restriction to common AR parameterization across all components can easily be relaxed via obvious changes to the recursive algorithm (4.28)–(4.30). Each AR component would then experience a local rank-1 update, and there would be no inter-component interaction. Such a model would be over-parameterized, as each component would then have unknown AR parameters *and* an unknown transformation g_i , causing identification problems. The common AR parameterization in the MEAR model overcomes this problem. It can be seen as a model-based regularization.

We have derived three variants of the identification algorithm: (i) Variational Bayes (VB), (ii) Quasi-Bayes (QB), and (iii) Viterbi-Like (VL). The VB algorithm is the optimal in the sense of KL minimization (Section 2.2.4), while the QB and VL variants are computationally simpler methods derived as approximations of the VB solution. All variants yield acceptable solutions in particular contexts, as demonstrated in Section 4.9.

The statistics, V_t , are updated by a structure of rank c in the VB and QB variants as stated already. This implies an interaction of regressors from each component, which appears to be a key benefit of the MEAR model, since it allows a small number of candidate models to span a larger transformation space. The concept of interaction between a finite set of components has been exploited in other techniques. The Kalman-based Interacting Multiple Models (IMMs) [94] linearly combine state vectors (i.e. certainty equivalents) evaluated using each filter, before using it in the Kalman updates. Again, however, this corresponds to a rank-1 update in our framework as in the VL variant (Section 4.6.1).

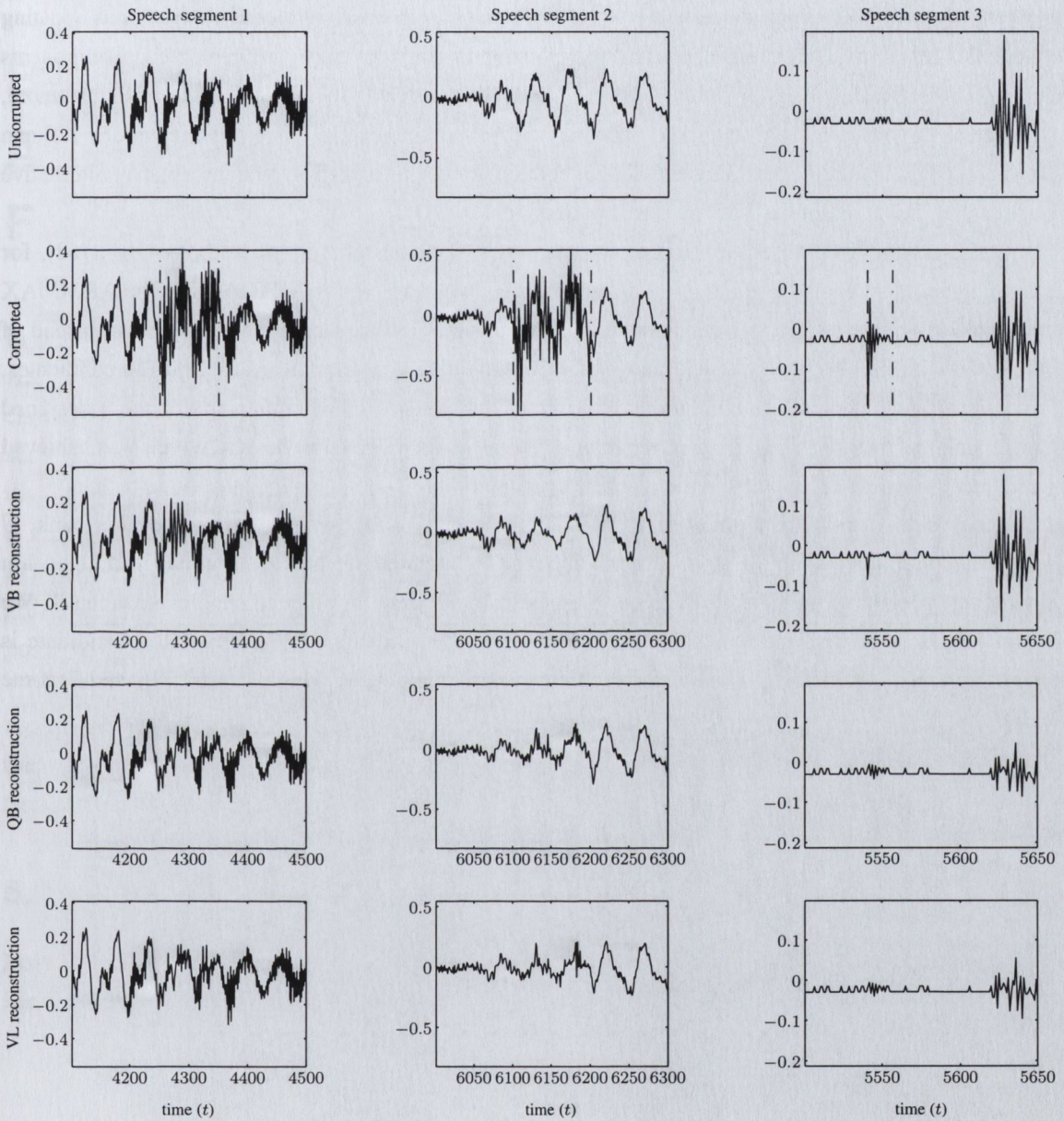


Figure 4.6.: Reconstruction of three sections of the `bbcnews.wav` speech file. In the second row, dash-dotted vertical lines delimit the beginning and end of each burst.

Bayesian identification unifies all tasks of inference into a single, model-consistent framework. In the burst noise example of Section 4.8, the MEAR algorithm combines the pre-processing tasks (of burst detection and signal reconstruction) with on-line identification. It is the dynamic weights (4.31) which balance the dyadic update contributed by each component at every step (4.28). This contrasts with the previously reported methods. For example, in [100], a Boolean detection decision is made concerning presence of outliers. During a detected burst, a Kalman filter is used for reconstruction, and updating of statistics is interrupted. In our approach, the updating of statistics is *never* interrupted. Components which, in effect, pre-process noisy data, contribute dyads constructed from *filtered* data. Furthermore, exponential forgetting is used to handle time-varying AR parameters, in place of the extended Kalman filter in [100]. In difficult cases, such as silence regions of speech, forgetting with informative alternative distributions (3.39) might be used, as it was in Section 4.9.

A Quasi-Bayes (QB)-based approximate update of sufficient statistics was employed in [62], for estimating an ARMA model using a mixture-based extension (known as ARMMAX). The ARMMAX model is a special case of the MEAR model, but with time-invariant component weights, instead of (4.11), and with moving-average whitening filters as candidate transformations (3.26). The candidates, G (4.7), used to represent the continuous multidimensional space of whitening filters, were designed using a simplex method. This is an example of a technique for filter-bank design, which was achieved at the price of loss of recursivity in the identification method.

In our work, we model the possible degradations of the AR process, and design the filter-bank, G (4.7), in an attempt to span all possibilities. The parallel architecture of the summed-dyad algorithm (Figure 4.1) permits extra candidates to be ‘plugged in’ with ease, in order to supplement the set. We saw in Section 4.8, for instance, how this can improve identification. When the extra candidate is not relevant, its contributing dyads are weighted by low component weights in (4.28), and become negligible.

Chapter 5.

Bayesian Inference of Non-stationary AutoRegressive Models Using Time-variant Forgetting

In Section 3.2.3, we reviewed the Bayesian inference of non-stationary parameters of the AR model. Analytical solution is available under the assumption of known forgetting factor, ϕ . The value of ϕ is chosen by the designer and always represents a trade-off: higher ϕ gives lower variance of estimates in stationary scenarios, and lower ϕ provides better tracking ability during non-stationary epochs. Intuitively, we would like to develop ‘smart’ forgetting, one which keeps ϕ high when the identified model is in agreement with the observed data, and which decreases ϕ when incoming data do not correspond to this model. This idea was studied in the context of window-based processing [101], and using a gradient-based MAP approach [102]. The min-max criterion approach of adaptive forgetting was proposed in [103]. The idea has also been studied in the context of Recursive Least Squares (RLS) [104].

In this Chapter, we seek a joint Bayesian inference of the non-stationary AR parameters of the multivariate AR model in tandem with the time-variant forgetting factor. The method will be extended to the MEAR model in Section 5.3. Progress in these areas is made possible via the VB-approximation of Section 2.2.4.

5.1. Bayesian Formulation

Following the Bayesian methodology, we treat uncertain ϕ_t as a random variable. We seek a joint identification of both θ_t and ϕ_t . From (3.11), (3.39) the joint posterior distribution is then

$$f(\theta_t, \phi_t | Y_t) \propto f(\mathbf{d}_t | \theta_t, Y_{t-1}, \mathbf{x}_t) f(\theta_t | Y_{t-1}, \phi_t) f(\phi_t | Y_{t-1}), \quad (5.1)$$

where the prior on ϕ_t is uniform, $\forall t$, in the interval $0 \leq \phi_t \leq 1$:

$$f(\phi_t | Y_{t-1}) = \mathcal{U}([0, 1]). \quad (5.2)$$

Note that (5.1) is conditionally independent of the previous parameters θ_{t-1} and ϕ_{t-1} , given Y_t . This is achieved via the forgetting operator (3.39) and the choice of prior (5.2). Hence, the proliferation of new random variables is avoided. Computationally feasible recursive identification is achieved if the

posterior distribution

$$f(\theta_t, \phi_t | Y_t) = f(\theta_t | \phi_t, Y_t) f(\phi_t | Y_t), \quad (5.3)$$

is chosen as conjugate with the observation model (3.29). This is possible for known forgetting, choosing $f(\theta_t | \phi_t, Y_{t-1})$ from \mathcal{NW} family (Section 3.2.3). However, this cannot be achieved for the joint posterior (5.3).

Therefore, we seek an approximate posterior in conditionally independent form:

$$\check{f}(\theta_t, \phi_t | Y_t) = \check{f}(\theta_t | Y_t) \check{f}(\phi_t | Y_t). \quad (5.4)$$

5.2. Variational Bayes (VB) Approximation

The conditional independence (5.4) is the basic assumption of the Variational Bayes (VB) approximation method (Section 2.2.4). In order to achieve recursive identification, we demand that the posterior distribution on parameters at time $t-1$ be of the same form as that at time t . The functional optimization achieved by the VB approximation allows us to choose the posterior distribution to be conjugate with the VB-optimized observation model (Section 2.3.3).

5.2.1. VB-conjugate Prior

Assume that the distribution of model parameters at time $t-1$ is of the form (5.4). It is updated by the observation model (3.11) to yield a posterior distribution. Then, the logarithm of the joint distribution is:

$$\begin{aligned} \ln f(\theta_t, \phi_t, \mathbf{d}_t | Y_{t-1}, \mathbf{x}_t) &= \ln f(\mathbf{d}_t | \theta_t, Y_{t-1}, \mathbf{x}_t) + \phi_t \ln \check{f}(\theta_t | Y_{t-1}) + \\ &+ (1 - \phi_t) \ln \bar{f}(\theta_t | Y_{t-1}) - \ln \zeta(\phi_t). \end{aligned} \quad (5.5)$$

Here $\zeta(\phi_t)$ is the ‘wildcard’ for normalizing constant of the forgetting operator (3.39), which depends on the form of the optimized distribution $\check{f}(\theta_t | \cdot)$. Using (5.5) and (5.4) in Theorem 2.1, the VB-optimized form of (5.4) is found in the following form:

$$\begin{aligned} \tilde{f}(\theta_t | Y_t) &\propto \exp \left(\ln f(\mathbf{d}_t | \theta_t, Y_{t-1}, \mathbf{x}_t) + \hat{\phi}_t \ln \check{f}(\theta_t | Y_{t-1}) + \right. \\ &\quad \left. + (1 - \hat{\phi}_t) \ln \bar{f}(\theta_t | Y_{t-1}) \right), \\ &\propto f(\mathbf{d}_t | \theta_t, Y_{t-1}, \mathbf{x}_t) \check{f}(\theta_t | Y_{t-1})^{\hat{\phi}_t} \bar{f}(\theta_t | Y_{t-1})^{1 - \hat{\phi}_t}, \end{aligned} \quad (5.6)$$

$$\tilde{f}(\phi_t | Y_t) \propto \exp \left(\phi_t E_{\theta_t} \left(\ln \check{f}(\theta_t | Y_{t-1}) - \ln \bar{f}(\theta_t | Y_{t-1}) \right) - \ln \zeta(\phi_t) \right). \quad (5.7)$$

Note that the VB-approximate update (5.6) is in the form the standard forgetting for the AR model (Section 3.2.3), with the forgetting factor given by $\hat{\phi}_t = E_{\phi_t | Y_t}(\phi_t)$. Therefore, using results from

Section 3.2.3, a conjugate update is possible if both $\tilde{f}(\theta_t|Y_{t-1})$, and $\bar{f}(\theta_t|Y_{t-1})$, are of the \mathcal{NW} form:

$$\tilde{f}(\theta_t|Y_{t-1}) = \mathcal{NW}(V_{t-1}, \nu_{t-1}), \quad (5.8)$$

$$\bar{f}(\theta_t|Y_{t-1}) = \mathcal{NW}(\bar{V}_{t-1}, \bar{\nu}_{t-1}). \quad (5.9)$$

5.2.2. VB-optimal Posterior Distribution

For the choice of priors (5.8) and (5.9), we can evaluate the ‘wildcard’ normalizing constant of the forgetting operator (3.39), as

$$\begin{aligned} \zeta(\phi_t) = \zeta_{\mathcal{NW}}(V(\phi_t), \nu(\phi_t)) &= \Gamma_p \left(\frac{1}{2} (\nu(\phi_t) - r + p + 1) \right) |\Lambda(\phi_t)|^{-\frac{1}{2}(\nu-r+p+1)} \\ &\quad |V_{\mathbf{aa}}(\phi_t)|^{-0.5p} 2^{0.5p(\nu(\phi_t)+p+1)} \pi^{\frac{r}{2}}. \end{aligned} \quad (5.10)$$

$$V(\phi_t) = \phi_t V_{t-1} + (1 - \phi_t) \bar{V}_{t-1}, \quad (5.11)$$

$$\nu(\phi_t) = \phi_t \nu_{t-1} + (1 - \phi_t) \bar{\nu}_{t-1}. \quad (5.12)$$

Equation (5.10) defines a complicated function in ϕ_t . Moreover, $\zeta(\phi_t)$ determines the approximate posterior distribution $\tilde{f}(\phi_t|Y_t)$ via its logarithm in (5.7). No standard distribution of this form is known to us. Moreover, evaluation of moments of $\tilde{f}(\phi_t|Y_t)$, involving (5.10), would be numerically intractable. Therefore, we seek an approximation of $\zeta(\phi_t)$. We take advantage of the fact that it is computationally simple to evaluate the normalizing coefficient of the \mathcal{NW} distribution $\zeta_{\mathcal{NW}}(\cdot)$ using LD decompositions (Section 3.2.1.1). Hence, we evaluate (5.10) at the extrema of its support:

$$\zeta(0) = \zeta_{\mathcal{NW}}(\bar{V}, \bar{\nu}), \quad (5.13)$$

$$\zeta(1) = \zeta_{\mathcal{NW}}(V_{t-1}, \nu_{t-1}). \quad (5.14)$$

Using these, we will now approximate (5.10) by interpolation between $\zeta(0)$ and $\zeta(1)$.

Proposition 5.1 (Approximate Normalization of the Forgetting Operator) *Let us choose the approximation of (5.10) in the following form:*

$$\zeta(\phi_t) = \exp(h_1 + h_2 \phi_t), \quad (5.15)$$

where h_1 and h_2 are unknown constants. Matching (5.15) at extrema (5.13), and (5.14) we obtain:

$$h_1 = \ln \zeta_{\mathcal{NW}}(\bar{V}_{t-1}, \bar{\nu}_{t-1}), \quad (5.16)$$

$$h_2 = \ln \zeta_{\mathcal{NW}}(V_{t-1}, \nu_{t-1}) - \ln \zeta_{\mathcal{NW}}(\bar{V}, \bar{\nu}). \quad (5.17)$$

Under this proposition, and using priors (5.8), (5.9), the joint log-distribution (5.5) is then approximated by

$$\begin{aligned} \ln f(\theta_t, \phi_t, \mathbf{d}_t|Y_{t-1}, \mathbf{x}_t) &\approx \ln \mathcal{N}(A\mathbf{x}_t, \Omega^{-1}) + \phi_t \mathcal{NW}_{A,\Omega}(V_{t-1}, \nu_{t-1}) + \\ &\quad + (1 - \phi_t) \mathcal{NW}_{A,\Omega}(\bar{V}_{t-1}, \bar{\nu}_{t-1}) + h_1 + h_2 \phi_t, \end{aligned} \quad (5.18)$$

where h_1 and h_2 are given by (5.16) and (5.17) respectively. Note that the choice (5.15) ensures that (5.18) is linear in ϕ_t .

Corollary 5.1 (Corollary 3 of Theorem 2.1, Variational Extreme for time-variant forgetting) *Using (5.4) and (5.18) in Theorem 2.1, the VB-optimal form of (5.4) is found via the following assignments:*

$$\tilde{f}(\theta_t|Y_t) = \mathcal{NW}(V_t, \nu_t), \quad (5.19)$$

$$\tilde{f}(\phi_t|Y_t) \approx \text{tExp}(b, [0, 1]), \quad (5.20)$$

with VB-statistics:

$$V_t = \hat{\phi}_t V_{t-1} + \mathbf{y}_t \mathbf{y}_t' + (1 - \hat{\phi}_t) \bar{V}, \quad (5.21)$$

$$\nu_t = \hat{\phi}_t \nu_{t-1} + 1 + (1 - \hat{\phi}_t) \bar{\nu}, \quad (5.22)$$

$$\begin{aligned} b = & -\frac{1}{2}(\nu_{t-1} - \bar{\nu}) \widehat{\ln|\Omega_t|} - \frac{1}{2} p \text{tr}((V_{aa;t-1} - \bar{V}_{aa}) V_{aa;t}^{-1}) \\ & - \ln \zeta_{\mathcal{NW}}(V_{t-1}, \nu_{t-1}) + \ln \zeta_{\mathcal{NW}}(\bar{V}, \bar{\nu}) \\ & - \frac{1}{2} \text{tr} \left((V_{t-1} - \bar{V}) \begin{bmatrix} -I_p, \hat{A}_t \end{bmatrix}' \hat{\Omega}_t \begin{bmatrix} -I_p, \hat{A}_t \end{bmatrix} \right). \end{aligned} \quad (5.23)$$

The required moments of the of the matrix Normal distribution (5.19), $\hat{A}_t, \hat{\Omega}_t, \widehat{\ln|\Omega_t|}$, and the first moment of the truncated Exponential distribution (5.20), $\hat{\phi}_t$, are given in Appendix A.1, and Appendix A.7 respectively.

Proof: (5.19), and the VB-statistics (5.21) and (5.22), follow from (5.6), using conjugacy of Normal distribution with Normal-Wishart, and closure of \mathcal{NW} distributions under geometric mean.

(5.20) follows from (5.7) evaluating the expected value in there, using (5.8), and (5.9)

$$\begin{aligned} E_{\theta_t} \left(\ln \tilde{f}(\theta_t|Y_{t-1}) - \ln \bar{f}(\theta_t|Y_{t-1}) \right) &= -\frac{1}{2}(\nu_{t-1} - \bar{\nu}) \widehat{\ln|\Omega_t|} \\ &\quad - \frac{1}{2} E_{\theta_t} \left(\text{tr} \left((V_{t-1} - \bar{V}) \begin{bmatrix} -I_p, \hat{A}_t \end{bmatrix}' \Omega_t \begin{bmatrix} -I_p, \hat{A}_t \end{bmatrix} \right) \right), \\ &= -\frac{1}{2}(\nu_{t-1} - \bar{\nu}) \widehat{\ln|\Omega_t|} - \frac{1}{2} p \text{tr} \left((V_{aa;t-1} - \bar{V}_{aa}) V_{aa;t}^{-1} \right) \\ &\quad - \frac{1}{2} E_{\Omega_t} \left(\text{tr} \left((V_{t-1} - \bar{V}) \begin{bmatrix} -I_p, \hat{A}_t \end{bmatrix}' \Omega_t \begin{bmatrix} -I_p, \hat{A}_t \end{bmatrix} \right) \right) \end{aligned} \quad (5.24)$$

We have used elementary properties of the trace operator, and property (A.2) of the Matrix Normal distribution were used. Moments $\hat{A}_t, \hat{\Omega}_t$ are given by (A.13), and (A.14) respectively. Truncation of the VB-posterior of ϕ_t to interval $[0, 1]$ follows from prior restriction (5.2). ■

The Variational Extreme (5.19), (5.20) can be found by iterating the implicit set of functions (5.21)–(5.23) to convergence via the VEM algorithm (Algorithm 2.2).

Remark 5.1 (Numerical Simplification) *In this case, it is difficult to find a numerical simplification of the VEM algorithm. The Restricted VB (Corollary 2.1) can be used. However, we cannot use the Quasi-Bayes principle (Remark 2.4), because exact marginal distributions are not tractable. Hence,*

instead we use a threshold for the number of iterations of the VEM algorithm. This was proposed in [35], as on-line VB for models with time-invariant parameters, where the number of iterations can be restricted to one. In our case, i.e. model with non-stationary parameters, one iteration per time step is equivalent to standard forgetting with an initial guess for the forgetting factor. In order to achieve an improvement, we fix the number of steps at each VEM iteration to two.

Remark 5.2 (MAP solution via EM algorithm) The classical EM algorithm for MAP estimation is similar to the Variational approximation, as described in Section 2.2.4. Specifically, the M-step involves maximization of (5.5) with respect to ϕ_t . Note, however, that under Proposition 5.1, (5.5) is linear in ϕ_t , and so the maximum is reached at one of the boundaries, i.e. 0 or 1. Thus MAP estimation via EM algorithm is possible only under a different approximation than that of Proposition 5.1.

5.3. Time-variant Forgetting for the MEAR model

Identification of the MEAR model with non-stationary parameters was discussed in Section 4.6.3, using time-invariant forgetting factors, $\phi_{\mathcal{NW}}$ and $\phi_{\mathcal{D}i}$. Note that the posterior distribution of the AR parameters of the MEAR model, A and Ω (4.24), is in \mathcal{NW} form which was studied in the previous Section. Hence, these results can be used for inference of an unknown forgetting factor $\phi_{\mathcal{NW}}$ for the MEAR model as follows:

$$f(\phi_{\mathcal{NW},t}|Y_t) \approx \text{tExp}(b_{\mathcal{NW}}, [0, 1]), \quad (5.25)$$

where $b_{\mathcal{NW}}$ is given by (5.23).

A time-variant forgetting factor of the Dirichlet distribution (4.25), $\phi_{\mathcal{D}i,t}$, can be derived in a similar manner to the one for $\phi_{\mathcal{NW},t}$. Once again, the normalizing constant of the forgetting operator for the Dirichlet distribution is not tractable. Hence, we invoke an approximation of the type in Proposition 5.1:

$$\zeta_{\mathcal{D}i}(\phi_{\mathcal{D}i,t}) \approx \exp(h_1 + \phi_{\mathcal{D}i,t}h_2), \quad (5.26)$$

where

$$\begin{aligned} h_1 &= \ln \zeta_{\mathcal{D}i}(\bar{\Phi}), \\ h_2 &= \ln \zeta_{\mathcal{D}i}(\Phi_{t-1}) - \ln \zeta_{\mathcal{D}i}(\bar{\Phi}). \end{aligned}$$

It is easy to verify that Theorem 2.1, applied to (4.23) extended by (5.26), yields the following approximate posterior:

$$\begin{aligned} f(\phi_{\mathcal{D}i,t}|Y_t) &\approx \text{tExp}(b_{\mathcal{D}i}, [0, 1]), \\ b_{\mathcal{D}i} &= -\ln \zeta_{\mathcal{D}i}(\Phi_{t-1}) + \ln \zeta_{\mathcal{D}i}(\bar{\Phi}) \\ &\quad - \frac{1}{2} \sum_{j=1}^c \sum_{i=1}^c (\Phi_{i,j;t-1} - \bar{\Phi}_{i,j}) E_{T_t|Y_t}(\ln t_{i,j}). \end{aligned} \quad (5.27)$$

The posterior distributions of the MEAR model parameters, A , Ω , T , \mathbf{l}_t , \mathbf{l}_{t-1} , are identical to (4.24)–(4.27) with VB-statistics (4.53)–(4.55), but with fixed values, $\phi_{\mathcal{N}\mathcal{W}}$ and $\phi_{\mathcal{D}i}$, replaced now by expected values, $\hat{\phi}_{\mathcal{N}\mathcal{W};t}$ and $\hat{\phi}_{\mathcal{D}i;t}$, from (5.25) and (5.27) respectively. This result is intuitively appealing.

It is necessary to mention that the Bayesian interpretation of forgetting [79] invokes a mixture-type model. The first component is the posterior distribution at time $t - 1$, and the second component is the alternative distribution (3.39). Hence, the forgetting factor ϕ_t plays the same role as label \mathbf{l}_t . These differ in two respects:

- ϕ_t is a continuous variable on support $[0, 1]$, while \mathbf{l}_t is a discrete random variable with c possible states $\{\mathbf{e}_1, \dots, \mathbf{e}_c\}$.
- each component of the MEAR model is an EAR model (3.29), which must be strictly data driven (that being achieved via the requirement of a non-zero Jacobian of (3.26) in Section 3.2.2). There is no such condition on the alternative distribution in forgetting (3.39). Typically, alternative distributions are chosen as non-committal priors, i.e. fixed and flat $\forall t$.

Remark 5.3 Note that the update of the VB-statistics, V_t (4.28), for the MEAR model has the following form:

$$V_t = \bar{V} + w_{1;t}V_{1,t} + \dots + w_{c;t}V_{c,t} + \phi_t (V_{t-1} - \bar{V}). \quad (5.28)$$

We have introduced the notation $V_{i,t} = \mathbf{y}_{i,t}\mathbf{y}'_{i,t}$. Note that the statistic at time t is, therefore, a weighted linear combination of statistics from different sources: (i) expert knowledge, \bar{V} , (ii) c transformations of the data source, $w_{i;t}V_{i,t}$, and (iii) accumulated statistics from the past data, V_{t-1} . The same is composition can also be shown for the remaining statistics, Φ_t , ν_t . This structure is common in algorithms for on-line identification of non-stationary models, however, all the weights are typically assumed to be known. In this Section, we have assumed that all weights involved in the update (5.28), i.e. $w_{1;t}, \dots, w_{c;t}$ and ϕ_t , are unknown. The resulting algorithm, balances, in effect, the contributions being made by past data, current data, and expert knowledge. It achieves this on-line.

From (5.28), we note that, in effect, the alternative distribution is balanced with respect to components in the MEAR model. Therefore, choice of the alternative distribution must be considered as a part of the filter-bank design.

5.4. Inference of an AR process with Switching Parameters

This experiment is designed to verify the ability of time-variant forgetting to detect sudden changes in parameters (changepoints) and to adjust the forgetting factor accordingly. A univariate second-order stable AR model (i.e. $\mathbf{x}_t = [d_{t-1}, d_{t-2}]'$) with parameters $\omega = 1$, and

$$A = \begin{cases} [1.8, -0.98] & \text{if } \text{mod}(t, 30) = \text{mod}(t, 60) \\ [-0.29, -0.98] & \text{if } \text{mod}(t, 30) \neq \text{mod}(t, 60) \end{cases},$$

where $\text{mod}(t, x)$ denotes the modulo function, i.e. remainder after division. The model was identified via VB-posteriors (5.19) and (5.20) using the VEM algorithm (Algorithm 2.2) to complete the asso-

ciated VB-statistics (5.21)–(5.23). In our simulation, we have chosen the alternative statistics to be

$$\bar{V} = \text{diag}([1, 0.001, 0.001]'), \quad \bar{\nu} = 10, \quad (5.29)$$

corresponding to the prior estimates $A_0 = [0, 0]$, $\text{var}(a_1) = \text{var}(a_2) = 1000$, $\omega_0 = 0.1$. The prior distribution is chosen equal to the alternative distribution. The initial value of the forgetting factor was $\hat{\phi}_t^{(1)} = 0.7$. The full VEM algorithm was stopped when $|\hat{\phi}_t^{(m)} - \hat{\phi}_t^{(m-1)}| < 0.001$. The restricted VEM algorithm was stopped after two steps. The results of identification are displayed in Figure 5.1.

Note that the method—to within one time-step—correctly detects a change of parameters and estimates the forgetting factor as low as $\hat{\phi}_t = 0.05$ (at $t = 33$), which achieves almost instant replacement of statistics V_t, ν_t by alternative (prior) values $\bar{V}, \bar{\nu}$. Thus, identification process is restarted. Note that number of iterations of the VEM algorithm is significantly higher at the changepoint. Therefore, at these points, the expected value of forgetting factor, $\hat{\phi}_t$, obtained using the restricted VEM algorithm (Remark 5.1), remains too high compared to the converged value of the full VEM algorithm (Figure 5.1). For comparison, the results of identification with stationary forgetting, $\phi_t = 0.9$, are displayed in Figure 5.1. The best parameter tracking is achieved using the VB posterior distributions evaluated via VEM iterated to convergence. Identification of the process using restricted VEM is acceptable if the parameter variations are not too rapid.

5.5. Inference of a Stationary AR Process using Time Variant Forgetting

The forgetting technique can be used even for on-line identification of stationary processes. In on-line scenario, the early estimates are heavily dependent on the chosen prior distribution, $f(\theta)$, which can negatively influence the convergence of the identification algorithm. Therefore, various *discount schedules* have been proposed to overcome this problem [35].

We now compare the performance of our method with the discount schedule proposed in of [35], which can be seen as a heuristic choice of forgetting factor in the form

$$\phi_t = 1 - \frac{1}{\eta_1(t-2) + \eta_2}, \quad (5.30)$$

where η_1 , and η_2 are *a priori* chosen constants. Note that $\phi_t \rightarrow 1$ as $t \rightarrow \infty$. The aim of this schedule is to discount the influence of the (possibly wrong) prior statistics at the beginning of identification. As the posterior becomes data-dominant, the forgetting factor approaches unity, resulting in standard AR identification. The rate of forgetting is, however, chosen by the designer via η_1 and η_2 . In our approach, this rate is inferred from data.

A univariate second-order, stable, AR model (i.e. $\mathbf{x}_t = [d_{t-1}, d_{t-2}]$) with parameters $A = [1.8, -0.98]'$, $\omega = 1$, was simulated. The results of parameter identification using VB posterior distributions (Corollary 5.1) are displayed in Figure 5.2. For comparison, identification using the discount factor (5.30) was also undertaken (Figure 5.2), via standard AR identification with forgetting (Section 3.2.3), for the choice of non-stationary forgetting factor (5.30) with $\eta_1 = \eta_2 = 1$.

5. BAYESIAN INFERENCE OF NON-STATIONARY AUTOREGRESSIVE MODELS USING TIME-VARIANT FORGETTING

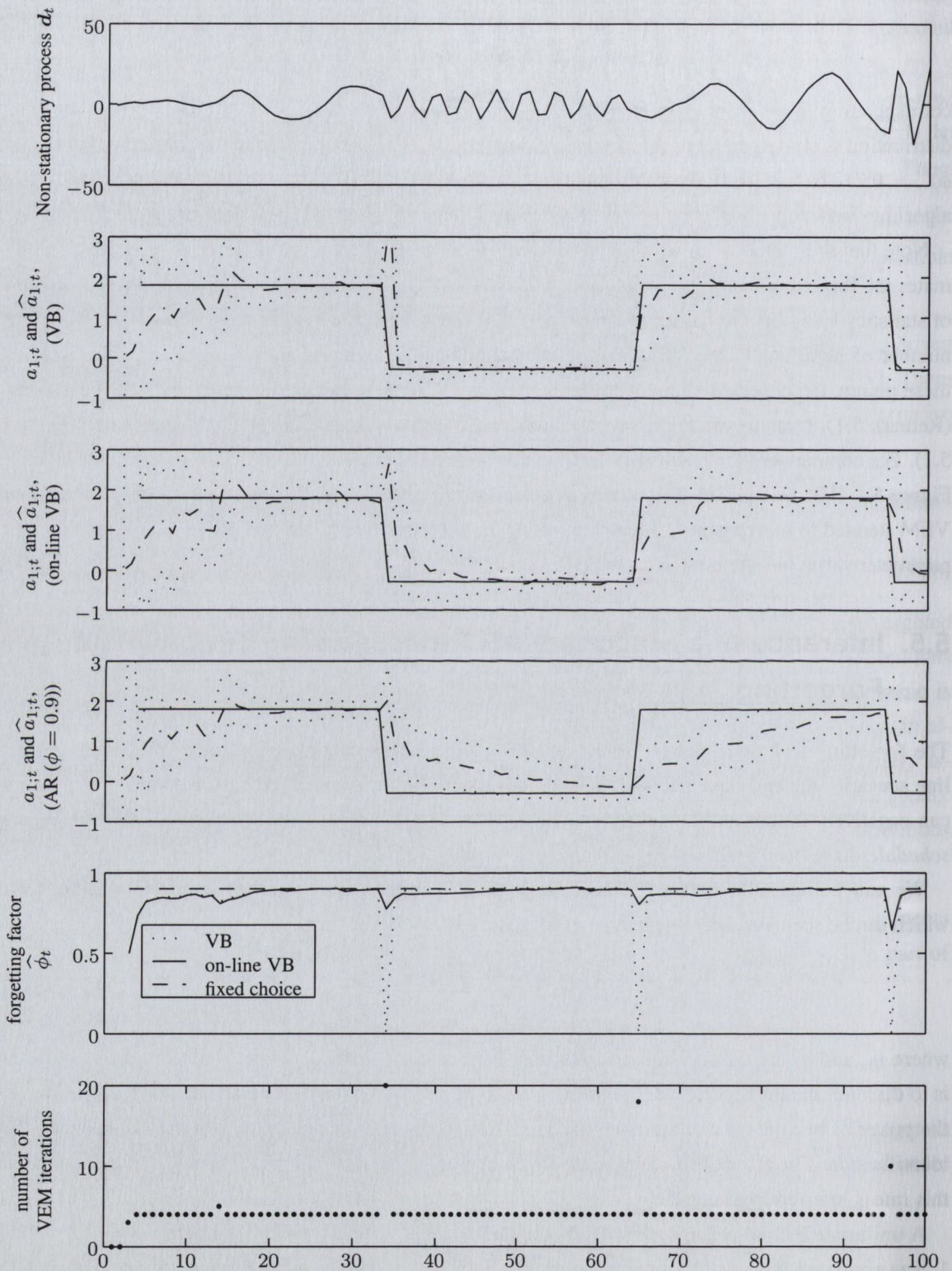


Figure 5.1.: Results of identification of a non-stationary process using time-variant forgetting. In sub-figures (i)–(iii), full lines denote simulated values of parameters, dashed lines denote posterior expected values, and dotted lines denote uncertainty bounds.

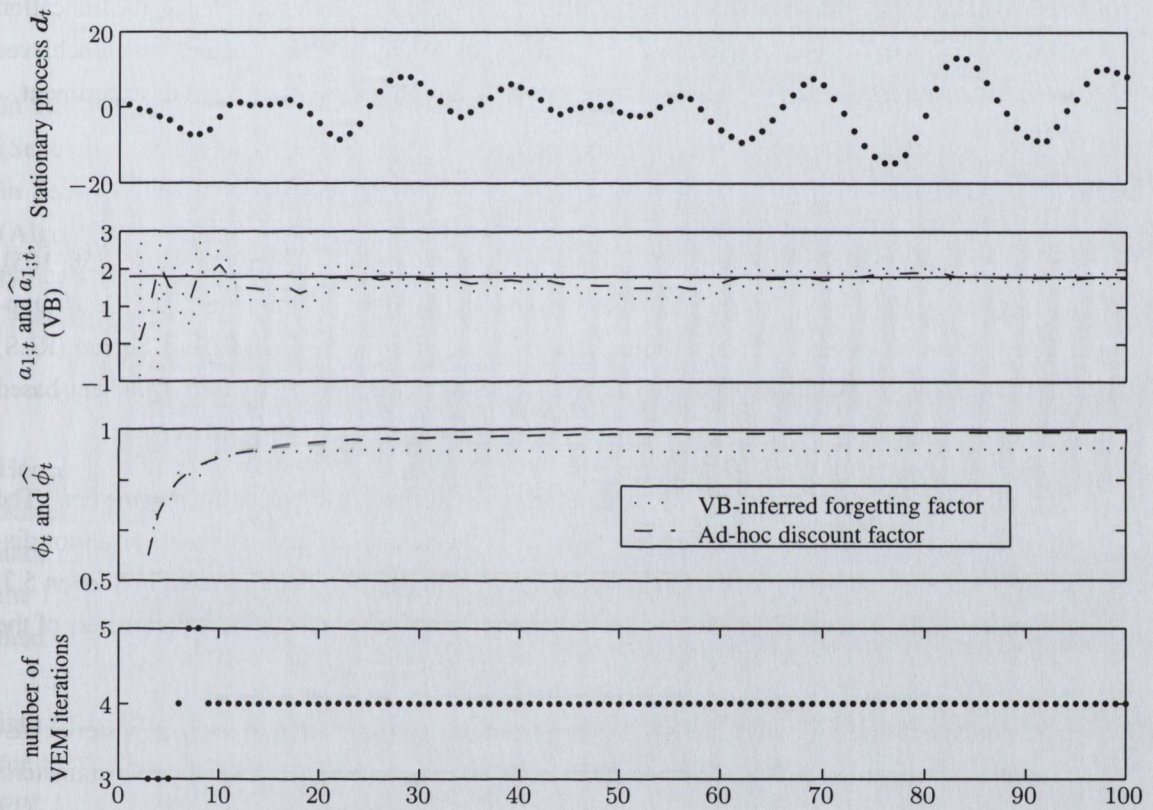


Figure 5.2.: Results of identification of a stationary process using time-variant forgetting. In the second sub-figure, full lines denote simulated values of parameters, dashed lines denote posterior expected values, and dotted lines denote uncertainty bounds.

Note that for $t < 15$, the expected value of the unknown forgetting factor, $\hat{\phi}_t$, is very close to the discount factor. However, as $t \rightarrow \infty$ the $\hat{\phi}_t$ does not converge to one but a smaller invariant value (in this simulation, $\phi_t \approx 0.92$ for $t > 20$). This is a consequence of the stationary alternative distribution, $\bar{f}(\cdot)$. Note that for stationary forgetting, $\phi_t = \phi$ (Section 3.2.3), parameters of the alternative distribution, \bar{V} and $\bar{\nu}$, are always present in V_{t-1}, ν_{t-1} (3.43), (3.44). Hence, $\lim_{t \rightarrow \infty} \hat{\phi}_t$ is determined by the chosen alternative $\bar{f}(\bar{V}, \bar{\nu})$. The alternative statistics were chosen as (5.29).

We believe that, in many practical applications, it is easier to choose a reasonable prior from expert knowledge, than to tune the discount schedule, via parameters η_1 and η_2 . The latter must be done experimentally, which may be time-consuming.

Note that number of iterations of the VEM algorithm is rather low (Figure 5.2). Hence, the truncation of VEM cycles (Remark 5.1) yields almost identical results to the full VB scheme, which achieves convergence at each time-step. The results of the latter were not, therefore, shown in this experiment.

5.6. Discussion

The technique of forgetting is used in many estimation methods for non-stationary processes [58, 105], with the forgetting factor considered to be time-invariant and known. Attempts to relax the assumption of *a priori* known forgetting factor were made, especially for the Recursive Least Square (RLS) algorithms [58]. The method presented in [102] is the closest to our approach. It is a gradient-based estimation of the forgetting factor for the RLS algorithm. We note the following differences:

- The RLS algorithm is based on the assumption of a Normal distribution of parameters. The Bayesian interpretation of forgetting (Section 3.2.3) can be applied to any class of posterior distributions that is closed under the geometric mean (3.39). This was demonstrated in Section 5.3, where we applied variable forgetting to the identification of the non-stationary parameters of the Markov model.
- In our approach, we minimize the KL distance from the approximating to the true posterior distribution at *each* time t . This allows for rapid changes (i.e. switching) of the model parameters. The criterion of asymptotic mean square error minimized in [102] addresses slower variations of parameters.
- The posterior inference of the forgetting factor (5.20) is sensitive to the chosen alternative pdf of the parameters, via \bar{V} and $\bar{\nu}$ in (5.23). They play a similar role to the tuning parameters (α and β) in [102]. The alternative distribution can be chosen using the available expert knowledge of the problem, via formal prior elicitation procedures [106]. The tuning parameters of [102] must be adjusted experimentally.

We have noted that the optimal posterior distribution of the forgetting factor is not tractable and it must be approximated to achieve a numerically efficient identification algorithm (Proposition 5.1). The choice of approximation, of course, influences the quality of results of the inference algorithm. The proposed approximation is simple and it may be inappropriate for certain tasks. Other approximations might be investigated in such cases, to improve performance.

Chapter 6.

Bayesian Treatment of Principal Component Analysis

In this Chapter, we study the Bayesian inference of parameters of the Probabilistic PCA (PPCA) model (Section 3.3). An approximate inference of the model, using the VB approximation, was reviewed in Section 3.3.3. The parameters of the posterior distribution are evaluated via the VEM algorithm (Algorithm 2.2), which is computationally intensive in the high-dimensional contents where PCA is typically applied. In this Chapter, we new VB identification algorithms, which are significantly faster.

Recall, that the PPCA model (3.50) is

$$f(D|A, X, \omega, r) = \mathcal{N}(AX', \omega^{-1}I_p \otimes I_t).$$

Hence, throughout the Chapter, we assume the identity covariance matrix of the additive noise. Results achieved in this Chapter can also be used even for colored, Normal distributed noise of *known* covariance matrix. In this case, the results are valid for a matrix of pre-processed \tilde{D} (3.88). Identification for the Factor Analysis model—i.e. the PPCA model with *unknown* covariance—will be addressed in the next Chapter.

The model is first studied at the lowest possible dimension, i.e. scalar variables, to gain insight into the problem. Detailed analysis of this *toy problem* leads to (i) faster evaluation of the posteriors for PPCA, and (ii) interest in the orthogonal parameterization of the PPCA model. The orthogonal PPCA model is then proposed and its Bayesian inference is developed. Performance of both methods is compared on simulated data. Application of the resulting algorithms to real data is deferred to Chapter 7.

6.1. Toy Problem: Scalar Decompositions

In this Section, we reduce the model (3.50) to the simplest case. Reducing (3.50) to minimal dimensions, i.e. $p = n = r = 1$, we obtain a scalar model:

$$d = ax + e. \tag{6.1}$$

Model (6.1) is clearly over-parameterized, with three unknown parameters (a, x, e) for one measurement, d . It expresses any additive/multiplicative decomposition of a real number. Separation of the

‘signal’, ax , from the ‘noise’, e , is not possible without further information, i.e. the model (6.1) must be regularized. Towards this end, let us assume that noise e is distributed as $\mathcal{N}(0, \sigma_e)$. Then,

$$f(d|a, x, \sigma_e) = \mathcal{N}(ax, \sigma_e), \quad (6.2)$$

where σ_e is assumed to be known.

The likelihood function for this model for $d = 1$, $\sigma_e = 1$ is displayed in the upper row of Figure 6.1, in surface plot (left) and contour plot (right) forms. The maximum of the likelihood is reached anywhere in the manifold defined by the signal estimate

$$\hat{a}\hat{x} = d. \quad (6.3)$$

This illustrates the rotational ambiguity problem (3.53) of Section 3.3. Further regularization is clearly required.

6.1.1. Bayesian Formulation

It can also be appreciated from Figure 6.1 (upper-left), that volume under the likelihood function is infinite. This means that $f(a, x|d, \sigma_e) \propto f(d|a, x, \sigma_e) f(a, x)$ is improper (unnormalizable) when the parameter prior $f(a, x)$ is itself improper (uniform in \mathbb{R}^2). Prior-based regularization is clearly required to achieve a proper posterior distribution via Bayes’ rule. Under the assignment,

$$f(a|\sigma_a) = \mathcal{N}(0, \sigma_a), \quad (6.4)$$

$$f(x|\sigma_x) = \mathcal{N}(0, \sigma_x), \quad (6.5)$$

the posterior distribution is:

$$f(a, x|d, \sigma_e, \sigma_a, \sigma_x) \propto \exp\left(-\frac{1}{2} \frac{(ax - d)^2}{\sigma_e} - \frac{1}{2} \frac{a^2}{\sigma_a} - \frac{1}{2} \frac{x^2}{\sigma_x}\right). \quad (6.6)$$

(6.6) is displayed in the lower row of Figure 6.1, for $\sigma_a = 10$, $\sigma_x = 20$, $d = 1$, $\sigma_e = 1$. The model is now regularized, with the consequence that the posterior (6.6) is normalizable (proper) with point maximizers (MAP estimates) as follows:

1. For $d > \frac{\sigma_e}{\sqrt{\sigma_a \sigma_x}}$, then

$$\hat{x} = \pm \left(d \sqrt{\frac{\sigma_x}{\sigma_a}} - \frac{\sigma_e}{\sigma_a} \right)^{\frac{1}{2}}, \quad (6.7)$$

$$\hat{a} = \pm \left(d \sqrt{\frac{\sigma_a}{\sigma_x}} - \frac{\sigma_e}{\sigma_x} \right)^{\frac{1}{2}}. \quad (6.8)$$

Note that product of maxima is

$$\hat{a}\hat{x} = d - \frac{\sigma_e}{\sqrt{\sigma_a \sigma_x}}. \quad (6.9)$$

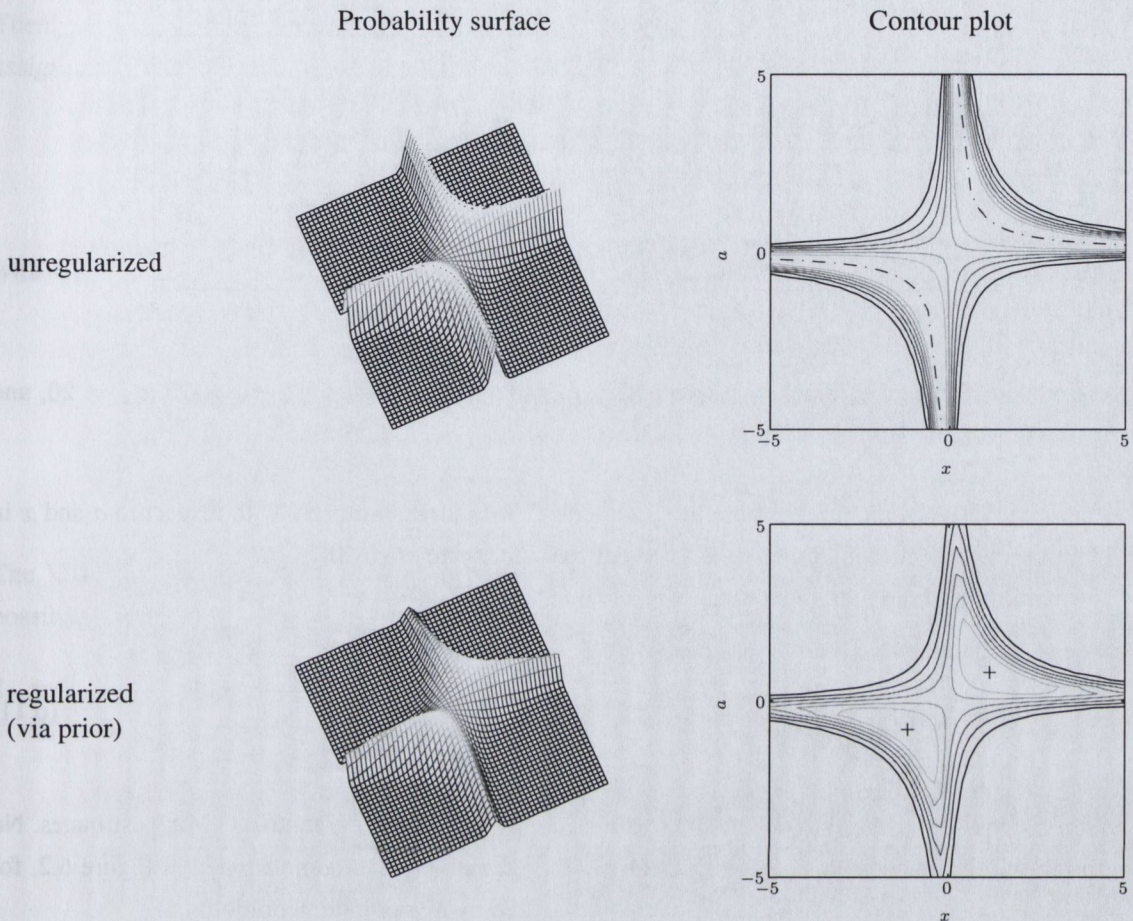


Figure 6.1.: Illustration of scaling ambiguity in the toy problem. **Upper row:** the likelihood function $f(d|a, x, \sigma_e)$ for $d = 1$ (dash-dotted line denotes manifold of maxima). **Lower row:** posterior pdf $f(a, x|d, \sigma_a, \sigma_x, \sigma_e)$ for $d = 1$, with priors $\sigma_e = 1$, $\sigma_a = 10$, $\sigma_x = 20$. Cross marks denote maxima.

From (6.9), the signal estimate has been shifted towards the coordinate origin compared to (6.3). For the choice, $\sigma_a \ll \sigma_e$ and $\sigma_x \ll \sigma_e$, the prior strongly influences the posterior and is therefore informative. For the choice, $\sigma_a \gg \sigma_e$ and $\sigma_x \gg \sigma_e$, the prior has negligible influence on the posterior and can be considered as non-committal.

$$2. \text{ For } d \leq \frac{\sigma_e}{\sqrt{\sigma_a \sigma_x}}, \hat{x} = \hat{a} = 0.$$

Clearly, the quantity $\tilde{d} = \frac{\sigma_e}{\sqrt{\sigma_a \sigma_x}}$ constitutes an important inferential breakpoint. For $d > \tilde{d}$, a non-zero signal is inferred, for $d \leq \tilde{d}$, the observation is considered to be purely noise.

6.1.2. Full Bayesian Solution

The posterior distribution (6.6) is normalizable, but the normalizing constant cannot be expressed in closed form. Integration of (6.6) over $x \in \mathfrak{R}$ yields the following marginal distribution for a :

$$f(a|d, \sigma_e, \sigma_a, \sigma_x) \propto \frac{1}{2} \exp\left(-\frac{1}{2} \frac{d^2 \sigma_a + a^4 \sigma_x + a^2}{\sigma_a (a^2 \sigma_x + 1)}\right) [\pi^2 \sigma_a (a^2 \sigma_x + 1)]^{-\frac{1}{2}}, \quad (6.10)$$

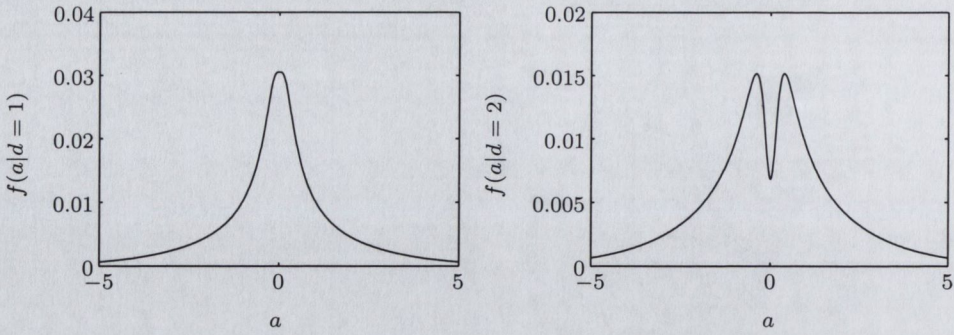


Figure 6.2.: Analytical marginals (of distribution from Figure 6.1). $\sigma_e = 1$, $\sigma_a = 10$, $\sigma_x = 20$, and $d = 1$ (left), $d = 2$ (right).

whose integral over a is not available in closed form. Structural symmetry with respect to a and x in (6.1) implies that the marginal inference for x has the same form as (6.10).

The maximum of the marginal (6.10) is reached for

$$\hat{a} = \begin{cases} \pm \left[\frac{-\sigma_a \sigma_x - 2\sigma_e + \sqrt{\sigma_a \sigma_x (\sigma_a \sigma_x + 4d^2)}}{2\sigma_x} \right]^{\frac{1}{2}} & \text{if } d > \sqrt{\frac{\sigma_e^2}{\sigma_a \sigma_x} + \sigma_e}, \\ 0 & \text{if } d \leq \sqrt{\frac{\sigma_e^2}{\sigma_a \sigma_x} + \sigma_e}. \end{cases} \quad (6.11)$$

The same symbol, \hat{a} , is used to denote the (distinct) joint (6.8) and marginal (6.11) MAP estimates. No confusion will be encountered. Both cases of (6.11), respectively, are demonstrated in Figure 6.2, for $d = 1$ (left) and $d = 2$ (right). The curves were normalized by numerical integration.

The only operation on the marginal posterior that can be evaluated analytically, is the maximum (6.11). Most importantly, analytical normalization of the marginal posteriors is not available, and so the moments of the posterior must be evaluated using numerical methods. Recall, that purpose of this analysis is to understand the Bayesian inference of the multivariate PPCA model (3.50). The full Bayesian solution presented in this section can, indeed, be extended into multivariate case [50]. The multivariate posterior distributions suffer the same difficulties as those of the toy problem: normalization of the marginal posteriors and their moments must be evaluated using numerical methods, such as MCMC (Section 2.2.6) [64]. Hence, we now seek an approximation of the posterior distribution using a Variational Bayes approximation.

6.1.3. Variational Bayes (VB) Approximation

Corollary 6.1 (Corollary 4 of Theorem 2.1) Consider the following conditionally independent factorization of (6.6):

$$\check{f}(a, x|d, \sigma_e, \sigma_a, \sigma_x) = \check{f}(a|d, \sigma_e, \sigma_a, \sigma_x) \check{f}(x|d, \sigma_e, \sigma_a, \sigma_x). \quad (6.12)$$

Then, using (6.12) and (6.6) in Theorem 2.1, the VB-optimal form of (6.12) via found in the following assignments:

$$\tilde{f}(x|d, \sigma_e, \sigma_a, \sigma_x) = \mathcal{N}(\hat{x}, \phi_x),$$

$$\tilde{f}(a|d, \sigma_e, \sigma_a, \sigma_x) = \mathcal{N}(\hat{a}, \phi_a),$$

with VB-statistics

$$\hat{x} = d\sigma_e^{-1}\phi_x\hat{a}, \quad (6.13)$$

$$\hat{a} = d\sigma_e^{-1}\phi_a\hat{x}, \quad (6.14)$$

$$\phi_x = (\sigma_e^{-1}(\phi_a + \hat{a}\hat{a}) + \sigma_x^{-1})^{-1}, \quad (6.15)$$

$$\phi_a = (\sigma_e^{-1}(\phi_x + \hat{x}\hat{x}) + \sigma_a^{-1})^{-1}. \quad (6.16)$$

The VB-statistics— \hat{a} , \hat{x} , ϕ_a , ϕ_x —can be found in *closed-form* from (6.13)–(6.16). There are three possible cases:

1. zero-signal inference:

$$\hat{x} = 0, \quad (6.17)$$

$$\hat{a} = 0,$$

$$\phi_x = \frac{\sigma_e}{2\sigma_a} \left(\sqrt{1 + \frac{4\sigma_a\sigma_x}{\sigma_e}} - 1 \right),$$

$$\phi_a = \frac{\sigma_e}{2\sigma_x} \left(\sqrt{1 + \frac{4\sigma_a\sigma_x}{\sigma_e}} - 1 \right),$$

2. and 3. non-zero signal inference:

$$\hat{x} = \pm \left[\frac{(d^2 - \sigma_e) \sqrt{\sigma_a\sigma_x} - d\sigma_e}{d\sigma_a} \right]^{\frac{1}{2}}, \quad (6.18)$$

$$\hat{a} = \pm \left[\frac{(d^2 - \sigma_e) \sqrt{\sigma_a\sigma_x} - d\sigma_e}{d\sigma_x} \right]^{\frac{1}{2}},$$

$$\phi_x = \frac{\sigma_e}{d} \sqrt{\frac{\sigma_x}{\sigma_a}} \operatorname{sgn}(d^2 - \sigma_e), \quad (6.19)$$

$$\phi_a = \frac{\sigma_e}{d} \sqrt{\frac{\sigma_a}{\sigma_x}} \operatorname{sgn}(d^2 - \sigma_e).$$

Here, $\operatorname{sgn}(\cdot)$ returns the sign of the argument.

From (6.19) we note, that extreme 2. and 3. is meaningful only for $d > \sqrt{\sigma_e}$. However, (6.18) collapses to $\hat{x} = 0$ (i.e. to the zero-signal inference) for

$$\tilde{d} = \frac{1}{2} \frac{\sigma_e + \sqrt{\sigma_e(\sigma_e + 4\sigma_a\sigma_x)}}{\sqrt{\sigma_a\sigma_x}} \approx \sqrt{\sigma_e} + \frac{\sigma_e}{2\sqrt{\sigma_a\sigma_x}}. \quad (6.20)$$

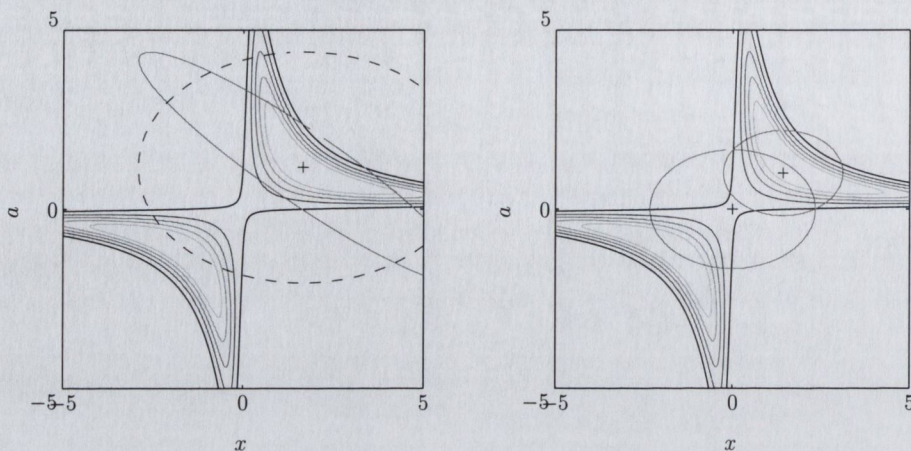


Figure 6.3.: Comparison of Laplace and Variational Bayes approximations (for distributions in Figure 6.1). $\sigma_e = 1$, $\sigma_a = 10$, $\sigma_x = 20$, $d = 2$. **Left:** (Laplace approximation): full line ellipse corresponds to 2-standard-deviation boundary of the joint Normal approximation; dashed line corresponds to product of Laplace marginal approximations (for comparison with VB). **Right:** (Closed-form solution of the VB approximation): VB approximation for the zero-signal (6.17) and non-zero-signal modes (6.18) are shown.

Hence, (6.20) denotes the VB-based breakpoint. For $d > \tilde{d}$, a non-zero signal is inferred (case 2. and 3.), for $d \leq \tilde{d}$, the observation is considered to be purely noise.

These solutions are illustrated in Figure 6.3 (right). This result illustrates a key consequence of the VB approximation, namely, absence of any cross-correlation between variables, in direct consequence of the conditional independence assumption. For comparison, the result of a Laplace approximation (Section 2.2.2) is displayed in Figure 6.3 (left). The Laplace approximation does model cross-correlation between variables.

The availability of a closed-form VB solution is rare. Therefore, we have the opportunity to study properties of the standard VEM algorithm in this case. Trajectory of the VEM iterations for mean values of VB-posteriors (6.17), (6.17), are shown in Figure 6.4, for $d = 2$ (left) and $d = 1$ (right). These two cases demonstrate the two distinct modes of solution of equations (6.13)–(6.16). Though there are no limiting conditions for the mode in origin, iterative algorithm typically converge to the positive (or negative) solution for $d > \sigma_e$. This suggest that for $d > \sigma_e$ the zero-signal solution is a local extreme of the KL distance.

These results, along with Colorary 6.1, suggest the following:

- The prior distribution is indispensable; as it regularizes the model, and necessary for yielding finite VB statistics. With uniform priors, i.e. $\sigma_a \rightarrow \infty$ and $\sigma_x \rightarrow \infty$, none of the derived solutions is valid.
- From (6.18), the ratio of the posterior expected values \hat{a}/\hat{x} is fixed by the priors.
- The inferential breakpoint, \tilde{d} —i.e. value of d above which a non-zero signal is inferred—depends on the product of $\sigma_a\sigma_x$ (6.20).

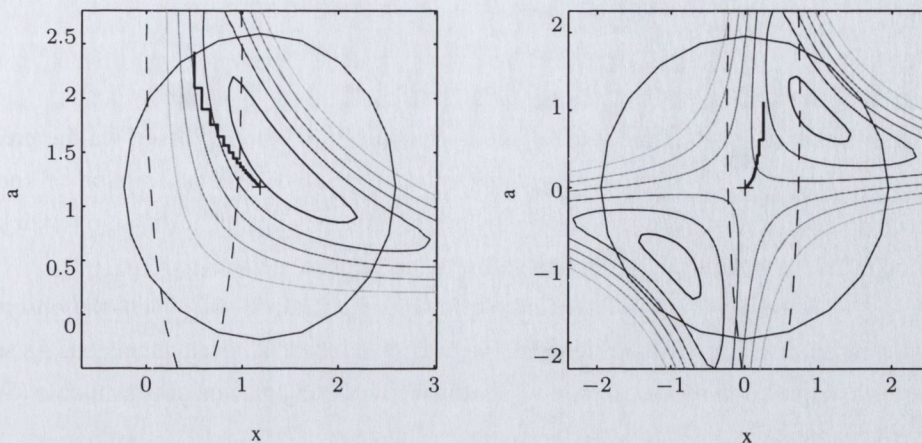


Figure 6.4.: VB approximations of the toy problem parameter distributions, using the VEM algorithm. Dashed line denote initial VB-posterior; full line denotes converged VB-posterior; trajectory of the VB means is also illustrated, $\sigma_e = 1$, $\sigma_a = 10$, $\sigma_x = 10$. **Left:** (non-zero-signal mode) $d = 2$. **Right:** (zero-signal-mode) $d = 1$.

Remark 6.1 (Alternative priors) The choice of priors in VPCA (3.65), (3.66) corresponds to the following choice of priors for the toy problem:

$$\sigma_x = 1, \quad f(\sigma_a | \alpha, \beta) = \mathcal{G}(\alpha, \beta),$$

where σ_x is fixed, but σ_a is considered as unknown (i.e. random variable of \mathcal{G} distribution with known hyper-parameters α, β) to be estimated jointly with other parameters. Using this prior structure for the toy problem, closed-form VB-statistics can also be found. In fact, the closed form VB-statistics can be found for many different choices. For example:

1. symmetric priors

$$f(\sigma_a | \alpha, \beta) = f(\sigma_x | \alpha, \beta) = \mathcal{G}(\alpha, \beta),$$

2. fixed σ_x , and σ_a conditioned by precision ω ,

$$\sigma_x = 1, \quad f(\sigma_a | \omega, \alpha, \beta) = \mathcal{G}(\alpha, \beta \omega).$$

This suggests, that an analytical solution may be achieved for a wider choice of multivariate priors. This may be significant for the development of numerically efficient algorithms for extended PPCA models.

6.1.4. Non-Degenerate Parameterization for the Toy Problem

The inference problems encountered in previous Sections are consequence of degenerate parameterization of the mean value: $m = ax$. In the multivariate case, this parameterization was used to model the rank restriction in $M_{(r)}$, $r < \min(n, p)$. In this scalar case, $n = p = r = 1$. Hence, the multiplicative

decomposition is not necessary, and we can work with a purely additive model:

$$d = m + e. \quad (6.21)$$

Once again, we assume that, e is Normally distributed: $f(e) = \mathcal{N}(0, \omega)$, expressed via the precision parameter, ω , for consistency with Section 3.2. Identification of m under the assumption of known ω constitutes the trivial identification task of inferring the mean value of a Normal distribution with known variance [17]. The problem is well-posed, and no regularizing prior is needed.

We, therefore, analyze the more complicated problem, where ω is unknown. Then, the observation model (6.21) is a Normal distribution conditioned on an unknown mean value and precision. As such, it is a special case of the regression model (Section 3.1), for which a conjugate prior is available (Section 3.2)

$$f(m, \omega) = \mathcal{NW}(V_0, \nu_0). \quad (6.22)$$

We choose the prior statistics of (6.22) to be $V_0 = \text{diag}([\varepsilon_1, \varepsilon_2]')$, $\nu_0 = \varepsilon_3$, where scalars $\varepsilon = [\varepsilon_1, \varepsilon_2, \varepsilon_3]'$ can be chosen small to yield a flat—i.e. non-committal—pdf. The posterior distribution is then

$$f(m, \omega | d, \varepsilon) = \mathcal{NW}(V_0 + [d, 1]'[d, 1], \nu_0 + 1). \quad (6.23)$$

From (6.21), and (6.22), the joint distribution is:

$$f(d, m, \omega | \varepsilon) = \omega^{\frac{1}{2}(1+\varepsilon_3)} \exp\left(-\frac{1}{2}(d-m)^2\omega - \frac{1}{2}m^2\varepsilon_2\omega - \frac{1}{2}\varepsilon_1\omega\right). \quad (6.24)$$

The posterior distribution is found using Bayes' rule:

$$f(m, \omega | d, \varepsilon) = \frac{f(d, m, \omega | \varepsilon)}{f(d | \varepsilon)}. \quad (6.25)$$

In this case, the analytical form of the posterior (6.25) is of the $\mathcal{NW}(\cdot)$ form (3.13), which is regular and for which moments are available in Appendix A.2. Recall, from (6.10), that this was not possible for the degenerate PPCA decomposition. Hence, for comparison, we now proceed with VB approximation of (6.25).

Consider the following conditional independent factorization of (6.25):

$$\check{f}(m, \omega | d) = \check{f}(m | d) \check{f}(\omega | d). \quad (6.26)$$

Using (6.26) with (6.25) in Theorem 2.1, the VB-optimal form of (6.26) is found via the following assignments:

$$\tilde{f}(m | d, \varepsilon) = \mathcal{N}_m\left(d(1 + \varepsilon_2)^{-1}, (1 + \varepsilon_2)\widehat{\omega}^{-1}\right), \quad (6.27)$$

$$\tilde{f}(\omega | d, \varepsilon) = \mathcal{G}_\omega\left(\frac{3}{2} + \varepsilon_3, \frac{1}{2}\left[(1 + \varepsilon_2)\widehat{m}^2 - 2d\widehat{m} + d^2 + \varepsilon_1\right]\right). \quad (6.28)$$

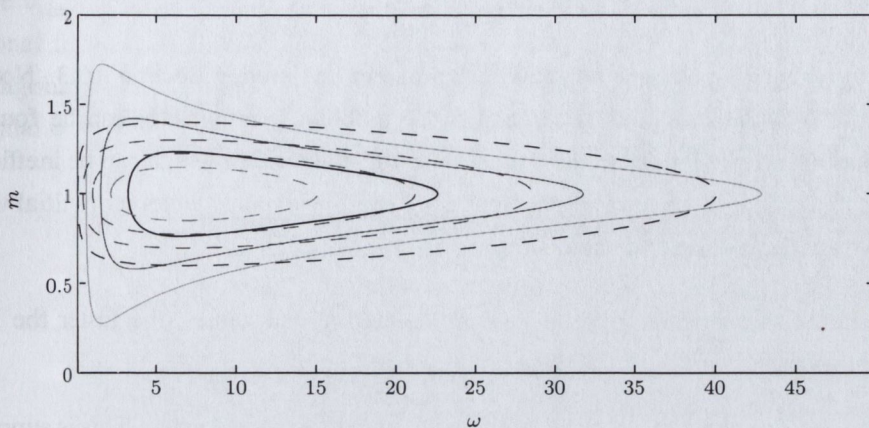


Figure 6.5.: Illustration of the VB-posterior, (6.27)–(6.28), for the scalar additive decomposition (dashed contour). Full contour lines denote the exact \mathcal{NW} posterior distribution (6.25).

For this simple case, $\tilde{f}(m, \omega | d, \varepsilon) = \mathcal{N}_m(\cdot) \mathcal{G}_\omega(\cdot)$ can be compared to the exact posterior (6.23). Graphical comparison for $d = 1$ is displayed in Figure 6.5.

We note the following:

- the prior distribution (6.22) can be chosen such that, $\varepsilon_2 = \varepsilon_3 = 0$, without loss of tractability. This choice corresponds to a uniform (improper) prior on parameter m . Hence, in contrast to the PPCA model (6.1), the prior on the mean value has no regularizing role.
- the hyperparameter, ε_1 , above has a regularizing effect. It is a hyper-parameter of the prior on the precision parameter ω . Note that we are inferring ω from a single observation only, hence the data does not contain any information about uncertainty and the inference must be regularized.

6.1.5. The Lessons Learnt

In this Section, we have studied scalar decomposition in an attempt to understand the nature of the VPCA approximation for the PPCA model (Section 3.3.3). We have noticed the following, which generalizes to the full multivariate context:

1. if the variance of the noise is known, the VB-statistics can be found in *closed-form*. This eliminates the need to evaluate VB-statistics via an iterative VEM algorithm;
2. inference of the model, $d = m + e$, where the mean is not degeneratively parameterized is predictably much simpler. Moreover, the correct posterior distributions of m and ω exhibits less correlation than is the case of a and x . Therefore, the conditional independence assumption of the VB approximation is less intrusive in the former case. More formal analysis related to this issue can be found in [84].

We will next explore these insights for the multivariate PPCA model (3.50).

6.2. Fast Variational PCA (FVPCA)

In this Section, we study the VB-posterior for the PPCA problem reviewed in Section 3.3.3. Notably, we exploit the fact that VB-statistics for the associated toy problem (Section 6.1) can be found in closed form. The solution reviewed in Section 3.3.3 uses the VEM algorithm which can be inefficient, especially when poor initial conditions are chosen. Hence, we begin with consideration of initial values for the VEM algorithm. Here, we use two simple ideas:

- the closer the initial value of the VB-statistics are to the optimal values, the faster the VEM algorithm will converge,
- the MAP estimate of a parameter is typically not too far from its expected value. This is supported by the fact that the approximate posterior distributions (3.72)–(3.73) are of Normal-type for which the mean and the maximum are identical.

Recall, that the maximum of the PPCA model (3.50) was reached for an orthogonal solution (3.51). Hence, we expect that solution of (3.76) will be very close to (3.55).

Proposition 6.1 (Orthogonal solution of VPCA) *Consider a special case of distributions of random variables A and X , with restricted first and second moments. The first moments, \hat{A} (3.76) and \hat{X} (3.78), are formed from scaled singular vectors of the data matrix, D (3.52),*

$$\hat{A} = U_{r;D} K_A, \quad (6.29)$$

$$\hat{X} = V_{r;D} K_X, \quad (6.30)$$

where $K_A = \text{diag}(\mathbf{k}_A) \in \mathbb{R}^{r \times r}$ and $K_X = \text{diag}(\mathbf{k}_X) \in \mathbb{R}^{r \times r}$ denote matrix constants of proportionality. The second moments (3.77), (3.79) are restricted to have a diagonal form:

$$\Sigma_A = \text{diag}(\boldsymbol{\sigma}_A), \quad (6.31)$$

$$\Sigma_X = \text{diag}(\boldsymbol{\sigma}_X). \quad (6.32)$$

Then, evaluation of the VB-statistics (3.76)–(3.83), via the VEM algorithm initialized with values in the form of (6.29)–(6.32), yields results also in the form of (6.29)–(6.32).

Proof: By induction: (i) the VEM algorithm is initialized in the form (6.29)–(6.32); (ii) substituting (6.29)–(6.32) into (3.76)–(3.83) yields results in the form of (6.29)–(6.32). ■

Note that, under Proposition 6.1, the distribution of A and X are determined by the constants of proportionality, \mathbf{k}_A and \mathbf{k}_X , and variances, $\boldsymbol{\sigma}_A$ and $\boldsymbol{\sigma}_X$, respectively. The iterative algorithm is then greatly simplified, since we need only iterate on the $4r$ degrees of freedom constituting \mathbf{k}_A , \mathbf{k}_X , $\boldsymbol{\sigma}_A$, and $\boldsymbol{\sigma}_X$ together, and not on \hat{A} , \hat{X} , Σ_A , Σ_X with $r(p+n+2r)$ degrees of freedom. Note that (3.76)–(3.83) now involve products of diagonal matrices. Hence, we need only evaluate diagonal elements, using identities of the kind

$$K_A K_X = \text{diag}(\mathbf{k}_A \circ \mathbf{k}_X),$$

where \circ denotes Hadamard product. Equations (3.76)–(3.83) can now be re-formulated in efficient diagonal form. Intuitively, the equation for the mean value μ_A (3.76) is replaced by an equation for its diagonal \mathbf{k}_A , the equation for the covariance matrix, Σ_A (3.77), is replaced by an equation for its diagonal σ_A , etc.

$$\mathbf{k}_A = \widehat{\omega} \mathbf{l}_{r;D} \circ \mathbf{k}_X \circ \sigma_A, \quad (6.33)$$

$$\sigma_A = (\widehat{\omega} n \sigma_X + \widehat{\omega} \mathbf{k}_X \circ \mathbf{k}_X + \widehat{\nu})^{-1}, \quad (6.34)$$

$$\mathbf{k}_X = \widehat{\omega} \sigma_X \circ \mathbf{k}_A \circ \mathbf{l}_{r;D}, \quad (6.35)$$

$$\sigma_X = (\widehat{\omega} p \sigma_A + \widehat{\omega} \mathbf{k}_A \circ \mathbf{k}_A + \mathbf{1}_{1,r})^{-1}, \quad (6.36)$$

$$\alpha_i = \alpha_0 + \frac{p}{2}, \quad i = 1, \dots, n, \quad (6.37)$$

$$\beta_i = \beta_0 + \frac{1}{2} (p \sigma_{A,i} + k_{A,i}^2), \quad i = 1, \dots, n, \quad (6.38)$$

$$\vartheta = \vartheta_0 + \frac{np}{2}, \quad (6.39)$$

$$\begin{aligned} \rho = \rho_0 + \frac{1}{2} & (\mathbf{l}'_D \mathbf{l}_D - 2\mathbf{l}'_{r;D} (\mathbf{k}_A \circ \mathbf{k}_X) + (\mathbf{k}_A \circ \mathbf{k}_X)' (\mathbf{k}_A \circ \mathbf{k}_X)), \\ & + \frac{1}{2} \widehat{\omega}^{-1} (p \sigma'_A (\mathbf{k}_X \circ \mathbf{k}_X) + pn \sigma'_A \sigma_X + n \sigma'_X (\mathbf{k}_A \circ \mathbf{k}_A)). \end{aligned} \quad (6.40)$$

Hence, equations (6.33)–(6.40) can be used as replacement for (3.76)–(3.83).

Note that elements of vectors σ and \mathbf{k} correspond in such a way that the i th element of one vector depends only on the i th elements of the remaining vectors, e.g.

$$k_{A;i} = \widehat{\omega} l_{D,i} k_{X,i} \sigma_{A,i}.$$

The only equation that makes them mutually dependent is (6.40), i.e. the expected value, $\widehat{\omega}$, of ω . If $\widehat{\omega}$ is known, the complexity of the problem is now reduced to the complexity of the toy problem (Section 6.1), which is analytically tractable.

Proposition 6.2 *Let the posterior expected value $\widehat{\omega}$ of ω be known. Then, equations (6.29)–(6.36) have an analytical solution with two modes:*

$$\text{the first one for } l_{D,i} \leq \frac{\sqrt{p} + \sqrt{n}}{\sqrt{\widehat{\omega}}} \sqrt{1 - \beta_0 \widehat{\omega}},$$

$$k_{A,i} = 0, \quad (6.41)$$

$$k_{X,i} = 0 \quad (6.42)$$

$$\sigma_{X,i} = \frac{1}{2} \frac{2n + -(n-p)\beta_0 \widehat{\omega} - \sqrt{\beta_0^2 \widehat{\omega}^2 (n-p)^2 + 4\beta_0 n p \widehat{\omega}}}{n(1 - \beta_0 \widehat{\omega})}, \quad (6.43)$$

$$\sigma_{A,i} = \frac{1 - \sigma_{X,i}}{\sigma_{X,i} p \widehat{\omega}}, \quad (6.44)$$

$$\beta_i = \frac{(\alpha_0 + \frac{1}{2}p)(\sigma_{X,i} - 1)}{(n(\sigma_{X,i} - 1) + p)\sigma_{X,i} \widehat{\omega}}, \quad (6.45)$$

and the second one for $l_{D,i} > \frac{\sqrt{p} + \sqrt{n}}{\sqrt{\hat{\omega}}} \sqrt{1 - \beta_0 \hat{\omega}}$,

$$k_{A,i} = z2, \quad (6.46)$$

$$k_{X,i} = \frac{(\hat{\omega} l_{D,i}^2 - p) k_{A,i}}{l_D (\hat{\omega} k_{A,i}^2 + 1)}, \quad (6.47)$$

$$\sigma_{A,i} = \frac{\hat{\omega} k_{A,i}^2 + 1}{\hat{\omega} (\hat{\omega} l_{D,i}^2 - p)}, \quad (6.48)$$

$$\sigma_{X,i} = \frac{(\hat{\omega} l_{D,i}^2 - p)}{\hat{\omega} l_{D,i} (\hat{\omega} k_{A,i}^2 + 1)} \quad (6.49)$$

$$\beta_i = \frac{(\alpha_0 + \frac{1}{2}p) (\hat{\omega} k_{A,i}^2 ((1 - \hat{\omega} \beta_0) (n - p) + l_{D,i}) + n + \beta_0 \hat{\omega}^2 l_{D,i}^2)}{\hat{\omega} k_{A,i}^2 (p - n) - n + \hat{\omega} l_{D,i}^2}, \quad (6.50)$$

where the expression denoted as $z2$ for $k_{A,i}$ is too long and can be found in Appendix B.

Proof: evaluated using the symbolic software package, Maple. See Appendix B for detailed analysis. ■

Conjecture 6.1 (Soft orthogonality constraints) Prior distributions on A and X , i.e. (3.65) and (3.66), were chosen with diagonal covariance matrices. This choice favours such matrices A and X whose product $A'A$ and $X'X$ is a diagonal matrix. Hence, the Variational Extreme (Corollary 3.1) converges to posterior distributions with orthogonal mean value, even if the VEM algorithm was initialized with non-orthogonal matrices.

Note that, using Proposition 6.2, the VEM algorithm can be greatly simplified. Substituting (6.41)–(6.45) into (6.33)–(6.38), the VB-statistics are determined up to the expected value of the precision (6.40). The VB-statistics can be evaluated via a simplified VEM algorithm with one degree of freedom, as follows:

Algorithm 6.1 (Fast VPCA)

1. SVD of the data matrix (3.52)
2. Choose initial value of $\hat{\omega}$ as $^l \hat{\omega} = \frac{n}{l_{p;D}}$.
3. Evaluate the breakpoint point:

$$\tilde{l}_D = \frac{\sqrt{p} + \sqrt{n}}{\sqrt{\hat{\omega}}} \sqrt{1 - \beta_0 \hat{\omega}}.$$

4. Split l_D into $l^{(1)} = \{l_{D,i} < \tilde{l}_D\}$ and $l^{(2)} = \{l_{D,i} \geq \tilde{l}_D\}$. In fact, the number of elements in $l^{(1)}$ determines the ARD property, and thus r_u .

5. Evaluate solutions of mode 1, (6.41)–(6.45), for $l^{(1)}$, and of mode 2, (6.46)–(6.50), for $l^{(2)}$.

¹This choice will be explained in Section 6.4.

6. Update estimate of $\hat{\omega}^{(t)} = \frac{\vartheta}{\rho}$, using (6.39) and (6.40).

7. If difference of $\hat{\omega}^{(t)} - \hat{\omega}^{(t-1)} > \text{threshold}$, go to 3.

Comparison with the standard model will be studied in Section 6.4.

6.3. Orthogonal Variational PCA (OVPCA)

In Section 3.3, we have shown that the ML estimation of parameters A and X of the PPCA model suffers from rotational ambiguity (Remark 3.6). It is a consequence of inappropriate modelling of low-rank matrix $M_{(r)}$, which is clearly over-parameterized. This is a complication in the Bayesian treatment, where the inference of A and X must be regularized via priors, as noted in Section 6.1.1.

However, from an analytical point-of-view, the model contains redundant parameters. For example, under the VB approximation, the posterior expected value of the mean,

$$E_{A,X}(M_{(r)}) = E_{A,X}(AX) = U_{r;D} \text{diag}(\mathbf{k}_{r;A} \circ \mathbf{k}_{r;X}) V'_{r;D},$$

is found in the SVD form, but the singular values are found as element-wise product of two vectors, \mathbf{k}_A and \mathbf{k}_X . Element-wise ratio of these two vectors is governed by the chosen priors. In this Section, we re-parameterize the model in a more compact way.

6.3.1. Orthogonal Parameterization of the PPCA Model

A standard tool for dealing with reduced-rank matrices is the ‘economic’ Singular Value Decomposition (SVD) [73]:

$$M_{(r)} = ALX'. \quad (6.51)$$

Since the rank r of the matrix $M_{(r)}$ is known², we can restrict matrices, A and X , to $\mathfrak{R}^{p \times r}$ and $\mathfrak{R}^{n \times r}$ respectively, with orthogonality restrictions $A'A = I_r$, $X'X = I_r$. Also $L = \text{diag}(\mathbf{l}) \in \mathfrak{R}^{r \times r}$ is a diagonal matrix of non-zero *singular values*, $\mathbf{l} = [l_1, \dots, l_r]'$, ordered, without loss of generality, as

$$l_1 > l_2 > \dots > l_r > 0. \quad (6.52)$$

The decomposition (6.51) is unique, up to the sign of the r singular vectors, (i.e. there are 2^r possible decompositions (6.51) satisfying the stated constraints, all equal to within a sign ambiguity).

Model (3.1), extended by (3.49), (6.51), yields:

$$f(D|A, L, X, \omega, r) = \mathcal{N}(ALX', \omega^{-1}I_p \otimes I_n). \quad (6.53)$$

To our knowledge, this model [107] has not been considered before in the literature. The maximum likelihood estimates of the model parameters, conditioned by known r , are given by

$$\left(\hat{A}, \hat{L}, \hat{X}, \hat{\omega} \right) = \arg \max_{A, L, X, \omega} f(D|A, L, X, \omega, r),$$

²At present we suppose it is known. This will be relaxed later.

with assignments

$$\hat{A} = U_{r;D}, \quad \hat{L} = L_{r,r;D}, \quad \hat{X} = V_{r;D}, \quad \hat{\omega} = \frac{pn}{\sum_{i=r+1}^p l_{D,i}^2}. \quad (6.54)$$

Here, $U_{r;D}$, and $V_{r;D}$ are the first r columns of the matrices U_D , and V_D of the SVD decomposition of the data matrix D (3.52) respectively, and $L_{r,r;D}$ is the $r \times r$ upper-left sub-block of matrix L_D .

6.3.2. Bayesian Formulation

The confinement, in the orthogonal model, ambiguity to only a sign-based ambiguity is an advantage gained at the expense of orthogonal restrictions which are generally difficult to handle. Specifically, parameters A and X are restricted to having orthonormal columns, i.e. $A'A = I_r$ and $X'X = I_r$ respectively. Intuitively, each column \mathbf{a}_i , $i = 1 \dots r$, of A belongs to the unit hyperball in p dimensions, i.e. $\mathbf{a}_i \in \mathcal{H}_p$. Hence, $A \in \mathcal{H}_p^r$, the Cartesian product of r p -dimensional unit hyperballs. However, the requirement of orthogonality—i.e. $\mathbf{a}_i' \mathbf{a}_j = 0$, $\forall i \neq j$ —confines the space further. The orthonormally constrained subset, $\mathcal{S}_{p,r} \subset \mathcal{H}_p^r$ is known as the Stiefel manifold [108]. Space $\mathcal{S}_{p,r}$ has finite area, which will be denoted as $\tau(p, r)$:

$$\tau(p, r) = \frac{2^r \pi^{\frac{1}{2}pr}}{\pi^{\frac{1}{4}r(r-1)} \prod_{j=1}^r \Gamma\left\{\frac{1}{2}(p-j+1)\right\}}, \quad (6.55)$$

where $\Gamma(\cdot)$ is the Gamma function [71]. Both the prior and posterior distributions have a support confined to $\mathcal{S}_{p,r}$.

We choose the priors on A and X to be the least informative, i.e. uniform on $\mathcal{S}_{p,r}$ and $\mathcal{S}_{n,r}$ respectively:

$$f(A) = \tau(p, r)^{-1} \chi(\mathcal{S}_{p,r}), \quad (6.56)$$

$$f(X) = \tau(n, r)^{-1} \chi(\mathcal{S}_{n,r}). \quad (6.57)$$

There is no upper bound on $\omega > 0$ (3.49). An appropriate prior is therefore (the improper) Jeffreys' prior on scale parameters [18]:

$$f(\omega) \propto \omega^{-1}. \quad (6.58)$$

Suppose that the sum of squares of elements of D is fixed, e.g.:

$$\sum_{i=p} \sum_{j=n} d_{i,j}^2 = \text{tr}(DD') = 1. \quad (6.59)$$

This can easily be achieved in a pre-processing step. (6.59) can be expressed, using (3.52), as:

$$\text{tr}(DD') = \text{tr}(U_D L_D L_D U_D') = \sum_{i=1}^p l_{D,i}^2 = 1.$$

This implies an upper bound on l :

$$\sum_{i=1}^r l_i^2 \leq \sum_{i=1}^p l_{D,i}^2 = 1. \quad (6.60)$$

This, together with (6.52), confines l to the space

$$\mathcal{L}_r = \left\{ l \mid l_1 > l_2 > \dots > l_r > 0, \sum_{i=1}^r l_i^2 \leq 1 \right\}, \quad (6.61)$$

which is a section of a unit hyperball. Constraint (6.60) forms a full hyperball \mathcal{H}_r , with volume

$$h_r = \pi^{\frac{r}{2}} / \Gamma\left(\frac{r}{2} + 1\right). \quad (6.62)$$

Positivity constraints restrict the allowed volume to $h_r/2^r$, and hyperplanes $\{l_i = l_j, \forall i, j = 1 \dots r\}$ partition the positive section of the hyperball into $r!$ sections with equal volume, only one of which satisfies condition (6.52). Hence, the volume of the support (6.61) is

$$\varphi_r = h_r \frac{1}{2^r (r!)} = \frac{\pi^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2} + 1\right) 2^r (r!)}.$$

Therefore, we choose the prior distribution on l to be non-committal—i.e. uniform—on support (6.61):

$$f(l) = \mathcal{U}(\mathcal{L}_r) = \varphi_r^{-1} \chi(\mathcal{L}_r). \quad (6.63)$$

Multiplying (6.53) by (6.56), (6.57), (6.58) and (6.63), and using the chain rule of probability, we obtain the joint distribution:

$$f(D, A, L, X, \omega | r) = \mathcal{N}(ALX, \omega^{-1} I_p \otimes I_t) \times \varphi_r^{-1} \tau(p, r)^{-1} \tau(n, r)^{-1}, \quad (6.64)$$

on support $\{A \in \mathcal{S}_{p,r}\} \times \{l \in \mathcal{L}_r\} \times \{X \in \mathcal{S}_{n,r}\} \times \{\omega > 0\}$.

The posterior distribution is then obtained using Bayes' rule:

$$f(A, L, X, \omega | D, r) = \frac{f(D, A, L, X, \omega | r)}{f(D | r)}. \quad (6.65)$$

Exact posterior inference from (6.64) is not available.

6.3.3. Variational Bayes (VB) Approximation

Corollary 6.2 (Corollary 5 of Theorem 2.1) Consider the following conditionally-independent factorization of (6.65):

$$\check{f}(A, L, X, \omega | D, r) = \check{f}(A | D, r) \check{f}(X | D, r) \check{f}(L | D, r) \check{f}(\omega | D, r) \quad (6.66)$$

Using (6.64) and (6.66) in Theorem 2.1, the VB-optimal form of (6.66) is found via the following assignments:

$$\tilde{f}(A|D, r) = \mathcal{M}(F_A), \quad (6.67)$$

$$\tilde{f}(X|D, r) = \mathcal{M}(F_X), \quad (6.68)$$

$$\tilde{f}(\mathbf{l}|D, r) = t\mathcal{N}(\boldsymbol{\mu}_l, s^2 I_r; \bar{\mathcal{L}}_r), \quad (6.69)$$

$$\tilde{f}(\omega|D, r) = \mathcal{G}(\vartheta, \rho). \quad (6.70)$$

Here, $\mathcal{M}(\cdot)$ denotes the von Mises-Fisher distribution (i.e. normal distribution restricted on the Stiefel manifold [108]). Their matrix parameters are $F_A \in \mathbb{R}^{p \times r}$ in (6.67), and $F_X \in \mathbb{R}^{n \times r}$ in (6.68). $t\mathcal{N}$ is the truncated Normal distribution with truncation points given by the lower and upper bounds of the prior (6.85).

The VB-statistics of (6.67)–(6.70) are:

$$F_A = \hat{\omega} D \hat{X} \hat{L}, \quad (6.71)$$

$$F_X = \hat{\omega} D' \hat{A} \hat{L}, \quad (6.72)$$

$$\boldsymbol{\mu}_l = \text{diag}^{-1}(\hat{X}' D' \hat{A}), \quad (6.73)$$

$$s^2 = \hat{\omega}^{-1}, \quad (6.74)$$

$$\vartheta = \frac{pn}{2}, \quad (6.75)$$

$$\rho = \frac{1}{2} \text{tr}(DD' - 2D\hat{X}\hat{L}\hat{A}') + \frac{1}{2} \hat{l}' \hat{l}. \quad (6.76)$$

These, in turn, are defined in terms of moments of distributions (6.67)–(6.70), namely \hat{A} , \hat{X} , \hat{l} , $\hat{l}' \hat{l}$ and $\hat{\omega}$. These are expressed via the SVD of parameters F_A (6.71) and F_X (6.72):

$$F_A = U_{F_A} L_{F_A} V_{F_A}', \quad (6.77)$$

$$F_X = U_{F_X} L_{F_X} V_{F_X}', \quad (6.78)$$

with L_{F_X} and L_{F_A} both in $\mathbb{R}^{r \times r}$. Then,

$$\hat{A} = U_{F_A} G(p, L_{F_A}) V_{F_A}', \quad (6.79)$$

$$\hat{X} = U_{F_X} G(n, L_{F_X}) V_{F_X}', \quad (6.80)$$

$$\hat{l} = \boldsymbol{\mu}_l + s \varphi(\boldsymbol{\mu}_l, s), \quad (6.81)$$

$$\hat{l}' \hat{l}_r = r s^2 + \boldsymbol{\mu}_l' \hat{l} - s \kappa(\boldsymbol{\mu}_l, s), \quad (6.82)$$

$$\hat{\omega} = \frac{\vartheta}{\rho}. \quad (6.83)$$

Moments of $\mathcal{M}(\cdot)$ and $t\mathcal{N}(\cdot)$ —from which (6.79)–(6.82) are derived—are reviewed in Appendices A.5 and A.4 respectively. Functions $G(\cdot, \cdot)$, $\varphi(\cdot, \cdot)$, and $\kappa(\cdot, \cdot)$ are also defined there.

Proof: Can be handled in the same way as proofs for the previous VB-related Corollaries. It is an easy but lengthy exercise in probability calculus. \blacksquare

Remark 6.2 (Approximate support for L) The correct distribution for l is

$$f(l|D, r) = t\mathcal{N}(\boldsymbol{\mu}_l, s^2 I_r; \mathcal{L}_r), \quad (6.84)$$

i.e. a Normal distribution truncated on support, \mathcal{L}_r (6.61). However, the moments of distribution (6.84) are difficult to evaluate, as \mathcal{L}_r forms a non-trivial section of the multivariate support. Therefore we approximate the support \mathcal{L}_r by its envelope $\mathcal{L}_r \approx \bar{\mathcal{L}}_r$. Note that (6.60) is maximized if $l_1 = l_2 = \dots = l_r$, $l_{r+1} = l_{r+2} = \dots = l_p = 0$. In this case $\sum_{i=1}^p l_i^2 = r l_r^2 < 1$, which defines an upper bound, $l_r < \bar{l}_r$ to be $\bar{l}_r = r^{-\frac{1}{2}}$. Hence, (6.61) has a rectangular envelope:

$$\bar{\mathcal{L}}_r = \left\{ l \mid 0 < l_i \leq \bar{l}_r = i^{-\frac{1}{2}}, \quad i = 1 \dots r \right\}. \quad (6.85)$$

(6.84) is then approximated by (6.69). Note that $\bar{\mathcal{L}}_r$ is a rectangular area. Hence, (6.69) can be written as the product of univariate truncated Normal distributions, moments of which are known (Appendix A.4). The error of approximation is largest at the boundaries, $l_i = l_j$, $i \neq j$, $i, j \in \{1 \dots r\}$, and is negligible when no two l_i 's are equal.

Once again, the general solution of the VEM algorithm (Algorithm 2.2) can be used. However, closer analysis of equations (6.71)–(6.83) reveals that the evaluation for our model can be simplified, as follows.

Proposition 6.3 (Orthogonal Variational PCA (OVPCA)) We search for a solution of \hat{A} (6.79) and \hat{X} (6.80) in the space of scaled singular vectors of matrix D (3.52):

$$\hat{A} = U_{r;D} K_A, \quad (6.86)$$

$$\hat{X} = V_{r;D} K_X. \quad (6.87)$$

U_D and V_D are given by (3.52). $K_A = \text{diag}(\mathbf{k}_A) \in \mathbb{R}^{r \times r}$ and $K_X = \text{diag}(\mathbf{k}_X) \in \mathbb{R}^{r \times r}$ denote constants of proportionality which must be determined. Then, each iteration using equations (6.71)–(6.83) will not leave this space: i.e. (6.86) and (6.87) are true at each iteration step.

Proof: Consider the t th iteration step, $t = 1, 2, \dots$, where superscript (t) denotes the optimized parameter values in this step. Assume that estimates, $\hat{A}^{(t-1)}$, $\hat{X}^{(t-1)}$, at the end of the previous step³, are of the form (6.86), (6.87); i.e.

$$\hat{A}_r^{(t-1)} = U_{r;D} K_A^{(t-1)}, \quad \hat{X}^{(t-1)} = V_{r;D} K_X^{(t-1)}. \quad (6.88)$$

Hence, von Mises-Fisher parameters F_A and F_X are updated, at iteration t , via (6.71) and (6.72) respectively:

$$F_A^{(t)} = \hat{\omega}^{(t-1)} (U_D L_D V_D') V_{r;D} K_X^{(t-1)} \hat{L}^{(t-1)} = \hat{\omega}^{(t-1)} U_{r;D} L_{r,r;D} K_X^{(t-1)} \hat{L}^{(t-1)}, \quad (6.89)$$

$$F_X^{(t)} = \hat{\omega}^{(t-1)} (U_D L_D V_D')' U_{r;D} K_A^{(t-1)} \hat{L}^{(t-1)} = \hat{\omega}^{(t-1)} V_{r;D} L_{r,r;D} K_A^{(t-1)} \hat{L}^{(t-1)}, \quad (6.90)$$

³Initial conditions, i.e. at $t = 0$, will be specified shortly.

using (3.52). These are in the SVD form of F_A (6.77), and F_X (6.78) respectively, with assignments:

$$U_{F_A} = U_{r;D}, \quad L_{F_A}^{(t)} = \widehat{\omega}^{(t-1)} L_{r,r;D} K_X^{(t-1)} \widehat{L}_r^{(t-1)}, \quad V_{F_A} = I_r, \quad (6.91)$$

$$U_{F_X} = V_{r;D}, \quad L_{F_X}^{(t)} = \widehat{\omega}^{(t-1)} L_{r,r;D} K_A^{(t-1)} \widehat{L}_r^{(t-1)}, \quad V_{F_X} = I_r. \quad (6.92)$$

Substituting (6.91) and (6.92) into (6.80) and (6.79) respectively:

$$\widehat{A}_r^{(t)} = U_{r;D} G \left(p, \widehat{\omega}^{(t-1)} L_{r,r;D} K_X^{(t-1)} \widehat{L}_r^{(t-1)} \right) I_r = U_{r;D} K_A^{(t)}, \quad (6.93)$$

$$\widehat{X}_r^{(t)} = V_{r;D} G \left(n, \widehat{\omega}^{(t-1)} L_{r,r;D} K_A^{(t-1)} \widehat{L}_r^{(t-1)} \right) I_r = V_{r;D} K_X^{(t)},$$

since function $G(\cdot, \cdot)$, with diagonal matrix argument, returns also a diagonal matrix (Appendix A.5.2). Therefore, new estimates remain of the same type (6.86), (6.87) with assignments:

$$K_A^{(t)} = G \left(p, \widehat{\omega}^{(t-1)} L_{r,r;D} K_X^{(t-1)} \widehat{L}_r^{(t-1)} \right), \quad (6.94)$$

$$K_X^{(t)} = G \left(n, \widehat{\omega}^{(t-1)} L_{r,r;D} K_A^{(t-1)} \widehat{L}_r^{(t-1)} \right). \quad (6.95)$$

■

Note that, under Proposition 6.3, the optimal values of \widehat{A} and \widehat{X} are determined up to the constants of proportionality, \mathbf{k}_A and \mathbf{k}_X . The iterative algorithm is then greatly simplified, since we need only iterate on the $2r$ degrees of freedom constituting K_A and K_X together, and not on \widehat{A} and \widehat{X} with $r(p + n - \frac{r-1}{2})$ degrees of freedom. To achieve this, we must, however, satisfy the requirement of Proposition 6.3, namely we must initialize the iterative scheme to satisfy (6.86) and (6.87), using any diagonal matrices K_A and K_X , with positive elements on their diagonals. In fact, for $K_A^{(0)} = K_X^{(0)} = I_r$, (6.86) and (6.87) are the ML solutions (6.54), and so an ML-initialized iteration is proposed, leading finally to the Orthogonal Variational PCA (OVPCA) algorithm.

Note that:

- initialization via the ML solution guarantees fast convergence to the unique solution, since (6.64) is likelihood-dominated by design.
- (6.71)–(6.83) now involve products of diagonal matrices. Equations (6.73)–(6.76), (6.94), and (6.95) can now be reformulated in efficient diagonal form.

The final OVPCA algorithm is as follows.

Algorithm 6.2 (OVPCA)

1. Initialize estimates using ML solution (6.54): i.e. $\mathbf{k}_A^{(0)} = \mathbf{k}_X^{(0)} = \mathbf{1}_{r,1}$, $\widehat{\mathbf{l}}^{(0)} = \mathbf{l}_{r;D}$, $\widehat{\omega}^{(0)} = \frac{pn}{\sum_{i=r+1}^p l_{D,i}^2}$.

2. Evaluate until convergence is reached:

$$\mathbf{k}_A^{(t)} = G\left(p, \widehat{\omega}^{(t-1)} \mathbf{l}_{r;D} \circ \mathbf{k}_X^{(t-1)} \circ \widehat{\mathbf{l}}^{(t-1)}\right), \quad (6.96)$$

$$\mathbf{k}_X^{(t)} = G\left(n, \widehat{\omega}^{(t-1)} \mathbf{l}_{r;D} \circ \mathbf{k}_A^{(t-1)} \circ \widehat{\mathbf{l}}^{(t-1)}\right), \quad (6.97)$$

$$\boldsymbol{\mu}_i^{(t)} = \mathbf{k}_X^{(t-1)} \circ \mathbf{l}_{r;D} \circ \mathbf{k}_A^{(t-1)}, \quad (6.98)$$

$$s^{(t)} = \left(\widehat{\omega}^{(t-1)}\right)^{-\frac{1}{2}}, \quad (6.99)$$

$$\widehat{\mathbf{l}}^{(t)} = \boldsymbol{\mu}_i^{(t-1)} + s^{(t-1)} \varphi\left(\boldsymbol{\mu}_i^{(t-1)}, s^{(t-1)}\right), \quad (6.100)$$

$$\widehat{\mathbf{l}}\widehat{\mathbf{l}}^{(t)} = \left(\boldsymbol{\mu}_i^{(t-1)}\right)' \widehat{\mathbf{l}}^{(t-1)} + r \left(s^{(t-1)}\right)^2 - s^{(t-1)} \kappa\left(\boldsymbol{\mu}_i^{(t-1)}, s^{(t-1)}\right)' \mathbf{1}_{1,r}, \quad (6.101)$$

$$\widehat{\omega}^{(t)} = pn \left[\mathbf{l}'_D \mathbf{l}_D - 2 \left(\mathbf{k}_X^{(t-1)} \circ \widehat{\mathbf{l}}^{(t-1)} \circ \mathbf{k}_A^{(t-1)} \right)' \mathbf{l}_{r;D} + \widehat{\mathbf{l}}\widehat{\mathbf{l}}^{(t-1)} \right]^{-1}. \quad (6.102)$$

Remark 6.3 (Automatic Rank Determination (ARD) Property of the OVPCA algorithm) It is observed that estimates of $k_{A,i}$ and $k_{X,i}$ typically converge to zero for $i > r_u$, for some empirical upper bound r_u . A similar property was used as a rank selection criterion for the OVPCA algorithm (Remark 3.8). There, the model order was chosen as $\hat{r} = r_u$ [7].

Remark 6.4 Equations (6.96)–(6.98) are satisfied for

$$\mathbf{k}_A = \mathbf{k}_X = \boldsymbol{\mu}_i = \mathbf{0}_{r,1}, \quad (6.103)$$

independently of data, i.e. independently of \mathbf{l}_D . The only parameter to be determined is ω . Solution (6.103) is appropriate for data formed only by realizations of homogeneous Gaussian noise without any signal, i.e. $r = 0$. This case will then be revealed by the ARD Property (Remark 6.3), i.e. r_u will be equal to zero. If the ARD Property yields a different estimate, i.e. $r_u \geq 1$, then solution (6.103) is a local maximum (or saddle point) and the true minimum of the KL distance has to be found by evaluation of (2.28) for both cases.

With respect to the original PCA (3.57) (Section 3.3), proposition 6.3 reveals an interesting analytical insight:

- Collinearity (6.86), (6.87) of the posterior mean with the respective ML estimate (i.e. PCA) means that uncertainty bounds on A are, in fact, uncertainty bounds on principal components $U_{r;D}$, (see Appendix A.5.3).
- We noted 2^r cases of SVD decomposition (6.51), distinct in terms of the signs of the singular vectors. Note, however, that Proposition 6.3 separates posterior mean values \widehat{A} (6.86) and \widehat{X} (6.87) into orthogonal and proportional parts. Only the latter (\mathbf{k}_A and \mathbf{k}_X) are estimated using the OVPCA algorithm (Algorithm 6.2). Since the function $G(\cdot, \cdot)$ is confined to the interval $[0, 1]$ (see Appendix A.5.2, Figure C.1), estimated values of \mathbf{k}_A and \mathbf{k}_X are always positive. In other words, the VB solution is unimodal, as though approximating only *one* of the possible 2^r modes. Due to symmetry of all modes, the solution is valid for all of them. This is important, as

the all-mode distribution of A is symmetric around the coordinate origin, which would consign the posterior mean to $\widehat{A} = \mathbf{0}_{p,r}$. Note that this symmetry is also reflected in the VB equations (6.96)–(6.102) (Remark 6.4).

- As a consequence of the ARD Property (Remark 6.3), the number of possible modes of the approximate posterior distribution is reduced to 2^{r_u} .

6.3.4. Inference of Rank

In the foregoing, we assumed that the rank, r , of the model (6.51) was known *a priori*. If this is not the case, then inference of this parameter can be made using Bayes' rule:

$$f(r|D) \propto f(D|r) f(r), \quad (6.104)$$

where $f(r)$ denotes the prior on r , typically uniform on $1 \leq r \leq p$. The marginal data posterior $f(D|r)$ can be approximated by a lower bound (Remark 2.3)

$$\begin{aligned} \ln f(D|r) &\approx \ln f(D|r) - KL\left(\tilde{f}(\theta|D, r) \parallel f(\theta|D, r)\right) \\ &= \int_{\Theta} \tilde{f}(\theta|D) \left(\ln f(D, \theta|r) - \ln \left(\tilde{f}(\theta|D, r) \right) \right) d\theta. \end{aligned} \quad (6.105)$$

The parameters are $\theta = \{A, L, X, \omega\}$, and $f(D, \theta)$ is given by (6.64). The optimal approximation, $\tilde{f}(\theta|D, r)$, is the conditionally independent model, obtained via the VB framework (6.67)–(6.73):

$$\tilde{f}(A, L, X, \omega|D) = \tilde{f}(A|D, r) \tilde{f}(L|D, r) \tilde{f}(X|D, r) \tilde{f}(\omega|D, r). \quad (6.106)$$

Substituting (6.67)–(6.73) into (6.106), and (6.64) into (6.105), then (6.104) yields:

$$\begin{aligned} f(r|D) &\propto \exp \left\{ -\frac{r}{2} \ln \pi + r \ln 2 + \ln \Gamma \left(\frac{r}{2} + 1 \right) + \ln (r!) \right. \\ &\quad + \frac{1}{2} s^{-2} \left(\boldsymbol{\mu}'_l \boldsymbol{\mu}_l - \widehat{\boldsymbol{l}}' \boldsymbol{\mu}_l - \boldsymbol{\mu}'_l \widehat{\boldsymbol{l}} + \widehat{\boldsymbol{l}}' \widehat{\boldsymbol{l}} \right) \\ &\quad + \ln {}_0F_1 \left(\frac{1}{2} p, \frac{1}{4} F_A F'_A \right) - \widehat{\omega} \left(\mathbf{k}_X \circ \widehat{\boldsymbol{l}} \circ \mathbf{k}_A \right)' \mathbf{l}_{r;D} \\ &\quad + \ln {}_0F_1 \left(\frac{1}{2} n, \frac{1}{4} F_X F'_X \right) - \widehat{\omega} \left(\mathbf{k}_X \circ \widehat{\boldsymbol{l}} \circ \mathbf{k}_A \right)' \mathbf{l}_{r;D} \\ &\quad + \sum_{j=1}^r \ln \left[\operatorname{erf} \left(\left(s\sqrt{2} \right)^{-1} \left(\overline{\bar{l}}_j - \mu_{j;l} \right) \right) + \operatorname{erf} \left(\left(s\sqrt{2} \right)^{-1} \mu_{j;l} \right) \right] \\ &\quad \left. + r \ln \left(s\sqrt{\pi/2} \right) - (\vartheta + 1) \ln \rho \right\}, \end{aligned} \quad (6.107)$$

where \mathbf{k}_A , \mathbf{k}_X , $\boldsymbol{\mu}_l$, $\widehat{\boldsymbol{l}}$, s and $\widehat{\omega}$ are the converged solutions of the OVPCA algorithm (Algorithm 6.2), and F_A and F_X are functions of these via (6.89) and (6.90) respectively. $\overline{\bar{l}}_j$ is the upper bound on support $\overline{\mathcal{L}}$ (6.85) of \boldsymbol{l} .

We note the following:

- One of the main algorithmic advantages of PCA is that a single evaluation of all p eigenvectors, i.e. U (3.54), provides with ease the PCA solution for any rank $r < p$, via the simple extraction of the first r columns, $U_{r;D}$ (3.52), of U_D . The OVPCA algorithm also enjoys this property, thanks to the linear dependence of solution (6.86) on U_D (3.57). Furthermore, V_D observes the same property. Therefore, in the OVPCA procedure, the optimal solution for given rank is obtained by simple extraction of $U_{r;D}$ and $V_{r;D}$, followed by iterations involving only scaling coefficients, \mathbf{k}_A and \mathbf{k}_X . Hence, $p \times (p + n)$ values (those of U_D and V_D) are determined rank-independently via the ML solution, and only $4r + 2$ values (those of \mathbf{k}_A , \mathbf{k}_X , $\boldsymbol{\mu}_l$, \hat{l} , s and $\hat{\omega}$ together) are involved in the rank-dependent iterations (6.96)–(6.102).
- As a consequence of the Automatic Rank Determination (ARD) property, Remark 6.3, values of all parameters, A, L, X, ω , inferred by the OVPCA algorithm, are almost identical for $r \geq r_u$. Therefore, it is reasonable to evaluate the OVPCA parameters for $r = p - 1$ (we cannot use $r = p$ because $\hat{\omega}_p$ is not valid (6.54)), and approximate $\hat{A}_r \approx \hat{A}_{p-1}$, $\forall r \geq r_u$ (and similarly for L, X and ω). This approximation can significantly reduce the number of runs of the OVPCA algorithm required for evaluation of $f(r|D)$ (6.107).
- The explicit posterior distribution on r , i.e. (6.107), was not provided by previously published approaches [4, 7]. In its place, the ARD property of *their* algorithms was used to infer rank. Since the OVPCA algorithm also possesses the ARD property (Remark 6.3), it will be used for comparison with the formal Bayesian solution (6.107) in the simulation studies that follow (Section 6.4).

6.3.5. Moments of the Model Parameters

The Bayesian solution provides an approximate posterior distribution of all involved parameters (6.67)–(6.70), and (6.107). Moments and uncertainty bounds are then inferable from these distributions.

The first moments of all involved parameters have already been presented, (6.80)–(6.81) and (6.83), since they are required by the OVPCA algorithm (Algorithm 6.2). The second non-central moment of l —i.e. $\hat{l}l$ —was also produced by the algorithm. Parameter ω is Gamma distributed (6.70), and so its confidence intervals are therefore available.

The difficult task is to determine uncertainty bounds on orthogonal parameters, A and X , which are von Mises-Fisher distributed (6.67), (6.68). To our knowledge, confidence intervals on this distribution are not published. Therefore, we develop approximate uncertainty bounds in Appendix A.5.3, using a maximum entropy (i.e. Gaussian-based) approach. The pdf of $X \in \mathbb{R}^{n \times r}$ is fully determined by the r -dimensional vector \mathbf{y}_X :

$$\mathbf{y}_X(X) = \text{diag}^{-1}(U'_{F_X} X V_{F_X}) = \text{diag}^{-1}(V'_{r;D} X). \quad (6.108)$$

Therefore, confidence intervals on X can be mapped to confidence intervals on \mathbf{y}_X by (6.108), as shown in Appendix A.5.2. The idea is illustrated graphically for $p = 2$ and $r = 1$ in Figure 6.6.

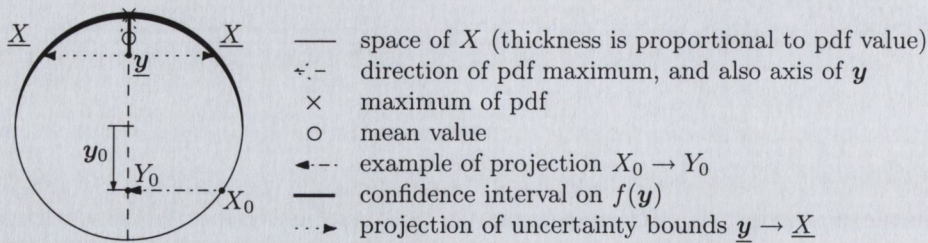


Figure 6.6.: Illustration of properties of von-Mises-Fisher distribution $X \sim \mathcal{M}(F)$, for $X, F \in \mathbb{R}^{2 \times 1}$.

Hence, lower and upper uncertainty bounds on X can be defined as follows:

$$\underline{X} = \{X : \mathbf{y}_X(X) = \underline{\mathbf{y}}_X\}, \quad (6.109)$$

$$\overline{X} = \{X : \mathbf{y}_X(X) = \overline{\mathbf{y}}_X\}, \quad (6.110)$$

using (6.108) and $\underline{\mathbf{y}}, \overline{\mathbf{y}}$ given in Appendix A.5.3, being bounds to the approximating Gaussian distribution of \mathbf{y}_X (A.42). In other words, uncertainty bounds (6.109) on X are those values of X that are projected onto the boundary of the confidence interval for \mathbf{y}_X .

Since A has the same distribution, \overline{A} and \underline{A} , are analogous.

6.4. Simulation Studies

In this section, we study properties of the algorithms described above in the context of simulated data. Artificial data were generated using model (3.2) for $p = 10$, $n = 100$, and $r = 3$. Simulated data are displayed in Figure 6.7. Three noise variances were considered: (i) $\omega = 100$, denoted as SIM1 (ii) $\omega = 25$, (SIM2) and (iii) $\omega = 10$, (SIM3).

6.4.1. VPCA vs. FVPCA

In Section 6.2, we have presented numerically efficient algorithm (FVPCA, Algorithm 6.1) for evaluation of VB-statistics (3.72)–(3.75) for Variational PCA (Corollary 3.1). A significant simplification of the algorithm was achieved by confining the space of possible solution to an orthogonal subspace, as formalized by Proposition 6.1. It is assumed that the chosen priors confine the space in such a way that the solution is found within this orthogonal subspace (Conjecture 6.1). The purpose of this experiment is to verify the validity of this Conjecture by simulation.

If Conjecture 6.1 is true, then the posterior moments, $\widehat{\mu}_X$ (3.78) and $\widehat{\mu}_A$ (3.76), converge to orthogonal (but *not* orthonormal) matrices for each possible initialization. A Monte Carlo study of 100 runs of the original VPCA algorithm (Section 3.3.3) was performed with random initial conditions. During the iterations, we tested orthogonality of the expected values of matrix X , via the assignment

$$O(\mu_X) = \|\widehat{\mu}_X' \widehat{\mu}_X\|,$$

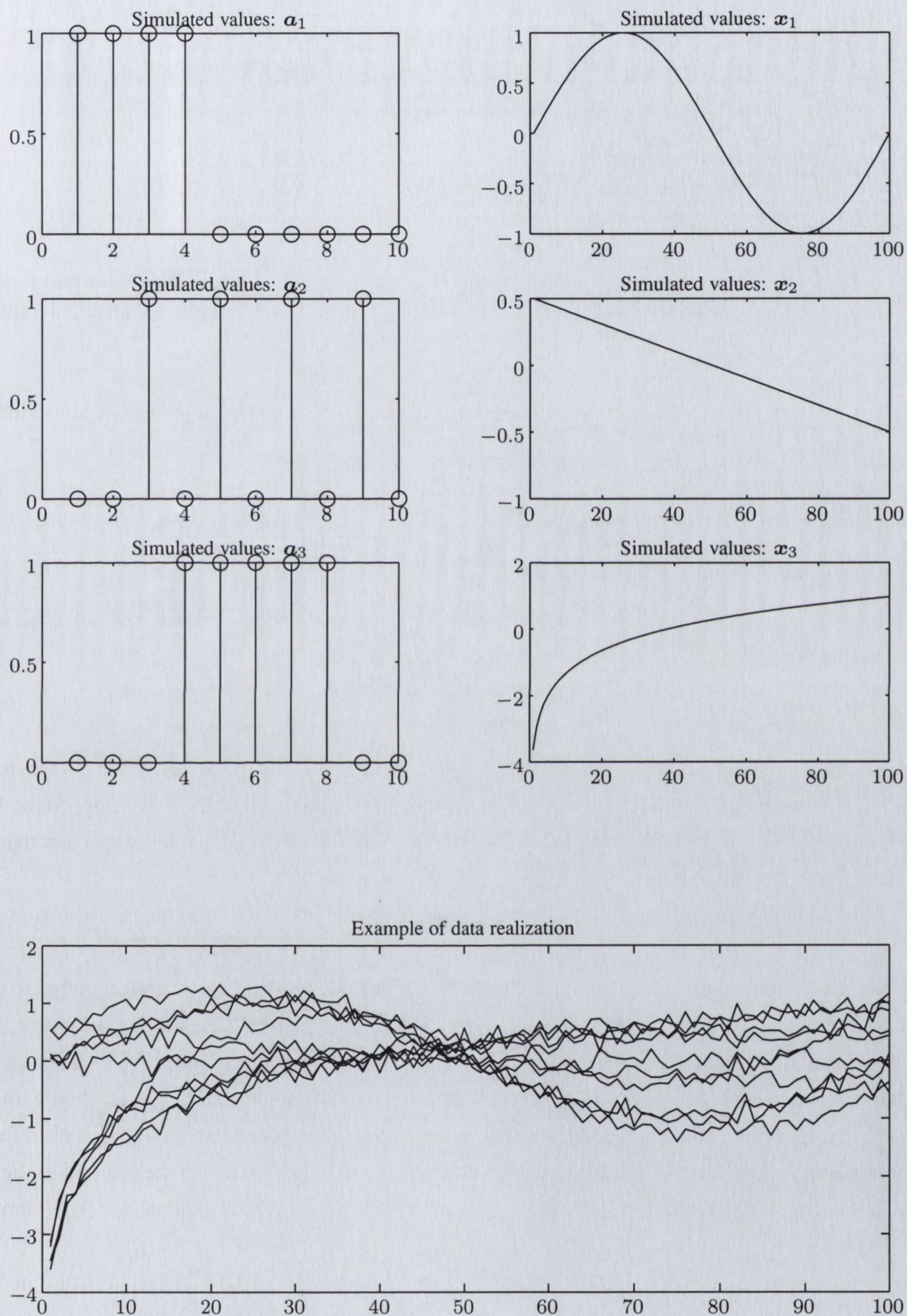


Figure 6.7.: Simulated data used for testing of PCA-based inference method.

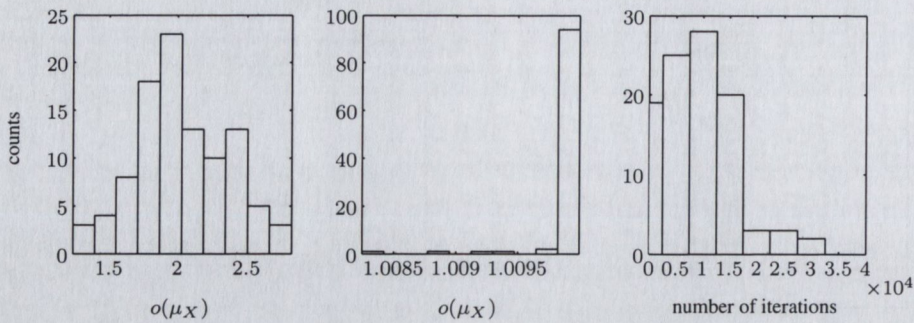


Figure 6.8.: Monte Carlo study (100 trials) to illustrate convergence of the VPCA algorithm to an orthogonal solution. **Left:** initial values of $o(\mu_X)$ (6.111). **Middle:** converged values of $o(\mu_X)$. **Right:** number of iterations required for convergence.

Table 6.1.: Comparison of converged values of k_A obtained from VPCA and FVPCA algorithms.

	VPCA, median	FVPCA
$k_{A,1}$	9.989	9.985
$k_{A,2}$	9.956	9.960
$k_{A,3}$	8.385	8.386

where $\|A\| = [|a_{i,j}|], \forall i, j$, denotes absolute value of a matrix applied element-wise. A criterion of diagonality is then

$$o(\mu_X) = \frac{\mathbf{1}'_{n,1} O(\mu_X) \mathbf{1}_{n,1}}{\mathbf{1}'_{n,1} \text{diag}^{-1}(O(\mu_X))}, \quad (6.111)$$

i.e. the ratio of the sum of all elements of $O(\mu_X)$ over the sum of its diagonal elements. Obviously, $o(\mu_X) = 1$ for a diagonal matrix, and $o(\mu_X) > 1$ for a non-diagonal matrix, $\widehat{\mu}_X$. We stopped the VEM algorithm for $o(\mu_X) < 1.01$, i.e. when the absolute value of non-diagonal elements was less than one percent of the diagonal elements.

In all simulated cases, this level was reached, though it took many iteration steps. Results are displayed in Figure 6.8: histograms of the orthogonality criterion (6.111) for initial values of matrix $\widehat{\mu}_X$ (left), and for its converged value (middle). The histogram of the number of iterations required to reach the stopping rule is displayed in Figure 6.8 (right). The middle picture seems to be redundant, since level $o(\mu_X) < 1.01$ was used as the stopping rule. This criterion is, however, important for comparison with the original proposal (threshold on moments of ω). The Variational Extreme is reached with high accuracy, even for non-orthogonal solutions (i.e. $o(\mu_X)$ above the chosen threshold). However, further iterations clearly push the solution towards the orthogonal one. This illustrates the flatness of the posterior distribution (see illustration for the toy problem Figure 6.2) and how expensive it is to iterate on the full space.

For comparison, values of k_A from VPCA⁴ and FVPCA for the SIM1 data set are listed in Table 6.1. Clearly, values of k_A obtained by VPCA and FVPCA converge to the same values. Tables of the remaining values (k_X, σ_A, σ_X) are not shown for conciseness.

⁴Values of VPCA are sorted here, since the method produces elements of k_A in random order.

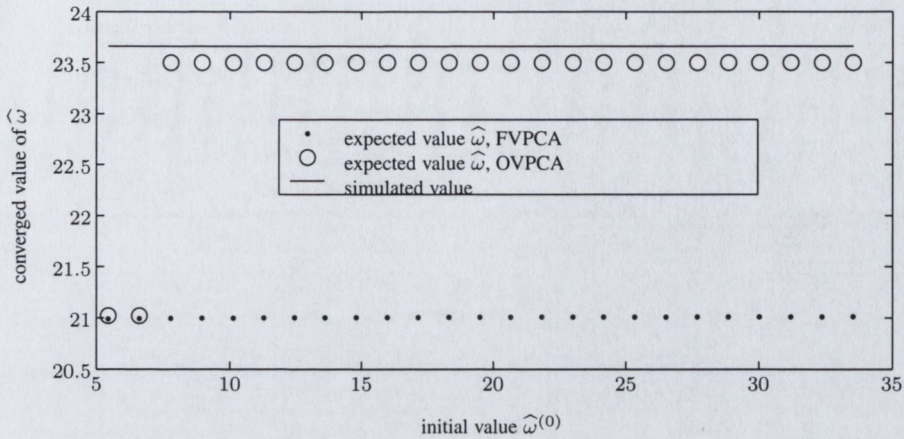


Figure 6.9.: Posterior estimates, $\hat{\omega}$, for the data set SIM2 with respect to different initial values $\hat{\omega}^{(0)}$.

Thus, we have demonstrated that the VPCA algorithm converges to the same values as FVPCA. Hence, in subsequent studies, we will consider FVPCA as a replacement for VPCA.

6.4.2. Initial conditions

Note that both algorithms, FVPCA and OVPCA, have to be initialized by expected value of ω , i.e. $\hat{\omega}^{(0)}$. In this Section, we study sensitivity of both algorithms to this choice. From asymptotic properties of PCA (Remark 3.7), we have a reasonable guess of the interval in which to search for $\hat{\omega}$:

$$\frac{np}{l'_D l_D} < \hat{\omega} < \frac{n}{l^2_{p,D}}. \quad (6.112)$$

We have tested initialization of both algorithms using values from the interval (6.112).

The posterior results for the data sets, SIM1 and SIM3, converged to the same value for all initial conditions in interval (6.112). However, for the data set SIM2, the results of both algorithm differ, as is displayed in Figure 6.9. Note that values of FVPCA are robust with respect to chosen initial conditions, but OVPCA results have two different modes: (i) the first two values (almost identical with FVPCA), and (ii) the majority of the interval, (very close to the simulated value). These two modes correspond to different values of the ARD property. FVPCA estimates $r_u = 2$, while OVPCA results are $r_u = 3$ or $r_u = 2$ for different initializations (6.9).

The basic idea of the VB approximation is minimization of the KL distance (Section 2.2.4). Hence, the value of the KL distance (2.25) for each mode can be used to choose the global minimum. In this case, the highest value of KL distance typically corresponds to the mode with highest precision (i.e. with the lowest variance). Hence, we choose to initialize the VEM algorithm with the upper bound

$$\hat{\omega}^{(0)} = \frac{n}{l^2_{p,D}}. \quad (6.113)$$

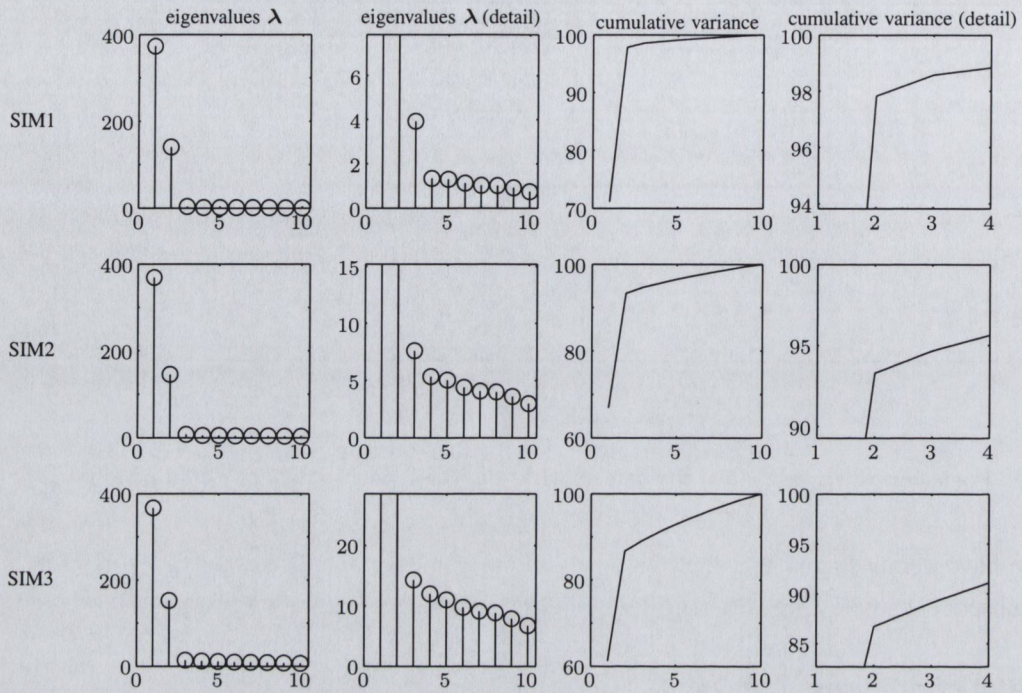


Figure 6.10.: *Ad hoc* methods for rank estimation for simulated data with different variance of noise. The methods of visual examination (Remark 3.7) is used for graphs of eigenvalues $\lambda = \text{eig}(DD')$ and cumulative variance (i.e. cumulative sum of eigenvalues λ).

6.4.3. Comparison of Methods for Inference of Rank

In this Section, we study the rank estimation properties of FVPCA and OVPCA. Note that results of both methods depend only on singular values l_D of the data matrix D .

The true dimensionality of the simulated data is $r = 3$. Many heuristic methods for choice of the number of relevant principal components are used in practice [80]. These methods are valuable, since they provide an intuitive insight into the problem. For example, the eigenvalues λ (3.57) of the simulated data, and the criterion of cumulative variation (i.e. cumulative sum of eigenvalues) are displayed in Figure 6.10.

Note that the first two eigenvectors are dominant (first column), while the third one is relatively small (it contains only 1% of total variation, see Figure 6.10 (right)). In the first row, i.e. case SIM1, the third eigenvalue is clearly distinct from the remaining ones. In the second case (SIM2), the difference is not very obvious, and it is completely lost in the third row (SIM3). A subjective *ad hoc* choice of dimensionality using visual inspection (Remark 3.7) and the method of cumulative variance is summarized in Table 6.2.

Next, we analyze the same data using formal methods. Results of FVPCA, OVPCA and Laplace approximation (Remark 3.9) are compared in Table 6.3.

Table 6.2.: Rank selection for the simulated data using *ad hoc* methods.

	SIM1	SIM2	SIM3
visual inspection	3	2-3	2
cumulative variance	2-3	2	2

Table 6.3.: Comparison of formal rank selection methods for simulated data.

	FVPCA	OVPCA					Laplace			
		ARD	$f(r D), r =$				$f(r D), r =$			
			2	3	4	5	2	3	4	5
SIM1	3	3	0	98.2	1.7	0.1	0	82	13	2
SIM2	2	3	96	3.5	0.2	0	70	25	3	0.5
SIM3	2	2	97	3.9	0.1	0	94	5	0.5	0.0

values of $f(r|D)$ not shown in the table are very close to zero < 0.001

Note that for high signal-to-noise ratio, all methods estimated the true dimensionality correctly. In this case, data were simulated according to the model. We therefore regard the results of all methods to be correct. The differences between posterior probabilities caused by different approximations are, in this case, insignificant. The differences will, however, become important for real data.

6.4.4. Comparison of Moments

A direct comparison of parameter moments for both methods is not possible. Therefore, we seek such a transformation of model parameters into a common, low dimensional, space in which it is possible to compare FVPCA and OVPCA. Note that the signal, $M_{(r)}$ (3.48), is (for both methods) a product of parameters, posterior distributions of which are all conditionally independent. Hence, the expected values (under different parameterization) of $M_{(r)}$ is:

$$\begin{aligned} E_{FVPCA}(M_{(r)}) &= \widehat{A}\widehat{X}' = U_{r;D}K_AK_XV'_{r;D}, \\ E_{OVPCA}(M_{(r)}) &= \widehat{A}\widehat{L}\widehat{X}' = U_{r;D}K_A\text{diag}(\boldsymbol{\mu}_l)K_XV'_{r;D}. \end{aligned}$$

Where K_A of the FVPCA method (6.29) is different from K_A of the OVPCA method (6.88). However, no confusion can arise, since these quantities are always used independently in their own context, i.e. all results related to FVPCA were evaluated using (6.29). Note that, for both methods, the expected values of $M_{(r)}$ are determined by singular vectors $U_{r;D}$ and $V_{r;D}$. This motivates us to introduce a transformed variable:

$$\tilde{M}_{(r)} = U'_{r;D}M_{(r)}V_{r;D} \in \mathfrak{R}^{r \times r}. \quad (6.114)$$

Using, linearity of expectations of the Matrix Normal distribution ((A.3) Appendix A.1) and invariance of von-Mises-Fisher distribution under orthogonal transformation (Appendix A.5), the expected values of $\tilde{M}_{(r)}$, under both methods, are:

$$\begin{aligned} E_{FVPCA}(\tilde{M}_{(r)}) &= K_AK_X, \\ E_{OVPCA}(\tilde{M}_{(r)}) &= K_A\text{diag}(\boldsymbol{\mu}_l)K_X. \end{aligned} \quad (6.115)$$

All elements on the right-hand side above are all diagonal matrices. Therefore, we can compare performance of FVPCA and OVPCA using moments of

$$\boldsymbol{\mu}_{\tilde{M}} = \text{diag}^{-1} \left(\tilde{M}_{(r)} \right) = \text{diag}^{-1} \left(U'_{r;D} M_{(r)} V_{r;D} \right).$$

We evaluate it separately for both methods.

FVPCA: The i th element $\mu_{i;\tilde{M}}$ of $\boldsymbol{\mu}_{\tilde{M}}$ is:

$$\mu_{i;\tilde{M}} = \mathbf{u}'_{i;D} \left(\sum_{j=1}^r \mathbf{a}_j \mathbf{x}'_j \right) \mathbf{v}_{i;D}, \quad i = 1, \dots, r \quad (6.116)$$

where $\mathbf{u}_{i;D}$, $\mathbf{v}_{i;D}$, \mathbf{a}_i , \mathbf{x}_i denotes the i th vector of matrices U_D , V_D , A and X , respectively. The first moment, i.e. expected values of (6.116), is then

$$\begin{aligned} \hat{\mu}_{i;\tilde{M}} &= E_{\tilde{f}(A|D)\tilde{f}(X|D)} \left(\mathbf{u}'_{i;D} \left(\sum_{j=1}^r \mathbf{a}_j \mathbf{x}'_j \right)' \mathbf{v}_{i;D} \right), \\ &= \mathbf{u}'_{i;D} \left(\sum_{j=1}^r \mathbf{u}_{j;D} k_{j;A} \mathbf{v}'_{j;D} k_{j;X} \right) \mathbf{v}_{i;D}, \\ &= k_{i;A} k_{i;X} \quad i = 1, \dots, r, \end{aligned} \quad (6.117)$$

using (6.29). The second equality in (6.117) follows from orthogonality of singular vectors $\mathbf{u}_i \mathbf{u}_j = 0$, $i \neq j$. The second non-central moment is:

$$\hat{\mu}^2_{i;\tilde{M}} = E_{\tilde{f}(A|D)\tilde{f}(X|D)} \left(\left[\mathbf{u}'_{i;D} \left(\sum_{j=1}^r \mathbf{a}_j \mathbf{x}'_j \right) \mathbf{v}_{i;D} \right]^2 \right), \quad (6.118)$$

$$= E_{\tilde{f}(A|D)\tilde{f}(X|D)} \left(\sum_{j=1}^r [\mathbf{u}'_{i;D} \mathbf{a}_j]^2 [\mathbf{x}'_j \mathbf{v}_{i;D}]^2 \right) \quad (6.119)$$

$$\begin{aligned} &= \sum_{j=1}^r E_{\tilde{f}(A|D)} \left(\mathbf{u}'_{i;D} \mathbf{a}_j \mathbf{a}'_j \mathbf{u}_{i;D} \right) E_{\tilde{f}(X|D)} \left(\mathbf{v}'_{i;D} \mathbf{x}_j \mathbf{x}'_j \mathbf{v}_{i;D} \right) \\ &= \sum_{j=1}^r \mathbf{u}'_{i;D} \left(\sigma_{j;A} I_p + k_{j;A}^2 \mathbf{u}_{j;D} \mathbf{u}'_{j;D} \right) \mathbf{u}_{i;D} \mathbf{v}'_{i;D} \left(\sigma_{j;X} I_n + k_{j;X}^2 \mathbf{v}_{j;D} \mathbf{v}'_{j;D} \right) \mathbf{u}_{i;D}, \\ &= \left(\sigma_{i;A} + k_{i;A}^2 \right) \left(\sigma_{i;X} I_n + k_{i;X}^2 \right) + \sum_{j=1, j \neq i}^r \sigma_{j;A} \sigma_{j;X}, \end{aligned} \quad (6.120)$$

The result was achieved using properties of Matrix Normal distribution (Appendix A.1)—specifically linearity of expectation (A.3), and second moment (A.2)—and orthogonality of singular vectors $U_{r;D}$ and $V_{r;D}$.

OVPCA: The i th element $\mu_{i;\tilde{M}}$ of $\boldsymbol{\mu}_{\tilde{M}}$ is:

$$\mu_{i;\tilde{M}} = \mathbf{u}'_{i;D} \left(\sum_{j=1}^r l_j \mathbf{a}_j \mathbf{x}'_j \right) \mathbf{v}_{i;D}, \quad i = 1, \dots, r \quad (6.121)$$

The first moment, i.e. expected value of (6.121), is then

$$\begin{aligned} \hat{\mu}_{i;\tilde{M}} &= E_{\tilde{f}(A|D)\tilde{f}(U|D)\tilde{f}(X|D)} \left(\mathbf{u}'_{i;D} \left(\sum_{j=1}^r l_j \mathbf{a}_j \mathbf{x}_j \right)' \mathbf{v}_{i;D} \right), \\ &= \mathbf{u}'_{i;D} \left(\sum_{j=1}^r l_j \mathbf{u}_{j;D} k_{j;A} \mathbf{v}'_{j;D} k_{j;X} \right) \mathbf{v}_{i;D}, \\ &= \hat{l}_{i;l} k_{i;A} k_{i;X} \quad i = 1, \dots, r, \end{aligned} \quad (6.122)$$

using (6.86), (6.87) and (6.81). The second non-central moment can be derived as follows:

$$\begin{aligned} \widehat{\mu^2}_{i;\tilde{M}} &= E_{\tilde{f}(A|D)\tilde{f}(U|D)\tilde{f}(X|D)} \left(\left[\mathbf{u}'_{i;D} \left(\sum_{j=1}^r l_j \mathbf{a}_j \mathbf{x}'_j \right) \mathbf{v}_{i;D} \right]^2 \right), \\ &= E_{\tilde{f}(A|D)\tilde{f}(U|D)\tilde{f}(X|D)} \left(\sum_{j=1}^r l_j^2 [\mathbf{u}'_{i;D} \mathbf{a}_j]^2 [\mathbf{x}'_j \mathbf{v}_{i;D}]^2 \right), \\ &= \sum_{j=1}^r \hat{l}^2_j E_{\tilde{f}(A|D)} (\mathbf{u}'_{i;D} \mathbf{a}_j \mathbf{a}'_j \mathbf{u}_{i;D}) E_{\tilde{f}(X|D)} (\mathbf{v}'_{i;D} \mathbf{x}_j \mathbf{x}'_j \mathbf{v}_{i;D}), \end{aligned} \quad (6.124)$$

here, we note that the second moment of the von-Mises-Fisher distribution is not available and we have chosen to approximate it by a Gaussian using maximum entropy principle (Appendix A.5.3). Therefore, evaluation of the expected values in (6.124) is equivalent to derivations for the FVPCA that follow after (6.119). Hence,

$$\widehat{\mu^2}_{i;\tilde{M}} = \hat{l}^2_i (\phi_{i;A} + k_{i;A}^2) (\phi_{i;X} I_n + k_{i;X}^2) + \sum_{j=1, j \neq i}^r \hat{l}^2_j \phi_{j;A} \phi_{j;X}, \quad (6.125)$$

where $\phi_{j;A}$ is a second moment of the von-Mises-Fisher distribution (6.67), evaluated via (A.40), Appendix A.5.3.

The uncertainty bounds on $\hat{\boldsymbol{\mu}}_{\tilde{M}}$ can be—for both methods—evaluated as 2-standard deviation interval, i.e.

$$\begin{aligned} \underline{\mu}_{i;\tilde{M}} &= \hat{\mu}_{i;\tilde{M}} - 2\sqrt{\widehat{\mu^2}_{i;\tilde{M}} - \hat{\mu}_{i;\tilde{M}}^2}, \\ \bar{\mu}_{i;\tilde{M}} &= \hat{\mu}_{i;\tilde{M}} + 2\sqrt{\widehat{\mu^2}_{i;\tilde{M}} - \hat{\mu}_{i;\tilde{M}}^2}, \end{aligned} \quad (6.126)$$

Table 6.4.: Comparison of inference of the diagonal $\mu_{\tilde{M}}$ of the transformed signal \tilde{M} (6.114), using FVPCA and OVPCA.

row number i	simulated	FVPCA			OVPCA		
		$\underline{\mu}_{i;\tilde{M}}$	$\hat{\mu}_{i;\tilde{M}}$	$\bar{\mu}_{i;\tilde{M}}$	$\underline{\mu}_{i;\tilde{M}}$	$\hat{\mu}_{i;\tilde{M}}$	$\bar{\mu}_{i;\tilde{M}}$
1	0.825	0.780	0.806	0.833	0.793	0.811	0.829
2	0.492	0.469	0.495	0.521	0.481	0.499	0.517
3	0.038	-0.001	0.000	0.001	0.023	0.035	0.046
4	0.000	-0.001	0.000	0.001	0.000	0.000	0.004
5	0.003	-0.001	0.000	0.001	0.000	0.000	0.004
6	0.002	-0.001	0.000	0.001	0.000	0.000	0.004
7	0.000	-0.001	0.000	0.001	0.000	0.000	0.004
8	0.002	-0.001	0.000	0.001	0.000	0.000	0.004
9	0.002	-0.001	0.000	0.001	0.000	0.000	0.004

where (6.117) and (6.120) are used for FVPCA, and (6.122) and (6.122) are used for OVPCA. The results are summarized in Table 6.4. Note that the space $\mu_{\tilde{M}}$ —on which we compare the methods—is determined by singular vectors ($U_{r;D}$ and $V_{r;D}$) of the data matrix D (3.52), which are not orthogonal with singular vectors of the simulated signal $M_{(r)}$. Hence, diagonal of the projected—via (6.114)—simulated signal contains non-zero values. These values should also be within uncertainty bounds (6.126).

As may be seen of Table 6.4, all projections of simulated values are within HPD regions for OVPCA. For FVPCA, the projected values are outside of the HPD regions for $i > 2$. The value $\mu_{3;\tilde{M}}$ is outside because of inaccurate estimation of r_u , and those for $i \geq 4$ are outside because uncertainty bounds for $i > r_u$ are too tight. It is worth noticing that these uncertainty bounds are dependent on hyper-parameter, β_0 , which was chosen very low in this simulation. OVPCA has no corresponding hyper-parameter, which—together with positivity constraints on l —yields more reliable results.

6.5. Discussion

In this Chapter, we have studied the Bayesian approach to Principal Component Analysis. Most of the published solutions, e.g. [7, 50] were based on the Probabilistic PCA model [48]. We have studied the simplest case of this model in Section 6.1.3, which we called the ‘toy problem’, and we compared the approximations available using published methods. The insight gained from this study (Section 6.1.5) has been explored in the full multivariate model, yielding two distinct algorithms: (i) fast evaluation of the VB-statistics of the VPCA (FVPCA algorithm) (Section 6.2); and (ii) the Variational approximation for the orthogonal parameterization (OVPCA) (Section 6.3). Both algorithms provides numerically efficient inference of the corresponding model parameters. We note the following differences:

- The orthogonal model (6.51) is a more compact parameterization of the signal $M_{(r)}$, as it does not require any regularizing prior distributions to be elicited. On the other hand, for FVPCA, the hyper-parameters coming from the priors must be used to regularize the model and find a solution.

- The orthogonal parameterization of the signal, $M_{(r)}$, is appropriate only if there are no other constraints imposed on the decomposition. For example, orthogonal decomposition of the signal under positivity constraints (Section 3.4) is not possible. Therefore, OVPCA can only be used for the task of denoising.
- The OVPCA algorithm converges faster (compared to the corresponding FVPCA algorithm) and the stopping rule on increments of parameter estimates (e.g. $\hat{\omega}$) can be set close to the machine precision. Convergence of the FVPCA algorithm is slower and it is sensitive to the choice of the stopping rule.
- The most numerically demanding operation in the FVPCA algorithm is evaluation of roots of a quadratic function which is an operation well supported by standard software. The OVPCA algorithm requires evaluations of hypergeometric functions of multivariate arguments, for which a reliable solution is not available, and which must be approximated (as presented in Appendix C).

From a practical (user-oriented) point of view, the FVPCA and OVPCA algorithms are almost equivalent. They yield comparable results (see Table 6.4) at comparable computational cost. The OVPCA algorithm appears to be more reliable on data with low signal-to-noise ratio (Table 6.4).

Although the orthogonal model (6.51) is a better parameterization—in the sense that it provides model-based regularization of the problem—it is complicated to extend it further, e.g. for the noise distribution as used in the factor analysis model (3.6). Formulation of the problem is straightforward, and so is the application of the Variational Bayes estimation method in this case. However, the resulting posterior distributions are of the generalized Bingham type [108], whose moments are not known to us. This suggests that efficient numerical evaluation—possible for OVPCA—cannot be achieved for the factor analysis model.

The original probabilistic PCA model [48] is based on the full factor analysis model [47], which explicitly includes a common non-zero mean value for the data columns, $E(d_i) = \mu$. This was not considered as a part of the linear model (3.2) in this thesis. It is easy to introduce a common mean value for applications where it is regarded as important, e.g. for mixtures of PCA model [109]. The model (3.2) can be readily extended to contain the common mean value, as follows:

$$M = AX + \mu \mathbf{1}_{1,n},$$

with $\mu \in \mathbb{R}^{p \times 1}$. The Variational Bayes (VB) approximation (Theorem 2.1) for this model requires more algebra, but is straightforward. However, Proposition 6.1 is valid only for fixed estimate $\hat{\mu}$. Hence, the VEM algorithm associated with this model has much higher computational complexity.

The orthogonal parameterization of the signal $M_{(r)}$ is appropriate only if there are no other

considering imposed on the decomposition. For example, orthogonal decomposition of the signal

under positivity constraints (Section 3.4) is not possible. Therefore, OVPFA can only be used for the task of denoising. $M_{(r)}$ is available, and which must be approximated (as presented in Appendix A.2) for use in the OVPFA algorithm. The OVPFA algorithm requires a quadratic function which is an optimization well supported by standard software. The OVPFA algorithm requires evaluation of hypergeometric functions of multivariate arguments, for which

The most numerically convenient operation in the OVPFA algorithm is evaluation of roots of a quadratic function which is an optimization well supported by standard software. The OVPFA

algorithm requires evaluation of hypergeometric functions of multivariate arguments, for which

reliable solution is not available, and which must be approximated (as presented in Appendix A.2) for use in the OVPFA algorithm. The OVPFA algorithm requires evaluation of roots of

quadratic function which is an optimization well supported by standard software. The OVPFA

algorithm requires evaluation of hypergeometric functions of multivariate arguments, for which

reliable solution is not available, and which must be approximated (as presented in Appendix A.2) for use in the OVPFA algorithm. The OVPFA algorithm requires evaluation of roots of

quadratic function which is an optimization well supported by standard software. The OVPFA

algorithm requires evaluation of hypergeometric functions of multivariate arguments, for which

reliable solution is not available, and which must be approximated (as presented in Appendix A.2) for use in the OVPFA algorithm. The OVPFA algorithm requires evaluation of roots of

quadratic function which is an optimization well supported by standard software. The OVPFA

algorithm requires evaluation of hypergeometric functions of multivariate arguments, for which

reliable solution is not available, and which must be approximated (as presented in Appendix A.2) for use in the OVPFA algorithm. The OVPFA algorithm requires evaluation of roots of

quadratic function which is an optimization well supported by standard software. The OVPFA

algorithm requires evaluation of hypergeometric functions of multivariate arguments, for which

reliable solution is not available, and which must be approximated (as presented in Appendix A.2) for use in the OVPFA algorithm. The OVPFA algorithm requires evaluation of roots of

quadratic function which is an optimization well supported by standard software. The OVPFA

algorithm requires evaluation of hypergeometric functions of multivariate arguments, for which

Chapter 7.

Bayesian Modelling for Functional Analysis of Medical Image Data

In this Chapter, we study the Bayesian inference of parameters of the FAMIS model introduced in Section 3.4. The standard parameter inference method for this problem is achieved in three steps in order to achieve computational tractability (Section 3.4). One of the steps is PCA, studied in previous Chapter. Hence, we now study application of the previous results in this context. The main concern of this Chapter is, however, derivation of a unified identification method for the FAMIS model (Section 3.4.5), using Variational Bayes (VB) method.

7.1. Bayesian Formulation

The FAMIS model (3.91) is, in essence, an extension of the PPCA model (3.50). We seek a Bayesian inference of the model parameters A, X, ω_p, ω_n . Following the Bayesian methodology, we complement the observation model (3.91) by priors on the model parameters.

Prior distributions on the precision matrices were chosen as follows:

$$f(\omega_p | \vartheta_p, \rho_p) = \prod_{i=1}^p \mathcal{G}(\vartheta_{i;p}, \rho_{i;p}), \tag{7.1}$$

$$f(\omega_n | \vartheta_n, \rho_n) = \prod_{i=1}^n \mathcal{G}(\vartheta_{i;n}, \rho_{i;n}),$$

with vector hyper-parameters $\vartheta_p = [\vartheta_{1;p}, \dots, \vartheta_{p;p}]$, $\rho_p = [\rho_{1;p}, \dots, \rho_{p;p}]$, $\vartheta_n = [\vartheta_{1;n}, \dots, \vartheta_{n;n}]$, $\rho_n = [\rho_{1;n}, \dots, \rho_{n;n}]$. These parameters can be chosen to yield a non-committal prior. Alternatively, recall—from Section 3.4.2—that pre-processing methods were derived by studying asymptotic properties of the noise. Hence, the asymptotic values can be used to elicit priors. These hyper-parameters can be seen as ‘knobs’ to tune the method to suit clinical practice.

The parameters, A and X , are modelled in the same way as those in Variational PCA, with the additional restriction of positivity. Prior distributions (3.65) and (3.66) then become

$$f(A | \mathbf{v}) = t\mathcal{N}(\mathbf{0}_{p,r}, I_p \otimes \Upsilon^{-1}, (\mathbb{R}^+)^{p,r}), \tag{7.2}$$

$$f(X) = t\mathcal{N}(\mathbf{0}_{r,n}, I_r \otimes I_n, (\mathbb{R}^+)^{r,n}), \tag{7.3}$$

$$f(v_i) = \mathcal{G}(\alpha_0, \beta_0), \quad i = 1, \dots, r, \tag{7.4}$$

with hyper-parameter $\Upsilon = \text{diag}(\mathbf{v}) \in \mathbb{R}^{r \times r}$, designed for inference of rank via the ARD property (Remark 3.8).

From (3.91), (7.1)–(7.4), the joint distribution is

$$\begin{aligned} f(D, A, X, \Omega_p, \Omega_n, \mathbf{v}) &= \mathcal{N}(AX, \Omega_p^{-1} \otimes \Omega_n^{-1}) \prod_{i=1}^p \mathcal{G}(\vartheta_{i;p}, \rho_{i;p}) \\ &\quad \prod_{i=1}^n \mathcal{G}(\vartheta_{i;n}, \rho_{i;n}) t\mathcal{N}(\mathbf{0}_{p,r}, I_p \otimes \Upsilon^{-1}, (\mathbb{R}^+)^{p,r}) \\ &\quad t\mathcal{N}(\mathbf{0}_{r,n}, I_r \otimes I_n, (\mathbb{R}^+)^{r,n}) \mathcal{G}(\alpha_0, \beta_0). \end{aligned} \quad (7.5)$$

The posterior distribution is then obtained using Bayes' rule:

$$f(A, X, \Omega_p, \Omega_n, \mathbf{v} | D) = \frac{f(D, A, X, \Omega_p, \Omega_n, \mathbf{v})}{f(D)}. \quad (7.6)$$

Exact Bayesian inference of this model is not tractable.

7.2. Variational Bayes (VB) Approximation

Following the Variational approximation, we can find an approximate posterior inference under the assumption of conditional independence, Theorem 2.1.

Corollary 7.1 (Corollary 6 of Theorem 2.1) *Consider the following conditional independence factorization of (7.6):*

$$\check{f}(A, X, \Omega_p, \Omega_n, \mathbf{v} | D) = \check{f}(A | D) \check{f}(X | D) \check{f}(\Omega_p | D) \check{f}(\Omega_n | D) \check{f}(\mathbf{v} | D). \quad (7.7)$$

Then, using (7.5) and (7.7) in Theorem 2.1, the VB-optimal form of (7.7) is found via the following assignments:

$$\tilde{f}(\text{vec}(A) | D) = t\mathcal{N}(\mu_A, \Sigma_A, (\mathbb{R}^+)^{pr}), \quad (7.8)$$

$$\tilde{f}(\text{vec}(X) | D) = t\mathcal{N}(\mu_X, \Sigma_X, (\mathbb{R}^+)^{nr}), \quad (7.9)$$

$$\tilde{f}(\mathbf{v} | D) = \prod_{i=1}^r \mathcal{G}(\alpha_i, \beta_i), \quad (7.10)$$

$$\tilde{f}(\omega_p | D) = \prod_{i=1}^p \mathcal{G}(\vartheta_{i;p}, \rho_{i;p}), \quad (7.11)$$

$$\tilde{f}(\omega_n | D) = \prod_{i=1}^n \mathcal{G}(\vartheta_{i;n}, \rho_{i;n}), \quad (7.12)$$

with VB-statistics

$$\mu_A = \Sigma_A \text{vec} \left(\widehat{\Omega}_p D \widehat{\Omega}_n \widehat{X}' \right), \quad (7.13)$$

$$\Sigma_A = \left(\mathbb{E}_{X|D} \left(X \widehat{\Omega}_n X' \right) \otimes \widehat{\Omega}_p + \text{diag}(\widehat{v}) \otimes I_p \right)^{-1}, \quad (7.14)$$

$$\mu_X = \Sigma_X \text{vec} \left(\widehat{A}' \widehat{\Omega}_p D \widehat{\Omega}_n \right), \quad (7.15)$$

$$\Sigma_X = \left(\mathbb{E}_{A|D} \left(A' \widehat{\Omega}_p A \right) \otimes \widehat{\Omega}_n + I_r \otimes I_n \right)^{-1}, \quad (7.16)$$

$$\alpha_i = \alpha_0 + \frac{1}{2} p, \quad i = 1, \dots, r,$$

$$\beta = \beta_0 + \frac{1}{2} \text{diag}^{-1} \left(\mathbb{E}_{A|D} \left(A' A \right) \right),$$

$$\vartheta_p = \vartheta_{0,p} + \frac{n}{2},$$

$$\begin{aligned} \rho_p = \rho_{0,p} + \frac{1}{2} \text{diag}^{-1} \left(D \widehat{\Omega}_n D' - \widehat{A} \widehat{X}' \widehat{\Omega}_n D' - D \widehat{\Omega}_n \widehat{X}' \widehat{A}' + \right. \\ \left. + \mathbb{E}_{A|D} \left(A \mathbb{E}_{X|D} \left(X \widehat{\Omega}_n X' \right) A' \right) \right), \end{aligned} \quad (7.17)$$

$$\vartheta_n = \vartheta_{0,n} + \frac{p}{2},$$

$$\begin{aligned} \rho_n = \rho_{0,n} + \frac{1}{2} \text{diag}^{-1} \left(D' \widehat{\Omega}_p D - D' \widehat{\Omega}_p \widehat{A} \widehat{X} - \widehat{X}' \widehat{A}' \widehat{\Omega}_p D \right. \\ \left. + \mathbb{E}_{X|D} \left(X' \mathbb{E}_{A|D} \left(A' \widehat{\Omega}_p A \right) X \right) \right). \end{aligned} \quad (7.18)$$

Posterior distributions of A and X are not of the matrix Normal distribution kind (Appendix A.1). Therefore, they are written in the form of the $\text{vec}(\cdot)$ operator (Appendix A.1).

Proof: Can be handled in the same way as proofs for the previous VB-related Corollaries. It is an easy but lengthy exercise in probability calculus. ■

Evaluation of moments of distributions (7.8) and (7.9) is complicated for two reasons:

1. the VB-statistics involve evaluation of large matrices, e.g. $\Sigma_A \in \mathbb{R}^{nr \times nr}$.

These matrices are block diagonal, hence all the operations involved in evaluation of (7.13)–(7.18) can be re-written in terms of blocks of these matrices. This operation is formally trivial but rather lengthy. Therefore, it will be omitted in this text.

2. moments of the truncated Normal distribution of vector argument are not known to us. Therefore, we approximate all involved moments of $\tilde{f}(A|D)$ and $\tilde{f}(X|D)$ by moments of

$$\bar{f}(A|D) = t\mathcal{N} \left(\mu_A, \text{diag} \left(\text{diag}^{-1}(\Sigma_A) \right), (\mathbb{R}^+)^{p \times r} \right),$$

$$\bar{f}(X|D) = t\mathcal{N} \left(\mu_X, \text{diag} \left(\text{diag}^{-1}(\Sigma_X) \right), (\mathbb{R}^+)^{r \times n} \right),$$

respectively. In effect, we neglect all covariances between elements of A , and ditto for X .

The Variational Extreme can be found by iterating (7.13)–(7.18) to convergence via the VEM algorithm (Algorithm 2.2).

7.3. Computational Simplifications

In principle, the model (7.5) is an extension of the PPCA model (3.50) for (i) unknown precision matrix of the noise, $\Omega_p \otimes \Omega_n$, and (ii) positivity constraints (7.2), (7.3). These extensions results in a significant increase in the computational complexity of the associated VEM algorithm, compared to the VEM algorithm associated with VPCA (Section 3.1). The main causes of this are as follows:

1. the structure of covariance matrices of the priors (7.2), (7.3) is different from the covariance structure of the model (3.91). Therefore, the covariance matrices, (7.14) and (7.16), of the posterior pdfs, (7.8) and (7.9), are not in the advantageous Kronecker-product form.
2. positivity restrictions, (7.2) and (7.3), are not satisfied for the orthogonal solution (3.51). Therefore, a simplification similar to Proposition 6.1 cannot be used. Moreover, no analytical result is known to us that can be used to invoke the Restricted VB (Corollary 2.1).

The first problem may be addressed by choice of different priors. The choice of covariance matrix, $I_p \otimes \Upsilon^{-1}$ in (7.8), and $I_r \otimes I_n$ in (7.9), is intuitively appealing. It is a simple choice which imposes the same prior on each pixel in the image, and, as a result, it acts as soft orthogonality constraint (Conjecture 6.1). However, as we have shown in the analysis of the toy problem (Section 6.1), analytical solution can be found for other choices of prior (Remark 6.1), as follows.

Proposition 7.1 (Alternative priors) *For the following choice of priors*

$$f(A|\mathbf{v}, \Omega_p) = t\mathcal{N}(\mathbf{0}_{p,r}, \Omega_p^{-1} \otimes \Upsilon^{-1}, (\mathbb{R}^+)^{p,r}), \quad (7.19)$$

$$f(X|\Omega_n) = t\mathcal{N}(\mathbf{0}_{r,n}, I_r \otimes \Omega_n^{-1}, (\mathbb{R}^+)^{r,n}), \quad (7.20)$$

in place of (7.2) and (7.3) the VB approximation (Corollary 7.1) yields the posterior results in the form of (7.8)–(7.12), with VB-statistics (7.14) and (7.16) replaced by

$$\Sigma_A = \left(\mathbb{E}_{X|D} \left(X \widehat{\Omega}_n X' \right) \otimes \widehat{\Omega}_p + \text{diag}(\widehat{\mathbf{v}}) \otimes \widehat{\Omega}_p \right)^{-1}, \quad (7.21)$$

$$\Sigma_X = \left(\mathbb{E}_{A|D} \left(A' \widehat{\Omega}_p A \right) \otimes \widehat{\Omega}_n + I_r \otimes \widehat{\Omega}_n \right)^{-1}, \quad (7.22)$$

respectively. These can be written in Kronecker product form:

$$\Sigma_A = \left(\mathbb{E}_{X|D} \left(X \widehat{\Omega}_n X' \right) + \text{diag}(\widehat{\mathbf{v}}) \right)^{-1} \otimes \widehat{\Omega}_p^{-1},$$

$$\Sigma_X = \left(\mathbb{E}_{A|D} \left(A' \widehat{\Omega}_p A \right) + I_r \right)^{-1} \otimes \widehat{\Omega}_n^{-1}.$$

Hence, the posterior distributions, (7.8) and (7.9), are, again, in the form of the (truncated) Matrix Normal distribution:

$$\tilde{f}(A|D) = t\mathcal{N}(\mu_A, \widehat{\Omega}_p^{-1} \otimes \Phi_A^{-1}, (\mathbb{R}^+)^{p \times r}), \quad (7.23)$$

$$\tilde{f}(X|D) = t\mathcal{N}(\mu_X, \Phi_X^{-1} \otimes \widehat{\Omega}_n^{-1}, (\mathbb{R}^+)^{r \times n}), \quad (7.24)$$

with VB-statistics

$$\begin{aligned}\mu_A &= D\widehat{\Omega}_n\widehat{X}'\Phi_A^{-1}, \\ \Phi_A &= E_{X|D}\left(X\widehat{\Omega}_nX'\right) + \text{diag}(\widehat{\nu}), \\ \mu_X &= \Phi_X^{-1}\widehat{A}'\widehat{\Omega}_pD, \\ \Phi_X &= E_{A|D}\left(A'\widehat{\Omega}_pA\right) + I_r.\end{aligned}$$

Note that the priors proposed in Proposition 7.1 are in the following mutually dependent form:

$$f(A, X, \Omega_n, \Omega_p) = f(A|\Omega_p) f(X|\Omega_n) f(\Omega_p) f(\Omega_n),$$

which is inconsistent with the assumption of conditional independence of posteriors as enforced by the VB approximation (7.7). However, the covariance matrices of the priors (7.19), (7.20) were chosen to reflect the structure of the covariance matrices arising from the observation model (as demonstrated by operations (7.14), (7.16)). Then, structure of the covariance matrices of the prior (7.19) is similar to that of the posterior (7.23), and the posterior is not prior-dominated.

Proposition 7.1 reduces the amount of computation associated with the VEM algorithm. However, no further analytical simplification can be made to decrease the number of iterations or number of parameters required in iterations (7.13)–(7.18). The main complication is the restriction of the support of the posterior Normal distributions (7.23), (7.24) to the set \mathfrak{R}^+ .

The VB-based identification of the FAMIS model is closely related to Independent Component Analysis (ICA). Specifically, FAMIS can be seen as a special case of *noisy ICA* [110]. In fact, our linear model $D = AX + E$, (3.1) and (3.2), is identical with that of ICA. In ICA, A is known as the *mixing matrix* and rows of X are called *sources*. Therefore, any method concerning this model may be called ICA. This broad meaning attached to ‘ICA’ makes categorical comparison rather difficult. However, the main keyword of ICA is the word *independent*. The method is typically defined in relation to the classical signal separation methods of Principal Component Analysis (PCA) and Factor Analysis (FA) [110]. In PCA and FA, the criteria used for signal separation are based on the sample covariance matrix of the signal, i.e. on the second moment of its distribution. In ICA, the criterion for signal separation is full statistical independence of sources, which corresponds to the assumption

$$f(\mathbf{x}_t) = \prod_{i=1}^r f(x_{i;t}), \quad t = 1, \dots, n, \quad (7.25)$$

in our notation. Note that this assumption does not imply any particular functional form of the probability distribution¹. Therefore, inference of parameters has to be adjusted for the chosen pdf. However, a template algorithm was developed for maximum-likelihood (ML) estimation [111], for which the change of the pdf influences only one operation in the algorithm. Bayesian inference is, however, much more complicated, and it is available only for a limited class of prior distributions. The solution for Gaussian and truncated Gaussian priors was presented in this thesis (Sections 3.3.3 and 7.2). The solution for mixtures of Gaussian priors was presented in [33].

¹In fact, for Gaussian distribution, the criterion (7.25) is identical to that of PCA.

From a computational point of view, the ICA method is currently receiving much attention and many interesting results has been reported. For example, efficient evaluation of the traditional Maximum Likelihood (ML) approach to ICA was achieved using fixed-point approximations [112]. Application of these ideas in the context of the Variational Bayes (VB) approach may bring significant computational savings.

Reduction of the computational cost for the VEM algorithm can also be achieved using heuristic techniques. For example:

- the traditional three-step methods (Section 3.4) can be used as a reasonable initial guess for the VEM iterative algorithm.
- for fixed estimates of the covariance matrices, $\widehat{\Omega}_p$ and $\widehat{\Omega}_n$, evaluation of the remaining parameters is somewhat simplified (in analogy to the simplification in Proposition 6.2). It may be useful to re-evaluate $\widehat{\Omega}_p$ and $\widehat{\Omega}_n$ only once for every q steps of the VEM algorithm (e.g. $q = 10$).

These *ad hoc* propositions have not been extensively tested and are left for further study.

7.4. Experiments

In this Section, we study performance of the Bayesian inference of the FAMIS model on a sequence of scintigraphic images of the chest. In this study, a radiotracer has been administered to the patient to highlight the kidneys and bladder. In this context, we perform two experiments: (i) we test application of Bayesian PCA in the orthogonal step of the standard approach (Section 3.4.3), and (ii) we test the performance of the proposed VB inference of the FAMIS model (Section 3.4.5).

7.4.1. Comparison of Methods for Inference of Rank in Orthogonal Analysis

The advantage of Bayesian PCA over the standard PCA is an explicit estimation of the number of relevant principal components (Section 6.3.4). In this study, a scintigraphic dynamic image sequence of the kidneys is considered. It contains $n = 120$ images, each of size 64×64 . These were analyzed using the standard approach, as follows:

Pre-processing:

- a rectangular area of $p = 525$ pixels was chosen as the region of interest at the same location in each image.
- data were scaled by the correspondence analysis method (3.89), which is optimal for scintigraphic data [89].

Orthogonal analysis:

Bayesian methods of inference for the PPCA model (Section 6) were tested. The expected mean value of both factor images and factor curves are identical with those obtained using standard PCA. The methods differ only in the estimated rank of the data.

Table 7.1.: Comparison of rank selection methods for scintigraphic image data.

OVPCA $f(r D)$	OVPCA ARD Property	FVPCA ARD Property	Laplace $f(r D)$
$\Pr(r = 17 D) = 0.0004$	$r_u = 45$	$r_u = 25$	$\Pr(r = 47 D) = 0.067$
$\Pr(r = 18 D) = 0.2761$			$\Pr(r = 48 D) = 0.622$
$\Pr(r = 19 D) = 0.7232$			$\Pr(r = 49 D) = 0.195$
$\Pr(r = 20 D) = 0.0002$			$\Pr(r = 50 D) = 0.089$

Note: where not listed, $f(r|D) < 3 \times 10^{-7}$

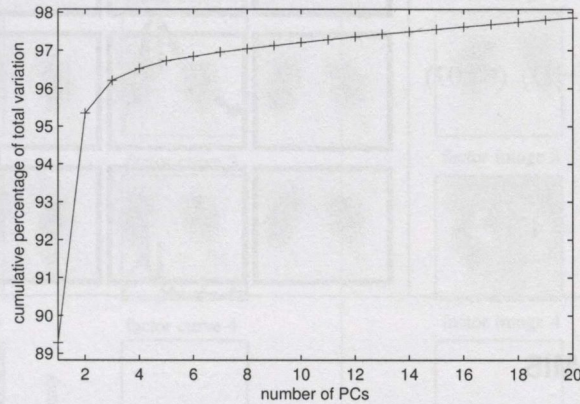


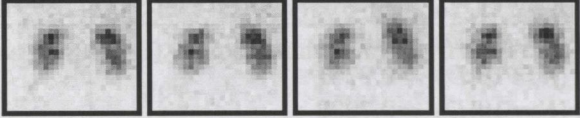
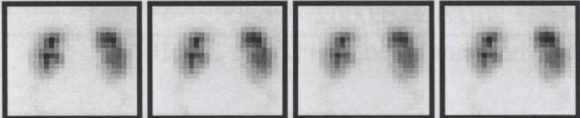

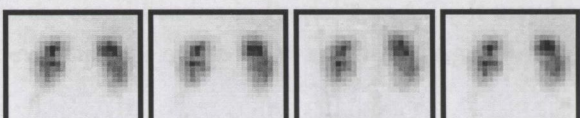
Figure 7.1.: Cumulative percentage of total variation for scintigraphic data. For clarity, only the first 20 elements are shown out of a total of $p = 120$.

Oblique analysis: this was not tested.

For these data, we compare methods for selection of relevant principal components. The OVPCA-based approximate posterior distribution of rank (6.107) and the ARD properties of both OVPCA (Remark 6.3) and VPCA (Remark 3.8) infer significantly different optimal rank (Table 7.1). For comparison, we also inferred the rank of the data via (i) Laplace approximation [84], and (ii) the *ad hoc* criterion of cumulative percentage of total variation [80] (Figure 7.1). Results are presented in Table 7.1. For method (ii), $r = 5$ was chosen.

It is difficult to compare performance of the methods since the true dimensionality is not known. From a medical point of view, the number of physiological factors should be 4 or 5. This estimate is supported by the *ad hoc* criterion (Figure 7.1). From this perspective, the formal methods appear to over-estimate significantly the number of relevant principal components (PCs). The reason for this can be understood by reconstructing the data using the number of PCs, r , recommended by each method (Table 7.2). Four consecutive frames of the actual scintigraphic data are displayed in the first row. Though the signal-to-noise ratio is poor, functional variation is visible in the central part of the left kidney and in the upper part of the right kidney, which cannot be accounted for by noise. The same frames of the sequence, reconstructed from $r = 5$ PCs (Table 7.2, second row), fail to capture this functional information. In contrast, the functional information is apparent on the sequence reconstructed using the Bayesian estimate—i.e. $r = 19$ PCs—and, indeed, on sequences reconstructed using $r > 19$ PCs, such as the $r = 45$ choice suggested by the ARD Property of OVPCA (Table 7.2, last row).

Table 7.2.: Reconstruction of scintigraphic data for different numbers of PCs

number of PCs used	frames 48–51 of the dynamic image sequence
Original images ($r = 120$)	
Ad hoc criterion ($r = 5$)	
Maximum of $f(r D)$ (6.107) ($r = 19$)	
ARD property ($r = 45$)	

7.4.2. Variational FAMIS

The FAMIS model is, in principle, the PPCA model (3.50) extended for unknown precision of noise (3.87), and restricted by positivity constraints (3.85). In the previous Section, the results of the PPCA model were presented. The precision of the noise was assumed known via the correspondence analysis (3.90).

Performance of the VEM algorithm for the FAMIS model (Corollary 7.1) was tested on the same data set used in Section 7.4.1, i.e. the scintigraphic study of kidneys of $n = 120$ images, each of size 64×64 , with selected region of interest of $p = 525$ pixels. First, we performed two experiments with *a priori* known precision matrices, Ω_p, Ω_n , corresponding to:

1. homogeneous noise: i.e. $\Omega_p = \omega I_p, \Omega_n = I_n$ (case 1).
2. correspondence analysis: $\Omega_p = \text{diag}(D\mathbf{1}_{n,1})^{-1}, \Omega_n = \text{diag}(D'\mathbf{1}_{p,1})^{-1}$ (case 2).

Note that this is the same pre-processing that was used in the standard analysis (Section 7.4.1).

The results of these experiments are displayed in Figure 7.2 (top-left) and (bottom-left) respectively.

Next, the VEM algorithm for the FAMIS model (Corollary 7.1) was used with non-committal prior ($\vartheta_p = \vartheta_n = \rho_n = \rho_p = 1e - 10 \times \mathbf{1}_{p,1}$ (7.1)). The algorithm was initialized using the same options as for the fixed matrices, i.e:

1. homogeneous noise: i.e. $\widehat{\Omega}_p^{(0)} = \omega I_p, \widehat{\Omega}_n^{(0)} = I_n$ (case 3)
2. correspondence analysis: $\widehat{\Omega}_p^{(0)} = \text{diag}(D\mathbf{1}_{n,1})^{-1}, \widehat{\Omega}_n^{(0)} = \text{diag}(D'\mathbf{1}_{p,1})^{-1}$ (case 4).

Results of these experiments are displayed in Figure 7.2 (top-right) and (bottom-right) respectively.

This experiment with real data leads to the following conclusions:

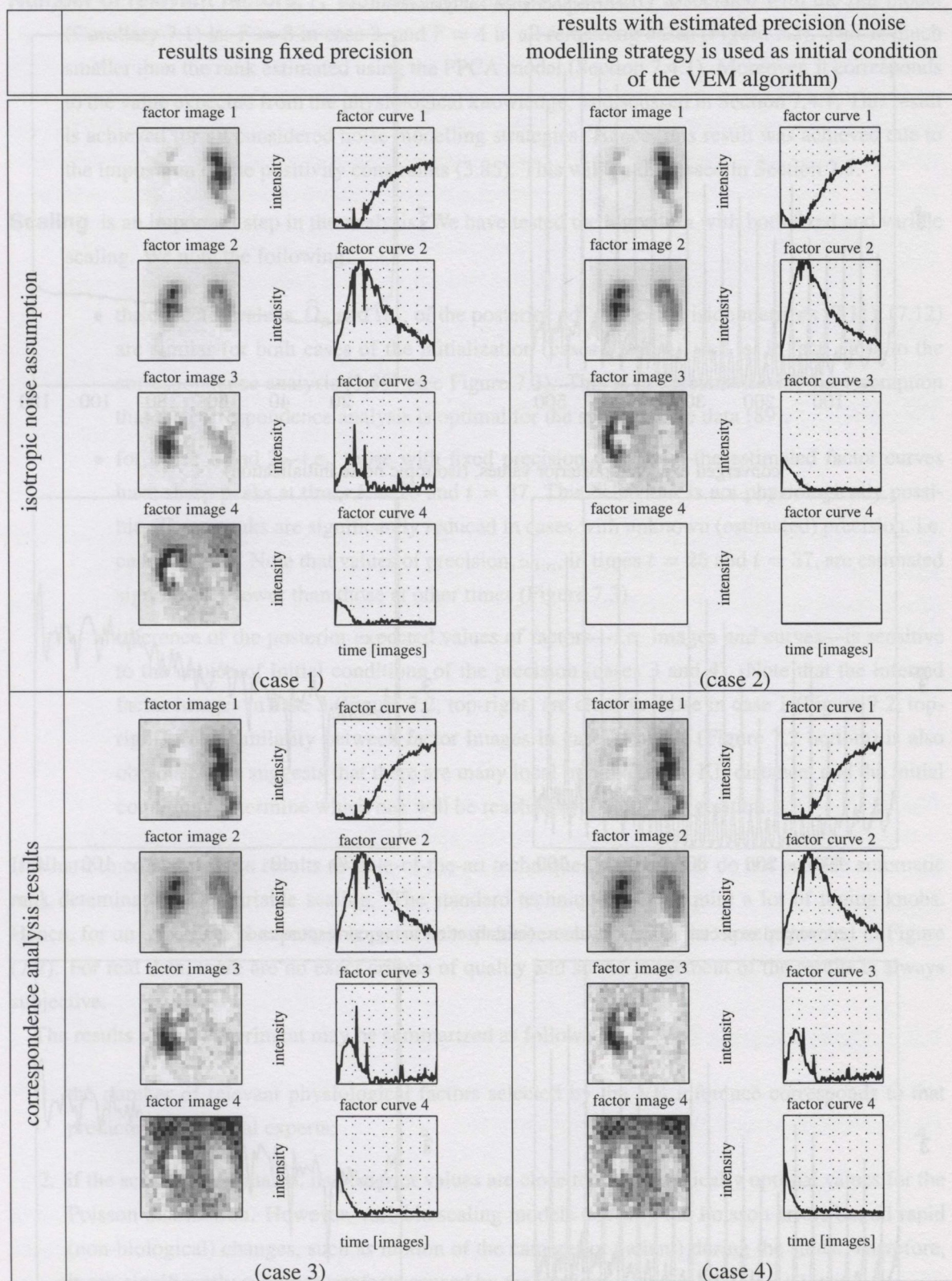


Figure 7.2.: Expected posterior values of factor images and factor curves for four noise modelling strategies.

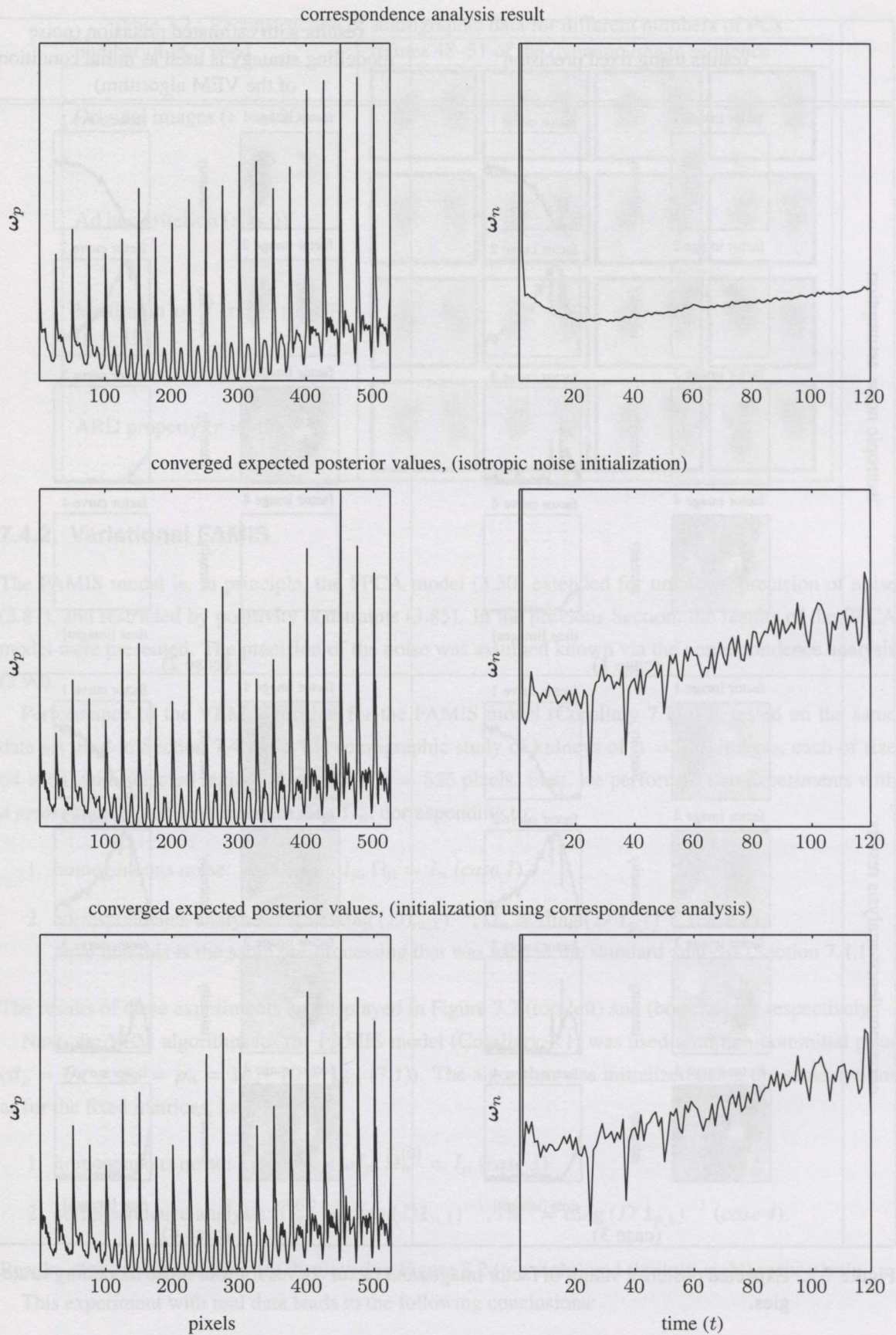


Figure 7.3.: Comparison of the posterior expected value of of the precision matrices Ω_p (left) and Ω_n (right), for different initializations.

Number of relevant factors, r , estimated using the ARD property associated with the full model (Corollary 7.1) is: $\hat{r} = 3$ in case 3, and $\hat{r} = 4$ in all remaining cases (Figure 7.2). This is much smaller than the rank estimated using the PPCA model (Section 7.4.1). Moreover, it corresponds to the value expected from the physiological knowledge, as discussed in Section 7.4.1. This result is achieved for all considered noise modelling strategies. Hence, this result was achieved due to the imposition of the positivity constraints (3.85). This will be discussed in Section 7.5.

Scaling is an important step in the analysis. We have tested the algorithm with both fixed and variable scaling. We note the following:

- the expected values, $\hat{\Omega}_p$ and $\hat{\Omega}_n$, of the posterior pdf of the precision matrices (7.11), (7.12) are similar for both cases of the initialization (cases 3 and 4), and is, in fact, close to the correspondence analysis (3.89) (see Figure 7.3). This is in agreement with the assumption that the correspondence analysis is optimal for the scintigraphic data [89].
- for cases 1 and 2—i.e. those with fixed precision matrices—the estimated factor curves have sharp peaks at times $t = 25$ and $t = 37$. This behaviour is not physiologically possible. These peaks are significantly reduced in cases with unknown (estimated) precision, i.e. cases 3 and 4. Note that values of precision, $\omega_{t,n}$, in times $t = 25$ and $t = 37$, are estimated significantly lower than those at other times (Figure 7.3).
- inference of the posterior expected values of factors—i.e. images *and* curves—is sensitive to the choice of initial conditions of the precision (cases 3 and 4). Note that the inferred factor images in case 3 (Figure 7.2, top-right) are close to those in case 1 (Figure 7.2, top-right). The similarity between factor images in cases 4 and 2 (Figure 7.2 bottom) is also obvious. This suggests that there are many local minima of the KL distance, and the initial conditions determine which one will be reached by the VEM algorithm.

It is hard to compare these results to state-of-the-art techniques, as the latter do not provide automatic rank determination nor variable scaling. The standard techniques also require a lot of tuning knobs. Hence, for an experienced expert, it is possible to produce results similar to those presented in Figure (7.2). For real data, there are no exact criteria of quality and so the judgement of the results is always subjective.

The results of this experiment may be summarized as follows:

1. the number of relevant physiological factors selected by the VB inference corresponds to that predicted by medical experts;
2. if the scaling is estimated, its posterior values are close to the theoretically optimal values for the Poisson distribution. However, variable scaling models not only the Poisson errors but all rapid (non-biological) changes, such as motion of the camera (or patient) during the study. Therefore, it can significantly suppress artefacts caused by the motion.
3. the method is too sensitive to the choice of initial conditions. Further work on initial conditions and convergence of the algorithm is required.

7.5. Discussion

One of the unsolved problems relating to the standard solution of the FAMIS model is estimation of the number of relevant principal components in orthogonal analysis (Section 3.4.3). This problem can be addressed using Bayesian PCA, as presented in Chapter 6. Therefore, we have tested the FVPCA and OVPCA algorithms, in this Chapter, in the context of Functional Analysis of Medical Image Sequences (FAMIS). We noted, in Section 7.4.1, that both algorithms significantly overestimated the number of relevant physiological factors. However, the VB inference of the unified FAMIS model (Section 3.4.5) yields physiologically acceptable results.

7.5.1. Model Matching

This last result can be explained by considering how each model is matched to the actual medical data. The real scintigraphic data are composed of three elements:

$$D = M + N + E.$$

M and E are the modelled elements, M being the rank-restricted mean value (3.48) with positivity constraints (3.85), and E being the Normally-distributed white noise (3.49). N is an *unmodelled* matrix of non-Gaussian noise and physiological residuals. Inevitably, then, the OVPCA and FAMIS methods provide estimates of the modelled parameters, M and E , corrupted by the residuals, N , as follows:

$$\begin{aligned}\hat{M} &= M + N_M, \\ \hat{E} &= E + N_E.\end{aligned}$$

N_M and N_E are method-dependent parts of the residual element, $N = N_M + N_E$.

If the criteria of separation are (i) rank-restriction with unknown r , and (ii) Gaussianity of the noise, then only a small part of N fulfills (ii), but a large part of N fulfills (i) as it has unknown rank. Consequently, the rank, r , is significantly over-estimated. However, if we now impose a *third* constraint, namely (iii) *positivity* of the signal M (3.85), we can expect that only a small part of N fulfills (iii), ‘pushing’ the larger part of N into the noise estimate, \hat{E} .

7.5.2. Consequence for Medical Applications

We have demonstrated experimentally that the joint Bayesian identification of the FAMIS model has the following advantages over the standard approach:

1. identification of the noise distribution—via parameters ω_p and ω_n —is more accurate than the standard scaling technique. Therefore, posterior estimates of the FAMIS model are less sensitive to non-standard noise distributions, and are more reliable under low Signal-to-Noise Ratio (SNR).
2. identification of the number of relevant physiological factors yields realistic and reliable results. No such method was available in the standard model; i.e., formerly, the number of relevant physiological factors was either chosen constant *a priori*, or it was selected by a human expert.

These results should, however, still be considered as preliminary. The method is not ready for application in medical imaging as it has not been showed to meet high standards of reliability and accuracy required in this area. The following problems must still be overcome:

- convergence of the algorithm is slow, and the overall computational cost is high (tens of minutes of computation on a 1GHz machine, for analysis of 100 frames of size 128×128). Possible approaches to this problem were discussed in Section 7.3;
- the posterior estimates are sensitive to the chosen initial conditions. Extensive experimental studies will probably be needed to choose the best initialization of the method;
- the only restriction imposed on each factor in the FAMIS model was positivity. Hence, the posterior estimates of the factor curves may contain sharp peaks (Figure 7.2) which are not physiologically possible. A further restriction of the factors is needed, such as the imposition of smoothness constraints on the factor curves [113].

These results should, however, still be considered as preliminary. The method proposed in this paper for medical imaging is not shown to meet high standards of reliability and accuracy. The following sections describe the proposed method for identifying relevant physiological factors. This problem can be reformulated as a sparse regression problem (Section 3.4.3). The proposed method is based on the following assumptions: (i) the number of relevant physiological factors is small, (ii) the signal-to-noise ratio (SNR) is high, and (iii) the noise is Gaussian. The proposed method is based on the following assumptions: (i) the number of relevant physiological factors is small, (ii) the signal-to-noise ratio (SNR) is high, and (iii) the noise is Gaussian. The proposed method is based on the following assumptions: (i) the number of relevant physiological factors is small, (ii) the signal-to-noise ratio (SNR) is high, and (iii) the noise is Gaussian.

The only restriction imposed on each factor in the FAMIS model was that it must be a non-negative function of the factor curves may contain sharp peaks (Figure 7.5) which are not physiological. A further restriction of the factor is needed, such as the imposition of smoothness constraints on the factor curves [113].

$$D = M + N + E.$$

M and E are the modelled elements, M being the rank-restricted mean value (3.48) with positivity constraints (3.45), and E being the Normally-distributed white noise (3.49). N is an unmodelled matrix of non-Gaussian noise and physiological residuals. Inevitably, then, the OLVPCA and FAMIS methods provide estimates of the modelled parameters, M and E , corrupted by the residuals, N , as follows:

$$\hat{M} = M + N_M,$$

$$\hat{E} = E + N_E.$$

N_M and N_E are model-dependent parts of the residual element, $N = N_M + N_E$. If the criteria of separation are (i) rank restriction with unknown r , and (ii) Gaussianity of the noise, then only a small part of N fulfils (i), but a large part of N fulfils (ii) as it has unknown rank. Consequently, the rank, r , is significantly over-estimated. However, if we now impose a third constraint, namely (iii) positivity of the signal M (3.45), we can expect that only a small part of N fulfils (iii), 'pushing' the larger part of N into the noise estimate, \hat{E} .

7.6.2. Consequence for Medical Applications

We have demonstrated experimentally that the joint Bayesian identification of the FAMIS model has the following advantages over the standard approach:

1. identification of the noise distribution—its parameters ω_σ and ω_μ —is more accurate than the standard scaling technique. Therefore, parameter estimates of the FAMIS model are less sensitive to non-standard noise distributions, and are more reliable under low Signal-to-Noise Ratio (SNR);
2. identification of the number of relevant physiological factors yields realistic and reliable results. No such method was available in the standard model, i.e., formerly, the number of relevant physiological factors was either chosen constant a priori, or it was selected by a human expert.

Chapter 8.

Conclusions

8.1. Discussion of the Work

The aim of this thesis was to extend the modelling capabilities of the linear model and to provide a numerically efficient Bayesian inference of the model parameters. Four different extensions of special cases of the linear model have been studied, each of them representing an important problem in Digital Signal Processing (DSP). In each case, we have derived a novel identification algorithm and shown its advantages over the existing techniques. In all cases, we have used the Variational Bayes (VB) approximation (Section 2.2.4) to obtain the posterior distributions of model parameters.

We have shown that the VB approximation is particularly appropriate for identification of a non-stationary process. A non-stationary process generates a new random variable at each time step. This leads to proliferation of random variables and computational intractability in exact Bayesian identification. The problem was circumvented by invoking the conditional independence assumption—which is the central assumption of the VB approximation—and optimizing it using that same VB procedure.

One of the main concerns of the thesis was computational efficiency, as is appropriate for work in DSP. We have shown that the Variational EM (VEM) algorithm (Algorithm 2.2)—i.e. the standard algorithm for evaluation of VB-statistics (which are the parameters of the VB-optimal posterior distribution)—may be computationally inefficient. In special cases (Chapter 6), the solution of the VB posterior distributions can be analytically simplified, yielding significantly faster identification algorithms. Computationally faster algorithms may also be achieved by further approximations of the VB method, such as the Restricted VB method (Corollary 2.1), of which Quasi-Bayes (QB) is a significant example. Under these simplified procedures, we showed that significant computational savings could be achieved at the price of only slight loss of accuracy.

We now discuss in more detail the key contributions of the work.

8.2. Key Contributions of the Thesis

Chapter 2: We reviewed the most common methods for approximation of Bayesian posterior distributions. The main emphasis was on the Variational Bayes (VB) approximation (Theorem 2.1). We introduced a Restricted VB approximation (Corollary 2.1) which yields parameters of the VB optimal posterior distribution in closed-form (closed-form solution for the non-restricted VB approximation is rare, an iterative VEM procedure is almost always implied). We showed that the

popular Quasi-Bayes (QB) approximation is a special case of this Restricted VB approximation method.

Chapter 3: We reviewed the most important special cases of the linear model and their Bayesian identification. Many well known multivariate methods—such as Factor Analysis (FA) or Principal Component Analysis (PCA)—had already been re-derived in the literature using the linear model. We showed that the method known as Factor Analysis of Medical Image Sequences (FAMIS) can also be understood this way.

Chapter 4: We extended the AutoRegressive (AR) model to embrace unknown transformations of its output. A tractable identification can be achieved only when there is a finite set of candidates. What follows is the Mixture-based Extended AR model (MEAR). The MEAR model is a mixture of AR components with common AR parameterization, each component modelling the AR process with respect to one possible data transformation. These transformations can be interpreted as a bank of filters, where each filter is used to pre-process the observed data. We have derived three algorithms for Bayesian identification of the underlying AR parameters: (i) Variational Bayes (VB), (ii) Quasi-Bayes (QB), and (iii) the Viterbi-Like (VL) algorithm. Each of these represents a different trade-off between numerical speed and accuracy. The VB algorithm is the most accurate, and the VL is computationally the least expensive. We present applications of these algorithms in identification of an AR process corrupted by outliers and burst noise respectively. The burst noise scenario was then considered in the real-data context of speech reconstruction.

Chapter 5: We have relaxed the standard assumption of *known* forgetting factor in on-line Bayesian identification of *non-stationary* AR processes. We derived an algorithm for on-line joint identification of (i) the unknown *time-variant* forgetting factor and (ii) the *non-stationary* parameters of the AR process. We showed that the resulting identification algorithm improves the parameter-tracking abilities of the standard fixed-forgetting Bayesian approach. This was demonstrated in simulation. The derived algorithm constitutes a data-driven procedure for steering the forgetting factor.

Later in Chapter 5, we also considered on-line identification of the MEAR model with time-variant parameters and unknown time-variant forgetting. The resulting algorithm, balances, in effect, the contributions being made by (i) past data (sufficient statistics), (ii) current data (dyadic update), and (iii) expert knowledge (parameters of the alternative distribution). It achieves this on-line.

Chapter 6: We derived two algorithms for Bayesian Principal Component Analysis (PCA). The first one was based on a standard Probabilistic PCA (PPCA) model. We showed that the standard VEM algorithm for evaluation of the VB-statistics can be analytically simplified. The computational cost of the new algorithm—called Fast Variational PCA (FVPCA)—was, therefore, significantly reduced.

We then introduced a new, *orthogonal*, parameterization of the PCA model and presented the associated VB solution: Orthogonal Variational PCA (OVPCA). The model avoids modelling ambiguities inherent in the PPCA model, and so no regularization via priors was needed. In

consequence, identification of this orthogonal model is robust with respect to choice of the prior distributions. Moreover, we were able to derive the posterior distribution of the model rank (i.e. the number of relevant principal components).

Chapter 7: Bayesian identification of the Factor Analysis of Medical Image Sequences (FAMIS) model—introduced in Chapter 3.1—was achieved using VB approximation. We showed that the Bayesian identification improves the standard certainty-equivalence-based method in two main ways: (i) better identification of the noise properties is achieved, and (ii) automatic estimation of the number of relevant physiological factors is possible. This is important in medical applications, since it allows more reliable results to be achieved with data manifesting low Signal-to-Noise Ratio (SNR). Furthermore, there is a reduced reliance on external information provided by medical doctors and imaging experts.

8.3. Further work

8.3.1. Short Term Extensions

1. In Sections 4.7 and 4.8, we used the MEAR model for identification of an AR process corrupted by additive outliers and burst noise respectively. In both cases, the resulting filter-bank was composed of linear filters. The MEAR model is, however, capable of dealing with non-linear distortions of data. It would be interesting to apply the MEAR model, for example, to an AR process suffering a memoryless non-linear distortion, such as occurs frequently in audio applications. A priority would be to design a suitable partitioning of the continuous space of distortions.
2. The standard VEM algorithm for evaluation of the VB-statistics is not guaranteed to converge in a finite number of iterations. Therefore, in an on-line scenario, additional treatment is required to achieve computational feasibility. For non-stationary forgetting (Section 5.2), we used the simple strategy of imposing a maximum allowable number of VEM iterations. We note that the space of unknown forgetting factors is confined to the interval $[0, 1]$. Therefore, more sophisticated strategies could be proposed, for exploration of this finite scalar interval.
3. Identification of an AR process using the non-stationary forgetting technique (Chapter 5) can be easily applied in changepoint detection in noisy speech.
4. Principal Component Analysis (PCA) is used as a standard data processing black-box in many scientific areas. In DSP, it finds an interesting context in the area of sub-space methods and spectral estimation, in algorithms such as MUSIC and ESPRIT [67]. Application of the results achieved in Chapter 6—namely, estimation of the number, r , of relevant principal components, and uncertainty bounds for A , X , and r —is straightforward in these areas, and could potentially bring significant added value.
5. In Section 7.3, we proposed various heuristic techniques for improving the numerical efficiency of the VEM algorithm for the FAMIS model. Implementation of these techniques is straightforward but time-consuming.

8.3.2. Future Research Directions

The Variational Bayes (VB) approximation method was chosen as a trade-off between accuracy and computational feasibility. It is not optimal from the statistical point of view (Section 2.2.4). Its key advantage is that the approximate posterior distributions are available in analytical form, providing computational feasibility. In some cases, such as on-line identification of mixture models, the statistically optimal (Section 2.2.3) approximation was shown to outperform the VB approximation [114]. Recent developments in mean field theory [34, 39] suggest that new, more accurate, approximations of the allowed posteriors may be found. Computationally efficient evaluation of the resulting approximations is a future challenge for the DSP community. Potentially, it is an intriguing one, as new kinds of computational algorithms and flow-of-control may be revealed.

Applicability of the MEAR model (Chapter 4) is limited by the assumption of an *a priori* known filter-bank. We showed that the filter-bank can be designed using analytical insight into the problem. However, this approach can be used only for a limited set of problems, such as 1-D and discretized function spaces. An automated approach would greatly extend applicability of the model. There has already been an attempt at automated filter-bank selection using simplex methods [62].

The VB-approximate posterior distribution of the time-variant forgetting factor in Chapter 5 was found to be intractable. Thus, further approximation was needed to achieve a numerically tractable solution (Proposition 5.1). The impact of this approximation has not been fully explored in this thesis. Also, performance of the method depends on the choice of alternative distribution (Section 3.2.3), which must be known *a priori*. Further work on the treatment of the alternative distribution would greatly enhance the applicability of the method in practice.

Preliminary experiments with the Bayesian identification of the FAMIS model (Chapter 7) are very promising. However, the list of problems that must be solved in order to apply this method in clinical practice is extensive (Section 7.3). For example, better physiological modelling, efficient numerical implementation and robustness improvements need to be addressed. Ultimately, clinical studies using this unified FAMIS framework are necessary.

The VB-based identification of the FAMIS model (Section 7.2) is closely related to the emerging class of algorithms known as Independent Component Analysis (ICA) (Section 3.2.1, Table 7.2). ICA is a statistical technique for decomposing a complicated dataset into independent sub-parts. The FAMIS model is, in fact, a special case of noisy ICA [110], with a non-stationary noise distribution. The ICA method is currently receiving much attention in signal processing because of its flexibility and ability to deal with non-linear models. Furthermore, numerically efficient evaluation of the traditional Maximum Likelihood (ML) approach [51] to ICA has been achieved using fixed-point approximations [112]. The possible use of these approximations within the Variational Bayes (VB) approach may bring significant computational savings in the future.

Appendix A.

Required Probability Distributions

A.1. Matrix Normal distribution

We say matrix X has a matrix Normal distribution, $f(X) = \mathcal{N}(\mu_X, \Sigma_p \otimes \Sigma_n)$, if the matrix $X \in \mathbb{R}^{p \times n}$ has the joint probability density

$$f(X) = (2\pi)^{-pn/2} |\Sigma_p|^{-n/2} |\Sigma_n|^{-p/2} \exp\left(-0.5 \text{tr}\left\{\Sigma_p^{-1} (X - \mu_X) (\Sigma_n^{-1})' (X - \mu_X)'\right\}\right), \quad (\text{A.1})$$

where $\Sigma_p \in \mathbb{R}^{p \times p}$ and $\Sigma_n \in \mathbb{R}^{n \times n}$ are symmetric, positive definite matrices.

The distribution has the following properties:

- first moment is $E_X(X) = \mu_X$,
- second non-central moments are

$$\begin{aligned} E_X(XZX') &= \text{tr}(Z\Sigma_n)\Sigma_p + \mu_X Z \mu_X', \\ E_X(X'ZX) &= \text{tr}(Z\Sigma_p)\Sigma_n + \mu_X' Z \mu_X, \end{aligned} \quad (\text{A.2})$$

where Z is an arbitrary matrix of appropriate sizes respectively,

- For any matrices $C \in \mathbb{R}^{c \times p}$ and $D \in \mathbb{R}^{n \times d}$ it holds:

$$f(CXD) = \mathcal{N}(C\mu_X D, C\Sigma_p C' \otimes D'\Sigma_n D). \quad (\text{A.3})$$

- distribution of $\text{vec}(X)$ is again Normal with

$$f(\text{vec}(X)) = \mathcal{N}(\text{vec}(\mu_X), \Sigma_n \otimes \Sigma_p).$$

Note that covariance matrix has changed the form compared to the matrix case. This notation is helpful as it allows to store the $pn \times pn$ covariance matrix in $p \times p$ and $n \times n$ structures.

This convention greatly simplifies notation, e.g. if columns \mathbf{x}_i of matrix X are independently distributed with the same Normal pdf

$$f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \prod_{i=1}^n \mathcal{N}(\mu_i, \Sigma) \equiv \mathcal{N}(\mu_X, \Sigma \otimes I_n) = f(X). \quad (\text{A.4})$$

Moreover, linear transformation of matrix argument X (A.3) preserves the decomposed form. This allows all operations on moments of X being done using matrix algebra.

A.2. Normal-Wishart Distribution

The Normal-Wishart distribution of variable $\theta = [A, \Omega]$ has pdf

$$\mathcal{NW}_{A,\Omega}(V, \nu) \equiv \frac{|\Omega|^{\frac{1}{2}\nu}}{\zeta_{\mathcal{NW}}(V, \nu)} \exp \left\{ -\frac{1}{2} \Omega [-I_p, A] V [-I_p, A]' \right\}, \quad (\text{A.5})$$

with normalizing constant

$$\zeta_{\mathcal{NW}}(V, \nu) = \Gamma_p \left(\frac{1}{2} (\nu - r + p + 1) \right) |\Lambda|^{-\frac{1}{2}(\nu - r + p + 1)} |V_{aa}|^{-0.5p} 2^{0.5p(\nu + p + 1)} \pi^{\frac{r}{2}}, \quad (\text{A.6})$$

and auxiliary values

$$V = \begin{bmatrix} V_{dd} & V'_{ad} \\ V_{ad} & V_{aa} \end{bmatrix}, \quad \Lambda = V_{dd} - V'_{ad} V_{aa}^{-1} V_{ad}, \quad (\text{A.7})$$

where (A.7) denotes partitioning of $V \in \mathfrak{R}^{(p+1) \times (p+1)}$ into blocks and V_{dd} is the upper left sub-block of size $p \times p$.

Marginal distributions of A and Ω are [16]:

$$f(A|\Omega, V, \nu) = \mathcal{N} \left(\hat{A}, \Omega^{-1} \otimes V_{aa}^{-1} \right), \quad (\text{A.8})$$

$$f(\Omega|V, \nu) = \mathcal{W} \left(\frac{\nu - p + 1}{2}, 2\Lambda^{-1} \right), \quad (\text{A.9})$$

$$f(A|V, \nu) = \mathcal{St} \left(\hat{A}, \frac{1}{\eta} \Lambda^{-1} \otimes V_{aa}^{-1}, \eta \right), \quad (\text{A.10})$$

with auxiliary constants

$$\hat{A} = V'_{ad} V_{aa}^{-1}, \quad (\text{A.11})$$

$$\eta = \nu - r + p + 1. \quad (\text{A.12})$$

\mathcal{St} denotes the matrix Student- t distribution with η degrees of freedom, and \mathcal{W} denotes the Wishart distribution [115].

The moments of these distributions are:

$$E_{A|\Omega}(A) = E_A(A) = \widehat{A}, \tag{A.13}$$

$$E_{\Omega}(\Omega) \equiv \widehat{\Omega}_t = \frac{1}{\eta} \Lambda^{-1}, \tag{A.14}$$

$$E_A \left((A - \widehat{A})(A - \widehat{A})' \right) = \frac{1}{\nu - r} \Lambda V_{aa}^{-1}, \tag{A.15}$$

$$E_{\Omega} \left((\Omega - \widehat{\Omega})^2 \right) = \frac{2\Lambda}{(\nu - r)^2}, \tag{A.16}$$

$$E_{\Omega}(\ln |\Omega|) = \frac{1}{2} \prod_{j=1}^p \psi \left(\frac{1}{2} (\nu - r + p - j) \right) \tag{A.17}$$

$$- \frac{1}{2} \ln |\Lambda| + \frac{1}{2} p \ln 2. \tag{A.18}$$

Here, conditioning by V, ν was dropped for simplicity.

A.3. Dirichlet Distribution

The Dirichlet distribution of the vector variable α has pdf

$$f(\alpha|\beta) = \mathcal{D}_{i\alpha}(\beta) = \frac{1}{\zeta_{\mathcal{D}_i}(\beta)} \prod_{i=1}^c \alpha_i^{\beta_i - 1}, \tag{A.19}$$

with vector parameter $\beta = [\beta_1, \beta_2, \dots, \beta_c]'$, normalizing constant

$$\zeta_{\mathcal{D}_i}(\beta) = \frac{\prod_{i=1}^c \Gamma(\beta_i)}{\Gamma(\gamma)}, \tag{A.20}$$

where $\gamma = \sum_{i=1}^c \beta_i$, and with first moment given by:

$$\hat{\alpha}_i = E_{\alpha|\beta}(\alpha_i) = \frac{\beta_i}{\gamma}, \quad i = 1, \dots, c. \tag{A.21}$$

Expected value of the logarithm is

$$\widehat{\ln \alpha_i} = E_{\alpha|\beta}(\ln \alpha_i) = \psi(\beta_i) - \psi(\gamma), \tag{A.22}$$

where $\psi(\beta) = \frac{\partial}{\partial \beta} \ln \Gamma(\beta)$.

A.4. Truncated Normal Distribution

The truncated normal distribution of scalar random variable x is defined as normal—with functional form $\mathcal{N}(\mu, s^2)$ —on a restricted support $a \leq x \leq b$. Its pdf is

$$f(x|s, a, b) = \frac{\sqrt{2} \exp\left(-\frac{1}{2} \left(\frac{x-\mu}{s}\right)^2\right)}{s\sqrt{\pi} (\operatorname{erf}(\beta) - \operatorname{erf}(\alpha))} \chi((a, b]), \quad (\text{A.23})$$

where $\alpha = \frac{a-\mu}{s\sqrt{2}}$, $\beta = \frac{b-\mu}{s\sqrt{2}}$. Moments of (A.23) are

$$\widehat{x} = \mu - s\varphi(\mu, s), \quad (\text{A.24})$$

$$\widehat{x^2} = s^2 + \mu\widehat{x} - s\kappa(\mu, s), \quad (\text{A.25})$$

with auxiliary functions:

$$\varphi(\mu, s) = \frac{\sqrt{2} [\exp(-\beta^2) - \exp(-\alpha^2)]}{\sqrt{\pi} (\operatorname{erf}(\beta) - \operatorname{erf}(\alpha))}, \quad (\text{A.26})$$

$$\kappa(\mu, s) = \frac{\sqrt{2} [b \exp(-\beta^2) - a \exp(-\alpha^2)]}{\sqrt{\pi} (\operatorname{erf}(\beta) - \operatorname{erf}(\alpha))}. \quad (\text{A.27})$$

(A.26) and (A.27) with vector arguments—e.g. $\kappa(\mathbf{m}, s)$ —are evaluated element-wise. Confidence intervals for this distribution can also be obtained. However, for simplicity, we use the first two moments, (A.24) and (A.25), to approximate (A.23) by a Gaussian. The Maximum Entropy (MaxEnt) principle [116] ensures, that uncertainty bounds on the MaxEnt Gaussian approximation of (A.23), enclose the uncertainty bounds of all distributions with the same first two moments. Hence,

$$\max\left(a, -2\sqrt{\widehat{x^2} - \widehat{x}^2}\right) < x - \widehat{x} < \min\left(b, 2\sqrt{\widehat{x^2} - \widehat{x}^2}\right). \quad (\text{A.28})$$

A.5. Von Mises-Fisher Matrix distribution

Moments of the von Mises-Fisher matrix distribution are now considered. Proofs of all unproven results are available in [108].

A.5.1. Definition

The von Mises-Fisher probability density function of matrix random variable, $Z \in \mathbb{R}^{p \times n}$, restricted to $Z^T Z = I_t$, is given by:

$$f(Z|F) = \mathcal{M}(F) = \frac{1}{\zeta(p, FF')} \exp(\operatorname{tr}(FZ')), \quad (\text{A.29})$$

$$\zeta(p, FF') = {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}FF'\right) C(p, n), \quad (\text{A.30})$$

where $F \in \mathfrak{R}^{p \times n}$ is a matrix parameter of the same dimensions as Z , and $p \geq n$. $\zeta(p, FF')$ is the normalizing coefficient. ${}_0F_1(\cdot)$ denotes a hypergeometric function of matrix argument FF' [117]. $C(p, r)$ denotes the area of the relevant Stiefel manifold $\mathcal{S}_{p,n}$ (6.55).

(A.29) is a Gaussian distribution with restriction $Z'Z = I_t$, re-normalized on $\mathcal{S}_{p,n}$. It is governed by a single matrix parameter F . Consider the (economic) SVD decomposition

$$F = U_F L_F V_F',$$

of the parameter F , where $U_F \in \mathfrak{R}^{p \times n}$, $L_F \in \mathfrak{R}^{n \times n}$, $V_F \in \mathfrak{R}^{n \times n}$. Then, maximum of (A.29) is reached at

$$\hat{Z} = U_F V_F'. \quad (\text{A.31})$$

Flatness of the distribution is controlled by L_F . When $\text{diag}(L_F) = \mathbf{0}_{n,1}$ the distribution is uniform on $\mathcal{S}_{p,n}$ [118]. For $l_{F,i} \rightarrow \infty$, $\forall i = 1 \dots n$ the distribution is a Dirac delta function at \hat{Z} (A.31).

A.5.2. First Moment

Let Y be the transformed variable

$$Y = U_F' Z V_F, \quad (\text{A.32})$$

It can be shown that $\zeta(p, FF') = \zeta(p, L_F^2)$. The pdf of Y is then:

$$f(Y|F) = \frac{1}{\zeta(p, L_F^2)} \exp(\text{tr}(L_F Y)) = \frac{1}{\zeta(p, L_F^2)} \exp(\mathbf{l}_F' \mathbf{y}), \quad (\text{A.33})$$

where $\mathbf{y} = \text{diag}(Y)$. Hence,

$$f(Y|F) \propto f(\mathbf{y}|\mathbf{l}_F). \quad (\text{A.34})$$

First moment of (A.33) is given by [108]:

$$\mathbb{E}(Y|L_F) = \Psi, \quad (\text{A.35})$$

where $\Psi = \text{diag}(\psi)$ is a diagonal matrix with diagonal:

$$\psi_i = \frac{\partial}{\partial l_{F,i}} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}L_F^2\right). \quad (\text{A.36})$$

We will denote function (A.36) as

$$\psi = G(p, \mathbf{l}_F), \quad (\text{A.37})$$

The mean value of the original random variable Z is then [119]:

$$\mathbb{E}(Z) = U_F \Psi V_F = U_F G(p, L_F) V_F, \quad (\text{A.38})$$

where $G(p, L_F) = \text{diag}(G(p, \mathbf{l}_F))$.

A.5.3. Second Moment and Uncertainty Bounds

The second central moment of the transformed variable $\mathbf{y} = \text{diag}(Y)$ (A.33) is given by

$$\mathbb{E}(\mathbf{y}\mathbf{y}' - \mathbb{E}(\mathbf{y})\mathbb{E}(\mathbf{y})') = \Phi, \quad (\text{A.39})$$

with elements,

$$\phi_{i,j} = \frac{\partial}{\partial l_{F,i} \partial l_{F,j}} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}L_F^2\right), \quad i, j = 1, \dots, r. \quad (\text{A.40})$$

Transformation (A.32) is one-to-one, with unit Jacobian. Hence, boundaries of confidence intervals on variables Y and Z can be mutually mapped using (A.32). However, mapping $\mathbf{y} = \text{diag}(Y)$ is many-to-one, and so $Z \rightarrow \mathbf{y}$ is surjective. Conversion of second moments (and uncertainty bounds) of \mathbf{y} to Z (via (A.32), (A.33)) is therefore available in implicit form only. For example, the upper bound subspace of Z is expressible as follows:

$$\bar{Z} = \{Z \mid \text{diag}(U_F' Z V_F) = \bar{\mathbf{y}}\},$$

where $\bar{\mathbf{y}}$ is an appropriately chosen upper bound on \mathbf{y} . The lower bound, Z , is similarly constructed via a bound $\underline{\mathbf{y}}$.

It remains then, to choose appropriately bounds $\underline{\mathbf{y}}$ and $\bar{\mathbf{y}}$ from (A.33). Exact confidence intervals for this multivariate distribution are not known to us. Therefore we use the first two moments, (A.35) and (A.39), to approximate (A.33) by a Gaussian. The Maximum Entropy (MaxEnt) principle [116] ensures, that uncertainty bounds on the MaxEnt Gaussian approximation of (A.33), enclose the uncertainty bounds of all distributions with the same first two moments. Confidence intervals for the Gaussian distribution, with moments (A.36), (A.40) are well known, e.g.

$$\Pr\left(-2\sqrt{\phi_i} < (y_i - \psi_i) < 2\sqrt{\phi_i}\right) \doteq 0.95, \quad (\text{A.41})$$

where ψ_i is given by (A.36), and ϕ_i by (A.40). Therefore, we choose

$$\bar{y}_i = \psi_i + 2\sqrt{\phi_i}, \quad (\text{A.42})$$

$$\underline{y}_i = \psi_i - 2\sqrt{\phi_i}. \quad (\text{A.43})$$

The required vector bounds are then constructed as $\bar{\mathbf{y}} = [\bar{y}_1, \dots, \bar{y}_r]'$, and ditto for $\underline{\mathbf{y}}$. The geometric relationship variables Z and \mathbf{y} is illustrated graphically for $p = 2$ and $n = 1$ in Figure 6.6.

A.6. Gamma Distribution

The Gamma distribution has pdf

$$f(x|a, b) = \mathcal{G}(a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \chi([0, \infty)), \quad (\text{A.44})$$

where $a > 0$, and $b > 0$, and $\Gamma(a)$ is the Gamma function [71] evaluated at a . Its first moment is:

$$\hat{x} = \frac{a}{b},$$

and the second central moment is:

$$E_x \left((x - \hat{x})^2 \right) = \frac{a}{b^2}.$$

A.7. Truncated Exponential Distribution

The Truncated Exponential Distribution has pdf

$$f(x|k, [a, b]) = \frac{k}{\exp(kb) - \exp(ka)} \exp(xk) \chi([a, b]), \quad (\text{A.45})$$

where $a < b$ are boundaries of the support. Its first moment is

$$\hat{x} = \frac{\exp(bk)(1 - bk) - \exp(ak)(1 - ak)}{k(\exp(ak) - \exp(bk))}, \quad (\text{A.46})$$

which is not defined for $k = 0$. Limit at this point is

$$\lim_{k \rightarrow 0} \hat{x} = \frac{a + b}{2}.$$

Which is not surprising, as the distribution becomes uniform on interval $[a, b]$.

A. REQUIRED PROBABILITY DISTRIBUTIONS

A.5.3i The second central moment of the transformed variable $y = \text{diag}(Y)$ (A.33) is given by

$$E(yy^T - Z(y)Z^T(y)) = \Phi, \quad \text{with the second central moment}$$

$$\frac{\partial^2}{\partial y_i \partial y_j} = \left(\frac{\partial}{\partial z} \right)^T \Phi \left(\frac{\partial}{\partial z} \right),$$

with elements

$$\phi_{ij} = \frac{\partial^2}{\partial z_i \partial z_j} \ln \mathcal{P}_i \left(\frac{1}{z_i}, \frac{1}{z_j} \right), \quad i, j = 1, \dots, r. \quad (\text{A.40})$$

Transformation (A.33) is one-to-one, with unit Jacobian. Hence, confidence intervals on variables Y and Z can be mutually mapped using (A.33) and its inverse. Conversion of equal moment (and uncertainty) bounds of y to Z (A.33), (A.35) is therefore (A.36) or (A.37) in the upper bound subspace of Z is expressible as follows

$$Z = \left\{ Z \mid \text{diag} \left(\frac{1}{Z} \right) \in \mathcal{Y} \right\} \quad (\text{A.36})$$

where \mathcal{Y} is an appropriately chosen upper bound on y . The lower bound Z , is similarly constructed via a bound \underline{y} .

It remains then, to choose appropriate bounds \underline{y} and \bar{y} from (A.33). Exact confidence intervals for this multivariate distribution are not known to us. Therefore we use the first two moments, (A.35) and (A.29), to approximate (A.33) by a Gaussian. The Maximum Entropy (MaxEnt) principle [116] ensures that uncertainty bounds are maximally conservative, i.e. they contain the true uncertainty bounds of all distributions with the same first two moments. Confidence intervals for the Gaussian distribution, with moments (A.35), (A.40) are well known, e.g.

$$\Pr \left(-2\sqrt{\phi_i} < (y_i - \psi_i) < 2\sqrt{\phi_i} \right) \approx 0.95, \quad (\text{A.41})$$

where ψ_i is given by (A.36), and ϕ_i by (A.40). Therefore, we choose

$$\bar{y}_i = \psi_i + 2\sqrt{\phi_i}, \quad (\text{A.42})$$

$$\underline{y}_i = \psi_i - 2\sqrt{\phi_i}. \quad (\text{A.43})$$

The required vector bounds are then constructed as $\bar{y} = [\bar{y}_1, \dots, \bar{y}_r]^T$, and ditto for \underline{y} . The geometric relationship variables Z and y is illustrated graphically for $r = 2$ and $n = 1$ in Figure 6.6.

A.6. Gamma Distribution

The Gamma distribution has pdf

$$f(x|a, b) = \mathcal{G}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \chi([0, \infty]), \quad (\text{A.44})$$

Appendix B.

Analytical Solution of Fast VPCA, Using MAPLE

Here, we provide analytical solution of Fast VPCA as obtained using software package Maple. Hence, we will present them as commented Maple code, i.e. equations are not numbered.

Initialize the Maple environment and invoke some basic assumptions.

```
> restart;
> assume(k[A]>0); assume(k[X]>0); assume(sigma[A]>0);
> assume(sigma[X]>0); assume(b>0);
> assume(omega>0); assume(n>0); assume(p>0);
> assume(ld>0); assume(alpha>0);
```

Variational equations for Fast VPCA

```
> eq1:=k[A]-omega*1[D]*k[X]*sigma[A]:
> eq2:=sigma[A] - 1/(omega*(n*sigma[X]+k[X]^2)+upsilon):
> eq3:=k[X]-omega*sigma[X]*k[A]*1[D]:
> eq4:=sigma[X] - 1/(omega*(p*sigma[A]+(k[A])^2)+1):
> eq5:=upsilon - (p)/(p*sigma[A]+k[A]^2+b):
> EQ:={eq1,eq2,eq3,eq4,eq5};
```

$$EQ := \left\{ k_A - \omega l_D k_X \sigma_A, v - \frac{p}{p \sigma_A + k_A^2 + b}, \sigma_X - \frac{1}{\omega (p \sigma_A + k_A^2) + 1}, k_X - \omega \sigma_X k_A l_D, \sigma_A - \frac{1}{\omega (n \sigma_X + k_X^2) + v} \right\}$$

```
> S:=solve(EQ,{k[A],sigma[A],k[X],sigma[X],upsilon}):
```

B.1. Closed-form Solution

The first mode (zero-centered):

```
> S1:=S[1];
```

$$S1 := \left\{ v = -\frac{(-n + n \%1 + p) \%1 \omega}{-1 + \%1}, \sigma_X = \%1, k_A = 0, k_X = 0, \right. \\ \left. \sigma_A = -\frac{-1 + \%1}{\%1 \omega p} \right\} \\ \%1 := \text{RootOf}((-n + bn\omega)_{-Z^2} + (2n - bn\omega + b\omega p)_{-Z} - n)$$

The second mode:

> S2:=S[2];

$$S2 := \left\{ \sigma_X = -\frac{p - \omega l_D^2}{\omega l_D^2 (\omega \%1^2 + 1)}, \sigma_A = -\frac{\omega \%1^2 + 1}{\omega (p - \omega l_D^2)}, k_A = \%1, v = \right. \\ \left. \frac{(-n\omega \%1^2 - n + \omega \%1^2 p + \omega l_D^2) n\omega}{(\%1^2 b\omega^2 p - \omega \%1^2 p - \%1^2 bn\omega^2 + \%1^2 \omega^2 l_D^2 - bn\omega + n\omega \%1^2 + n + b\omega^2 l_D^2)}, \right. \\ \left. k_X = -\frac{(p - \omega l_D^2) \%1}{l_D (\omega \%1^2 + 1)} \right\} \\ \%1 := \text{RootOf}(_Z^4 n\omega^3 l_D^2 + (-p^2 \omega - bn\omega^2 p + 2\omega^2 l_D^2 p + bn\omega^3 l_D^2 \\ + n\omega^2 l_D^2 - \omega^3 l_D^4 - b\omega^3 l_D^2 p + n\omega p + b\omega^2 p^2)_{-Z^2} - b\omega^3 l_D^4 \\ + pn - bn\omega p + b\omega^2 l_D^2 p + bn\omega^2 l_D^2)$$

B.2. Determination of Acceptable Solutions

Note that solution of both modes depends on evaluation of fourth order polynomial in variable $_Z$. However, it can be rewritten as second order polynomial in $_Z^2$. Roots of second order polynomials are easy to evaluate, using formula:

$$> z = (-a[1] + \sqrt{R}) / (2*a[2]); \quad R := a[1]^2 - 4*a[0]*a[2];$$

$$z = \frac{1}{2} \frac{-a_1 + \sqrt{R}}{a_2}$$

$$R := a_1^2 - 4a_0 a_2$$

An important limiting condition is then value of auxiliary variable R, which must be positive. We will analyze this for both modes.

Zero-centered mode

In this mode, mean values are zero. It remains to make sure that variances are real and positive.

$$> a[2] := (-n + b*n*\omega): \quad a[1] := (2*n - b*n*\omega + b*\omega*p): \quad a[0] := -n:$$

$$> R := a[1]^2 - 4*a[2]*a[0]: R := \text{collect}(\text{simplify}(R), \{b, \omega\});$$

$$R := (n^2 - 2pn + p^2)\omega^2 b^2 + 4bn\omega p$$

Clearly, R is always positive, hence, roots are always real.

Values of the roots are then:

```

> z1:=(-a[1]+sqrt(R))/(2*a[2]);
      z1 := 1/2 * (-2n + bn*omega - b*omega*p + sqrt(4*bn*omega*p + b^2*n^2*omega^2 - 2*b^2*n*omega^2*p + b^2*omega^2*p^2)) / (-n + bn*omega)
> z2:=(-a[1]-sqrt(R))/(2*a[2]);
      z2 := 1/2 * (-2n + bn*omega - b*omega*p - sqrt(4*bn*omega*p + b^2*n^2*omega^2 - 2*b^2*n*omega^2*p + b^2*omega^2*p^2)) / (-n + bn*omega)
> limit(z1,b=0);limit(z2,b=0);
      1
      1
    
```

Note, that the roots evaluated above are, in fact, values of σ_X , which has prior value set to 1. Not surprisingly, the limit for very small b is one. The first root, z_1 , approaching from above, the second one, z_2 , from below. Note that sign of σ_A depends on $(-\sigma_X + 1)$, hence only the lower root, z_2 , is valid.

The second mode:

First, let us analyze roots of the polynomial, using the same formula as above.

```

> a[2]:=n*omega^3*[D]^2;
> a[1]:=(n*omega*p-p^2*omega+2*omega^2*[D]^2*p+n*omega^2*[D]^2+b*n*ome
> ga^3*[D]^2-omega^3*[D]^4-b*n*omega^2*p-b*omega^3*[D]^2*p+b*omega^2*
> p^2);
> a[0]:=n*p-b*omega^3*[D]^4-b*n*omega*p+b*omega^2*[D]^2*p+b*n*omega^2*
> 1[D]^2;
> R:=a[1]^2-4*a[2]*a[0];
    
```

$$\begin{aligned}
 R := & (-p^2\omega - bn\omega^2p + 2\omega^2l_D^2p + bn\omega^3l_D^2 + n\omega^2l_D^2 - \omega^3l_D^4 \\
 & - b\omega^3l_D^2p + n\omega p + b\omega^2p^2)^2 \\
 & - 4n\omega^3l_D^2(pn - b\omega^3l_D^4 - bn\omega p + b\omega^2l_D^2p + bn\omega^2l_D^2)
 \end{aligned}$$

```

> sp:=solve({R},{1[D]});
    
```

$$\begin{aligned}
 sp := & \{l_D = \frac{\sqrt{\omega p}}{\omega}\}, \{l_D = -\frac{\sqrt{\omega p}}{\omega}\}, \{l_D = \frac{\sqrt{\omega p}}{\omega}\}, \{l_D = -\frac{\sqrt{\omega p}}{\omega}\}, \\
 & \{l_D = \frac{\sqrt{-\omega(b\omega p - p - n + 2\sqrt{pn} + bn\omega - 2\omega b\sqrt{pn})}}{\omega}\}, \\
 & \{l_D = -\frac{\sqrt{-\omega(b\omega p - p - n + 2\sqrt{pn} + bn\omega - 2\omega b\sqrt{pn})}}{\omega}\}, \\
 & \{l_D = \frac{\sqrt{-\omega(-p - n - 2\sqrt{pn} + b\omega p + bn\omega + 2\omega b\sqrt{pn})}}{\omega}\}, \\
 & \{l_D = -\frac{\sqrt{-\omega(-p - n - 2\sqrt{pn} + b\omega p + bn\omega + 2\omega b\sqrt{pn})}}{\omega}\}
 \end{aligned}$$

Clearly only some of these are positive:

```

> simplify({sp[1,1],sp[5,1],sp[7,1]});
    
```

$$\{l_D = \frac{(\sqrt{p} + \sqrt{n})\sqrt{1 - \omega b}}{\sqrt{\omega}}, l_D = \frac{\sqrt{1 - \omega b}|\sqrt{p} - \sqrt{n}|}{\sqrt{\omega}}, l_D = \frac{\sqrt{p}}{\sqrt{\omega}}\}$$

The first singular point is notable, as it is also singular point for values of σ_A, σ_X . Value of l_D must be higher than this limit.

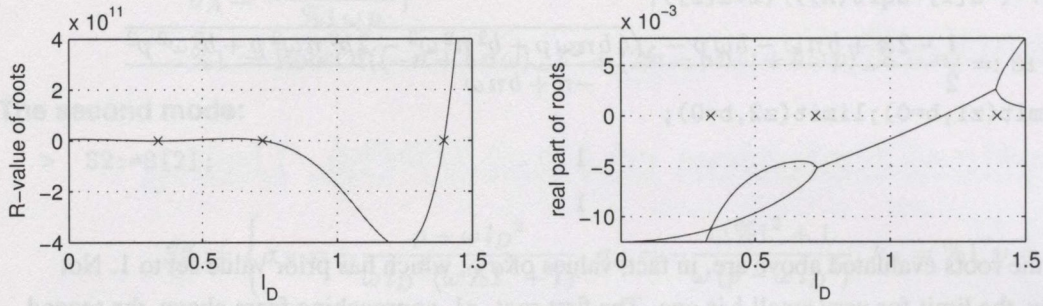


Figure B.1.: Analysis of singular points of roots of second mode solution of fast VPCA, singular points are denoted by cross.

An example of R-value and both roots for certain numerical values is displayed in Figure B.1. Hence, the boundary for acceptable solution is at:

$$l_D = \frac{(\sqrt{n} + \sqrt{p})\sqrt{1 - b\omega}}{\sqrt{\omega}}$$

For values lower than this bound, the second mode has no acceptable solution, and iterative algorithm than converge to zero-centered solutions.

Singular point of upsilin is:

```
> spu:=n+n*omega*RootOf(_Z^4*n*omega^3*1[D]^2+(n*omega*p-p^2*omega+2*p*
> omega^2*1[D]^2+n*omega^2*1[D]^2+b*n*omega^3*1[D]^2-omega^3*1[D]^4-b*n*
> omega^2*p-b*omega^3*1[D]^2*p+b*omega^2*p^2)*_Z^2+n*p-b*omega^3*1[D]^4-
> b*n*omega*p+b*omega^2*1[D]^2*p+b*n*omega^2*1[D]^2)^2-omega*RootOf(_Z^4
> *n*omega^3*1[D]^2+(n*omega*p-p^2*omega+2*p*omega^2*1[D]^2+n*omega^2*1[
> D]^2+b*n*omega^3*1[D]^2-omega^3*1[D]^4-b*n*omega^2*p-b*omega^3*1[D]^2*
> p+b*omega^2*p^2)*_Z^2+n*p-b*omega^3*1[D]^4-b*n*omega*p+b*omega^2*1[D]^
> 2*p+b*n*omega^2*1[D]^2)^2*p-omega*1[D]^2:
> solve(spu, {1[D]});
```

$$\{l_D = \frac{\sqrt{\omega p}}{\omega}\}, \{l_D = -\frac{\sqrt{\omega p}}{\omega}\}, \{l_D = 0\}$$

Clearly, it is the same singular point as for σ_A and σ_X , and it is to be compared with the singular point for R-value. However, we expect b to be chosen as small as possible, hence, we will consider

$$l_D = \frac{(\sqrt{n} + \sqrt{p})}{\sqrt{\omega}}$$

to be the limit for acceptable solution of the second mode.

Roots of the polynomial are then:

- > z1:=(-a[1]+sqrt(R))/(2*a[2]):
- > z2:=(-a[1]-sqrt(R))/(2*a[2]):

With limits:

- > limit(sqrt(z1)/l[D],l[D]=infinity);
 $\frac{1}{\sqrt{n}}$
- > limit(sqrt(z2)/l[D],l[D]=infinity);
 0

We see, that the first root is asymptotically approaching l_D/\sqrt{n} , which is intuitively appealing. The second possible root is approaching zero quite rapidly.

C.1. Hypergeometric Function of Scalar Argument

The natural logarithm (\ln) of the hypergeometric function, ${}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right)$, of a scalar argument can be expressed as

$$\ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = \ln B\left(\frac{1}{2}p - 1, s\right) + \left(\frac{1}{2}p - 1\right) (\ln 2 - \ln(s)) - \ln \Gamma\left(\frac{1}{2}p\right), \quad (C.1)$$

where B denotes the modified Bessel function of the first kind [71]. (C.1) is plotted as a function of s in Figure C.1 (left), for $p = 5$. The first two derivatives of (C.1),

$$\frac{d}{ds} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = -\frac{B\left(\frac{1}{2}p, 2s\right)}{B\left(\frac{1}{2}p - 1, 2s\right)}, \quad (C.2)$$

$$\frac{d^2}{ds^2} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = -\frac{B\left(\frac{1}{2}p + 1, 2s\right)}{B\left(\frac{1}{2}p - 1, 2s\right)} - 4 \left[\frac{B\left(\frac{1}{2}p, 2s\right)}{B\left(\frac{1}{2}p - 1, 2s\right)} \right]^2 + 2 \frac{B\left(\frac{1}{2}p, 2s\right)}{s B\left(\frac{1}{2}p - 1, 2s\right)}. \quad (C.3)$$

The first derivative is illustrated in Figure C.1 (right), for the same case, $p = 5$. (C.2) can be expressed

B. ANALYTICAL SOLUTION OF FAST VPCA, USING MAPLE

The first singular point is notable, as it is also singular point for α and β and is higher than this limit.

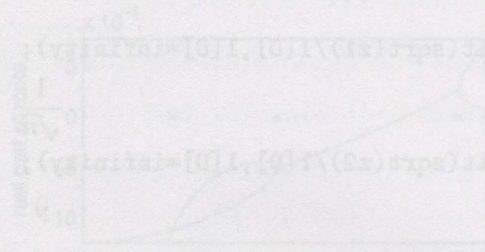
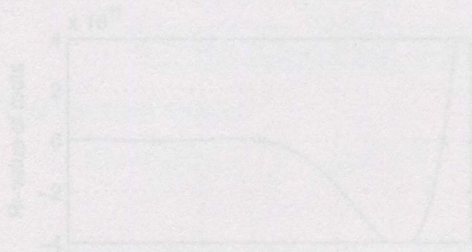


Figure B.1: Analysis of singular points of roots of \sqrt{b} , which is intuitively expected. We see that the first root is asymptotically approaching \sqrt{b} , while the second possible root is approaching zero quite rapidly. The singular point for \sqrt{b} is denoted by a vertical dashed line.

An example of R-value and both roots for certain numerical values is displayed in Figure B.1. Hence, the boundary for acceptable solution is at:

$$l_D = \frac{(\sqrt{b} + \sqrt{p})\sqrt{b-p}}{\sqrt{d}}$$

For values lower than this bound, the second mode has no acceptable solution, and iterative algorithm that converge to zero-centered solutions.

Singular point of ω is:

```

> sp1:=b+n*omega+RootOf(2^2*a+n*omega^3+1[D]^2*(u+omega+p-p^2+omega+2*p*
> omega^2+1[D]^2+n*omega^2+1[D]^3+n*omega^3+1[D]^2-omega^3+1[D]^4-b+n*
> omega^2*p-b+omega^3+1[D]^2+p+b*omega^2*p^2)+2^2+n+p-b+omega^3+1[D]^4-
> b+n*omega+p+b+omega^2+1[D]^2+p+b*omega^2+1[D]^2)-b*omega+RootOf(2^2*
> n*omega^3+1[D]^2*(n+omega*p-p^2+omega+2*p*omega^2+1[D]^2+n*omega^2+1[
> D]^2+b+n*omega^3+1[D]^2-omega^3+1[D]^4-b+n*omega^2*p-b+omega^3+1[D]^2+
> p+b*omega^2*p^2)+2^2+n+p-b+omega^3+1[D]^4-b+n*omega+p+b+omega^3+1[D]^
> 2*p+b+n*omega^2+1[D]^2+2*p-omega+1[D]^2);
> solve(sp1,1[D]);

```

$$l_D = \frac{\sqrt{b}}{\omega}, \quad l_D = -\frac{\sqrt{b}}{\omega}, \quad l_D = 0$$

Clearly, it is the same singular point as for α and β , and it is to be compared with the singular point for R-value. However, we expect b to be chosen as small as possible, hence, we will consider

$$l_D = \frac{(\sqrt{n} + \sqrt{p})}{\sqrt{d}}$$

to be the limit for acceptable solution of the second mode.

Roots of the polynomial are then:

Appendix C.

Hypergeometric Functions

Numerical evaluation of the OVPCA algorithm requires evaluation of the following transformations of the hypergeometric function ${}_0F_1$ of matrix argument: (i) its natural logarithm (ln), for Bayesian rank selection (6.104), and (ii) the first derivative of the ln, required for the first moment of the von Mises-Fisher distribution (A.36). Analytical closed form solutions are not known to us. Recently, a very good approximation of ${}_0F_1$ of matrix argument was developed [120]. It is based on the Laplace approximation at the saddle point. It yields reliable results for use in (i). Unfortunately, the first derivative of ln of this approximation for higher singular values, i.e. $l_i \gg 1$, are greater than one, thus placing the corresponding mean value $E(y_i|l_F)$ (A.36) outside of the unit circle, which is not permissible (Figure 6.6). Therefore, we now develop an approximation which overcomes this difficulty, by first considering the hypergeometric function ${}_0F_1$ of scalar argument.

C.1. Hypergeometric Function of Scalar Argument

The natural logarithm (ln) of the hypergeometric function, ${}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right)$, of a scalar argument can be expressed as

$$\ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = \ln \mathcal{B}\left(\frac{1}{2}p - 1, s\right) + \left(\frac{1}{2}p - 1\right) (\ln 2 - \ln(s)) + \ln \Gamma\left(\frac{1}{2}p\right), \quad (C.1)$$

where \mathcal{B} denotes the modified Bessel function of the first kind [71]. (C.1) is plotted as a function of s in Figure C.1 (left), for $p = 5$. The first two derivatives of (C.1):

$$\frac{d}{ds} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = 2 \frac{\mathcal{B}\left(\frac{1}{2}p, 2s\right)}{\mathcal{B}\left(\frac{1}{2}p - 1, 2s\right)}, \quad (C.2)$$

$$\frac{d^2}{ds^2} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = 4 \frac{\mathcal{B}\left(\frac{1}{2}p + 1, 2s\right)}{\mathcal{B}\left(\frac{1}{2}p - 1, 2s\right)} - 4 \left[\frac{\mathcal{B}\left(\frac{1}{2}p, 2s\right)}{\mathcal{B}\left(\frac{1}{2}p - 1, 2s\right)} \right]^2 + 2 \frac{\mathcal{B}\left(\frac{1}{2}p, 2s\right)}{s \mathcal{B}\left(\frac{1}{2}p - 1, 2s\right)}. \quad (C.3)$$

The first derivative is illustrated in Figure C.1 (right), for the same case, $p = 5$. (C.2) can be expressed

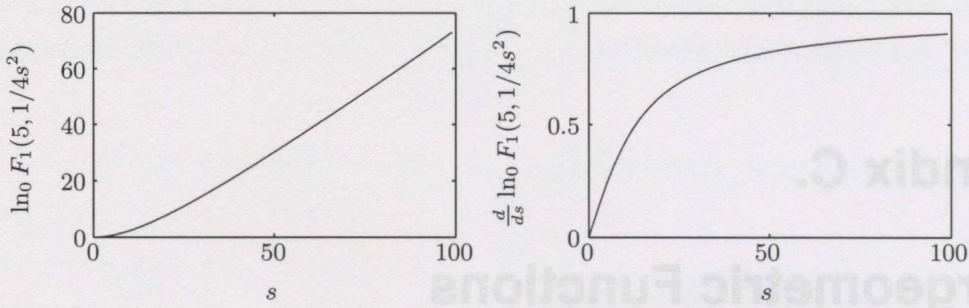


Figure C.1.: $\ln {}_0F_1\left(5, \frac{1}{4}s^2\right)$ of scalar argument s (left), and $\frac{d}{ds} {}_0F_1\left(5, \frac{1}{4}s^2\right)$ (right).

as a continuous fraction expansion [71]:

$$\frac{d}{ds} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = \frac{s_i}{\frac{p}{2} \left[1 + \frac{\frac{1}{4}s_i^2}{\left(\frac{p}{2}+1\right)\left(\frac{p}{2}+2\right) \left[1 + \frac{\frac{1}{4}s_i^2}{\left(\frac{p}{2}+2\right)\left(\frac{p}{2}+3\right) + [1+\dots]} \right]} \right]}. \quad (C.4)$$

Furthermore, (C.3) can be expressed in terms of (C.2) and, therefore (C.4). The evaluation of expansion (C.4) converges very fast for $s < p$. However, when $s \gg p$ (say $s > 10p$) the convergence is quite slow. For large s , a more numerically efficient approximation is obtained via a Taylor expansion of (C.2) at $s \rightarrow \infty$:

$$\frac{d}{ds} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = 1 - \left(\frac{p-1}{2s}\right) \exp\left(-\frac{p-3}{4s}\right) + o(5). \quad (C.5)$$

Here, $o(5)$ denotes elements of the serie in terms of s^{-5} . This expansion provides an excellent approximation in the case $s \gg p$.

C.2. Approximation of ${}_0F_1$ of Matrix Argument by ${}_0F_1$ of Scalar Arguments

Consider the special case of the von Mises Fisher matrix distribution (A.29) with $Z = [z_1, z_2] \in \mathbb{R}^{p \times 2}$, and parameter $F = [f_1, f_2] \in \mathbb{R}^{p \times 2}$, with added constraint that f_1, f_2 are mutually orthogonal: $f_1' f_2 = 0$. Then, the marginal distribution of z_1 is [108]:

$$f(z_1|F) = \frac{{}_0F_1\left(\frac{1}{2}(p-1), \frac{1}{4}(I_p - z_1 z_1') f_2 f_2'\right) \exp(\text{tr}(f_1' z_1))}{{}_0F_1\left(\frac{1}{2}p, \frac{1}{4}F F'\right) C(p, 1)}. \quad (C.6)$$

Note that maximum likelihood estimate of z_1 (A.31), i.e.:

$$\hat{z}_1 = \arg \max_{z_1} f(Z|F) = f_1 / \sqrt{f_1' f_1},$$

This is orthogonal to f_2 , i.e. $\hat{z}_1' f_2 = 0$, [108]. Therefore, the contribution of the quadratic term in the argument of ${}_0F_1$ in the numerator of (C.6) would be negligible for values of z_1 around \hat{z}_1 . Hence, we

approximate

$${}_0F_1\left(\frac{1}{2}(p-1), \frac{1}{4}(I_p - z_1 z_1') f_2 f_2'\right) \approx {}_0F_1\left(\frac{1}{2}(p-1), \frac{1}{4} f_2 f_2'\right), \quad (C.7)$$

which will be satisfied when $f(z_1|F)$ is not diffuse, i.e. when $f_1 \rightarrow \infty$ (see Section A.5.1). Under this approximation, the leading fraction in (C.6) is independent of z_1 , and thus acts as a normalizing coefficient. Distribution (C.6) is then of the von Mises-Fisher type, namely $f(z_1|F) \approx f(z_1|f_1) = \mathcal{M}(f_1)$ (A.29). Comparing the normalizing coefficient in (C.6) with that in (A.30) yields

$${}_0F_1\left(\frac{1}{2}p, \frac{1}{4}F_2 F_2'\right) \approx {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}f_1 f_1'\right) {}_0F_1\left(\frac{1}{2}(p-1), \frac{1}{4}f_2 f_2'\right). \quad (C.8)$$

Extending (C.6) into higher dimension and using the chain rule of pdfs we obtain an approximation of the following type:

$${}_0F_1\left(\frac{1}{2}p, \frac{1}{4}L_F^2\right) \approx \prod_{i=1}^p {}_0F_1\left(\frac{1}{2}p - i + 1, \frac{1}{4}l_{F,i}^2\right). \quad (C.9)$$

Figure C.1 shows two plots. The left plot shows the function ${}_0F_1\left(\frac{1}{2}, -\frac{1}{2}x^2\right)$ for $x \in [0, 1]$. The right plot shows the function ${}_0F_1\left(\frac{1}{2}, -\frac{1}{2}x^2\right)$ for $x \in [0, 1]$. The text discusses the approximation of the von Mises-Fisher matrix distribution (A.29) with the leading fraction in (C.2) being independent of x_i , and thus being a normalizing coefficient. Comparing the normalizing coefficient in (C.6) with that in (A.30) yields

$$(C.3) \quad \frac{d}{ds} \ln {}_0F_1\left(\frac{1}{2}, -\frac{1}{2}x^2\right) = \frac{d}{ds} \ln \left[\frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2} + s\right)} \left(1 - \frac{1}{2}x^2\right)^{-s} \right]$$

Exceeding (C.3) into higher dimension and using the chain rule of parts we obtain an approximation of the following type:

$$(C.4) \quad \frac{d}{ds} \ln {}_0F_1\left(\frac{1}{2}, -\frac{1}{2}x^2\right) = \frac{d}{ds} \ln \left[\frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2} + s\right)} \left(1 - \frac{1}{2}x^2\right)^{-s} \right] = \frac{d}{ds} \ln \left[\frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2} + s\right)} \right] - s \frac{d}{ds} \ln \left(1 - \frac{1}{2}x^2\right)$$

Furthermore, (C.3) can be expressed in terms of (C.2) and, therefore (C.4). The evaluation of expansion (C.4) converges very fast for $s < p$. However, when $s \gg p$ (say $s > 10p$) the convergence is quite slow. For large s , a more numerically efficient approximation is obtained via a Taylor expansion of (C.2) at $s \rightarrow \infty$:

$$(C.5) \quad \frac{d}{ds} \ln {}_0F_1\left(\frac{1}{2}, -\frac{1}{2}x^2\right) = 1 - \left(\frac{p-1}{2s}\right) \exp\left(-\frac{p-3}{4s}\right) + o(s^{-2})$$

Here, $o(s^{-2})$ denotes elements of the series in terms of s^{-3} . This expansion provides an excellent approximation in the case $s \gg p$.

C.2. Approximation of ${}_0F_1$ of Matrix Argument by ${}_0F_1$ of Scalar Arguments

Consider the special case of the von Mises-Fisher matrix distribution (A.29) with $Z = (z_1, z_2) \in \mathbb{R}^{2 \times 2}$ and parameter $F = (f_1, f_2) \in \mathbb{R}^{2 \times 2}$, with added constraint that f_1, f_2 are mutually orthogonal; $f_1^T f_2 = 0$. Then, the marginal distribution of z_1 is [108]:

$$(C.6) \quad f(z_1|F) = \frac{{}_0F_1\left(\frac{1}{2}, -\frac{1}{2}(z_1 - z_1^*)^T f_1 f_1^T (z_1 - z_1^*)\right)}{{}_0F_1\left(\frac{1}{2}, -\frac{1}{2}F^T F\right) C(p, 1)} \exp\{tr(f_1^T z_1)\}$$

Note that maximum likelihood estimate of z_1 (A.31), i.e.:

$$\hat{z}_1 = \arg \max_{z_1} f(z_1|F) = f_1 / \sqrt{f_1^T f_1}$$

This is orthogonal to f_2 , i.e. $\hat{z}_1^T f_2 = 0$ [108]. Therefore, the contribution of the quadratic term in the argument of ${}_0F_1$ in the numerator of (C.6) would be negligible for values of z_1 around \hat{z}_1 . Hence, we

Bibliography

- [1] S. M. Kay, *Fundamentals Of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.
- [2] P. Gebouský, M. Kárný, and A. Quinn, “Lymphoscintigraphy of Upper Limbs: A Bayesian Framework,” in *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, and J. O. Berger, eds.), pp. 543–552, Oxford: University Press, 2003.
- [3] J. Heřmanská, M. Kárný, J. Zimák, L. Jirsa, and M. Šámal, “Improved prediction of therapeutic absorbed doses of radioiodine in the treatment of thyroid carcinoma,” *The Journal of Nuclear Medicine*, vol. 42, no. 7, pp. 1084–1090, 2001.
- [4] Z. Ghahramani and M. J. Beal, “Variational inference for bayesian mixtures of factor analyzers,” *Neural Information Processing Systems*, vol. 12, pp. 449–455, 2000.
- [5] G. Parisi, *Statistical Field Theory*. Reading Massachusetts: Addison Wesley, 1988.
- [6] L. K. Saul, T. S. Jaakkola, and M. I. Jordan, “Mean field theory for sigmoid belief networks.,” *Journal of Artificial Intelligence Research*, vol. 4, pp. 61–76, 1996.
- [7] C. M. Bishop, “Variational principal components,” in *Proceedings of the Ninth International Conference on Artificial Neural Networks*, (ICANN), 1999.
- [8] H. Attias, “A Variational Bayesian framework for graphical models.,” in *Advances in Neural Information Processing Systems* (T. Leen, ed.), vol. 12, MIT Press, 2000.
- [9] M. Opper and D. Saad, *Advanced Mean Field Methods: Theory and Practice*. Cambridge, Massachusetts: The MIT Press, 2001.
- [10] J. Bernardo and A. Smith, *Bayesian Theory*. Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons, 1997. 2nd edition.
- [11] B. de Finetti, *Theory of Probability: A Critical Introductory Treatment*. New York: J. Wiley, 1970.
- [12] M. Kárný and R. Kulhavý, “Structure determination of regression-type models for adaptive prediction and control,” in *Bayesian Analysis of Time Series and Dynamic Models* (J. Spall, ed.), New York: Marcel Dekker, 1988. Chapter 12.

- [13] A. Quinn, "Regularized signal identification using Bayesian techniques," in *Signal Analysis and Prediction*, Birkhäuser Boston Inc., 1998.
- [14] M. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill Company, 1970.
- [15] J. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1985.
- [16] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System identification* (P. Eykhoff, ed.), pp. 239–304, Oxford: Pergamon Press, 1981.
- [17] Box and Tiao, *Bayesian Statistics*. Oxford: Oxford, 1961.
- [18] H. Jeffreys, *Theory of Probability*. Oxford University Press, 3 ed., 1961.
- [19] B. Koopman, "On distributions admitting a sufficient statistic," *Transactions of American Mathematical Society*, vol. 39, p. 399, 1936.
- [20] J. Pratt, H. Raiffa, and R. Schlaifer, *Introduction to Statistical Decision Theory*. MIT Press, 1995.
- [21] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39 Series (B), pp. 1–38, 1977.
- [22] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse and other variants.,," in *Learning in Graphical Models* (M. I. Jordan, ed.), pp. 355–369, Kluwer, 1998.
- [23] C.F.J.Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, pp. 95–103, 1983.
- [24] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of American Statistical Association*, vol. 90, pp. 773–795, 1995.
- [25] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixtures*. New York: John Wiley & Sons, 1985.
- [26] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.
- [27] J. M. Bernardo, "Expected information as expected utility," *The Annals of Statistics*, vol. 7, no. 3, pp. 686–690, 1979.
- [28] S. Amari, S. Ikeda, and H. Shimokawa, "Information geometry of α -projection in mean field approximation," in *Advanced Mean Field Methods* (M. Opper and D. Saad, eds.), (Cambridge, Massachusetts), The MIT Press, 2001.
- [29] R. Kulhavý, "Recursive nonlinear estimation: A geometric approach," *Automatica*, vol. 26, no. 3, pp. 545–555, 1990.
- [30] A. Rakar, P. N. Tatiana V. Guy, M. Kárný, and D. Juričić, "Advisory system productool: Case study on gas conditioning unit," *Journal of Process Control*, 2003. accepted.

- [31] A. Quinn, P. Ettler, L. Jirsa, I. Nagy, and P. Nedoma, "Probabilistic advisory systems for data-intensive applications," *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 2, pp. 133–148, 2003.
- [32] D. J. C. MacKay, "Probable networks and plausible predictions — a review of practical bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, pp. 469–505, 1995.
- [33] J. W. Miskin, *Ensemble Learning for Independent Component Analysis*. PhD thesis, University of Cambridge, 2000.
- [34] M. Opper and O. Winther, "From naive mean field theory to the TAP equations," in *Advanced Mean Field Methods* (M. Opper and D. Saad, eds.), (Cambridge, Massachusetts), The MIT Press, 2001.
- [35] M. Sato, "Online model selection based on the variational bayes," *Neural Computation*, vol. 13, pp. 1649–1681, 2001.
- [36] S. Amari, *Differential-Geometrical Methods in Statistics*. Springer-Verlag, 1985.
- [37] M. J. Beal and Z. Ghahramani, "The variational bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics 7* (J. M. et. al. Bernardo, ed.), Oxford University Press, 2003.
- [38] S. J. Godsill, *Digital Audio Restoration: A Statistical Model Based Approach*. Springer Verlag, 1998.
- [39] J. S. Yedidia, "An idiosyncratic journey beyond mean field theory," in *Advanced Mean Field Methods* (M. Opper and D. Saad, eds.), (Cambridge, Massachusetts), The MIT Press, 2001.
- [40] R. Kulhavý, "Recursive Bayesian estimation under memory limitations," *Kybernetika*, vol. 26, pp. 1–20, 1990.
- [41] R. Kulhavý, "A Bayes-closed approximation of recursive non-linear estimation," *International Journal Adaptive Control and Signal Processing*, vol. 4, pp. 271–285, 1990.
- [42] J. M. Bernardo, "Approximations in statistics from a decision-theoretical viewpoint," in *Probability and Bayesian Statistics* (R. Viertl, ed.), pp. 53–60, New York: Plenum, 1987.
- [43] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag, 1997.
- [44] P. M. Lee, *Bayesian Statistics, an Introduction*. Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons, 1997. 2nd edition.
- [45] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

- [46] S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2002.
- [47] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, 1971.
- [48] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 61, pp. 611–622, 1998.
- [49] D. Rubin and D. Thayer, "EM algorithms for ML factor analysis," *Psychometrica*, vol. 47, pp. 69–76, 1982.
- [50] S. J. Press and K. Shigemasu, "Bayesian inference in factor analysis," in *Contributions to Probability and Statistics* (L. J. Glesser, ed.), ch. 15, Springer Verlag, New York, 1989.
- [51] P. Comon, "Independent component analysis: A new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.
- [52] H. Attias, "EM algorithms for independent component analysis," in *Neural Networks for Signal Processing* (M. Niranjan, ed.), pp. 132–141, IEEE, 1998.
- [53] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, pp. 803–851, 1999.
- [54] M. West, P. J. Harrison, and H. S. Migon, "Dynamic generalized linear models and Bayesian forecasting," *Journal of the American Statistical Association*, vol. 80, no. 389, 1985.
- [55] M. West, "Robust sequential approximate bayesian estimation," *Journal of the Royal Statistical Society. Series B*, vol. 43, no. 2, pp. 157–166, 1981.
- [56] C. Cargnoni, P. Muller, and M. West, "Bayesian forecasting of multinomial series through conditional gaussian dynamic models," *Journal of American Statistical Association*, vol. 92, no. 438, pp. 640–647, 1997.
- [57] S. J. Godsill, A. Doucet, and M. West, "Maximum a posteriori sequence estimation using monte carlo particle filters," *Annals of the Institute of Statistical Mathematics*, vol. 53, pp. 82–96, 2001.
- [58] L. Ljung and T. Söderström, *Theory and practice of recursive identification*. Cambridge; London: MIT Press, 1983.
- [59] L. Ljung, *System Identification-Theory for the User*. Prentice-hall. Englewood Cliffs, N.J: D. van Nostrand Company Inc., 1987.
- [60] Z. Ghahramani and M. Beal, "Graphical models and variational methods," in *Advanced Mean Field Methods* (M. Opper and D. Saad, eds.), (Cambridge, Massachusetts), The MIT Press, 2001.
- [61] V. Peterka, "Real-time parameter estimation and output prediction for ARMA type system models," *Kybernetika*, vol. 17, no. 6, pp. 526–533, 1981.

- [62] L. He and M. Kárný, "Estimation and prediction with ARMMAX model: a mixture of ARMAX models with common ARX part," *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 4, pp. 265–283, 2003.
- [63] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," tech. rep., Aston University, 1997.
- [64] D. B. Rowe and S. J. Press, "Gibbs sampling and hill climbing in bayesian factor analysis," tech. rep., University of California, Riverside, 1998.
- [65] B. Porat, *Digital processing of random signals: theory and methods*. Englewood Cliffs, N.J.: Prentice-Hall, 1994.
- [66] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.
- [67] S. Kay, *Modern spectral estimation*. New Jersey: Prentice-Hall, 1988.
- [68] P. Wellstead and M. Zarrop, *Self-tuning Systems*. Chichester: John Wiley & Sons, 1991.
- [69] J. Rajan, P. Rayner, and S. Godsill, "Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler," *Vision, Image and Signal Processing, IEE Proceedings*, vol. 144, no. 4, pp. 249–256, 1997.
- [70] B. Widrow and S. Stearns, *Adaptive Signal Processing*. Prentice-Hall, 1985.
- [71] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover Publications, Inc., 1972.
- [72] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems*, pp. 758–764, 2001.
- [73] G. Golub and C. VanLoan, *Matrix Computations*. Baltimore, London: The John Hopkins University Press, 1989.
- [74] G. Bierman, *Factorization Methods for Discrete Sequential Estimation*. New York: Academic Press, 1977.
- [75] L. Berc, *Model Structure Identification: Global and Local Views. Bayesian Solution*. PhD thesis, Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Prague, Czech Republic, 1998.
- [76] R. H. Middleton, G. C. Goodwin, D. J. Hill, and D. Q. Mayne, "Design issues in adaptive control," *IEEE Transactions on Automatic Control*, vol. 33, no. 1, pp. 50–58, 1988.
- [77] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic Press, 1979.
- [78] G. V. Moustakides, "Locally optimum adaptive signal processing algorithms," *IEEE Transactions on Signal Processing*, vol. 46, no. 12, pp. 3315–3325, 1998.

- [79] R. Kulhavý and M.B.Zarrop, "On general concept of forgetting," *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.
- [80] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 2nd ed., 2002.
- [81] I. Buvat, H. Benali, and R. Di Paola, "Statistical distribution of factors and factor images in factor analysis of medical image sequences," *Physics in Medicine and Biology*, vol. 43, no. 6, pp. 1695–1711, 1998.
- [82] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, vol. 2, pp. 559–572, 1901.
- [83] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.
- [84] T. P. Minka, "Automatic choice of dimensionality for PCA," tech. rep., MIT, 2000.
- [85] T. W. Anderson, "Estimating linear statistical relationships," *Annals of Statistics*, vol. 12, pp. 1–45, 1984.
- [86] A. P. J. Fine, "Asymptotic study of the multivariate functional model. application to the metric of choice in principal component analysis," *Statistics*, vol. 23, pp. 63–83, 1992.
- [87] F. Pedersen, M. Bergstroem, E. Bengtsson, and B. Langstroem, "Principal component analysis of dynamic positron emission tomography studies," *European Journal of Nuclear Medicine*, vol. 21, pp. 1285–1292, 1994.
- [88] F. Hermansen and A. A. Lammertsma, "Linear dimension reduction of sequences of medical images: I. optimal inner products," *Physics in Medicine and Biology*, vol. 40, pp. 1909–1920, 1995.
- [89] H. Benali, I. Buvat, F. Frouin, J. P. Bazin, and R. Di Paola, "A statistical model for the determination of the optimal metric in factor analysis of medical image sequences (FAMIS)," *Physics in Medicine and Biology*, vol. 38, no. 8, pp. 1065–1080, 1993.
- [90] M. Šámal, M. Kárný, H. Benali, W. Backfrieder, A. Todd-Pokropek, and H. Bergmann, "Experimental comparison of data transformation procedures for analysis of principal components," *Physics in Medicine and Biology*, vol. 44, pp. 2821–2834, 1999.
- [91] M. Šámal, M. Kárný, H. Šurová, E. Maříková, and Z. Dienstbier, "Rotation to simple structure in factor analysis of dynamic radionuclide studies," *Physics in Medicine and Biology*, vol. 32, no. 3, pp. 371–382, 1987.
- [92] M. Kárný, M. Šámal, and J. Böhm, "Rotation to physiological factors revised," *Kybernetika*, vol. 34, no. 2, pp. 171–179, 1998.
- [93] G. D. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

- [94] X. R. Li and Y. Bar-Shalom, "Multiple-model estimation with variable structure," *IEEE Transactions on Automatic Control*, vol. 41, no. 4, pp. 478–493, 1996.
- [95] D. G. Lainiotis, "Partitioning: A unifying framework for adaptive systems, I: Estimation," *Proceedings of the IEEE*, vol. 64, no. 8, pp. 1126–1143, 1976.
- [96] H. A. P. Blom and Y. Bar-Shalom, "The interacting multiple model algorithm for systems with switching coefficients," *IEEE Transactions on Automatic Control*, vol. 33, pp. 780–783, 1988.
- [97] M. G. Kendall, A. Stuart, and K. Ord, *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Edward Arnold, 6th ed., 1998.
- [98] T. Söderström and R. Stoica, *System Identification*. Prentice-Hall, 1989.
- [99] M. Kárný, J. Böhm, T. V. Guy, and P. Nedoma, "Mixture-based adaptive probabilistic control," *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 2, pp. 119–132, 2003.
- [100] M. Niedźwiecki and K. Cisowski, "Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals," *IEEE Transactions on Signal Processing*, vol. 44, no. 3, 1996.
- [101] V. Katkovnik, K. Egiazarian, and J. Astola, "Application of the ICI principle to window size adaptive median filtering," *Signal Processing*, vol. 83, pp. 251–257, 2003.
- [102] C.-F. So, S. C. Ng, and S. H. Leung, "Gradient based variable forgetting factor rls algorithm," *Signal Processing*, vol. 83, pp. 1163–1175, 2003.
- [103] R. Kulhavý, "Restricted exponential forgetting in real-time identification," *Automatica*, vol. 23, no. 5, pp. 589–600, 1987.
- [104] B. Toplis and S. Pasupathy, "Tracking improvements in fast rls algorithms using a variable forgetting factor," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 2, 1988.
- [105] M. Kárný, J. Andryšek, L. Tesař, and P. Nedoma, "Mixture-based adaptive probabilistic control," *International Journal of Adaptive Control and Signal Processing*, 2003. draft of the paper.
- [106] M. Kárný, N. Khailova, P. Nedoma, and J. Böhm, "Quantification of prior information revised," *International Journal of Adaptive Control and Signal Processing*, vol. 15, no. 1, pp. 65–84, 2001.
- [107] V. Šmídl and A. Quinn, "Orthogonal variational PCA," *Journal of the American Statistical Society*, 2003, under review.
- [108] C. G. Khatri and K. V. Mardia, "The von Mises-Fisher distribution in orientation statistics," *Journal of Royal Statistical Society B*, vol. 39, pp. 95–106, 1977.
- [109] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," tech. rep., Aston University, 1998.

- [110] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.
- [111] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.
- [112] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, pp. 626–634, 1999.
- [113] M. v. V. Šmídl, M. Kárný, "Smoothing prior information in principal component analysis of dynamic image data," in *Preprints of the 17th International Conference on Information Processing in Medical Imaging*, vol. 2082 of *Lecture Notes in Computer Science*, (Davis CA), June 2001.
- [114] J. Andryšek, "Projection Based Estimation of Dynamic Probabilistic Mixtures," Tech. Rep. 2098, ÚTIA AV ČR, Praha, 2004.
- [115] S. Kotz and N. Johnson, *Encyclopedia of statistical sciences*. New York: John Wiley, 1985.
- [116] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
- [117] A. T. James, "Distribution of matrix variates and latent roots derived from normal samples," *Annals of Mathematical Statistics*, vol. 35, pp. 475–501, 1964.
- [118] K. Mardia and P. E. Jupp, *Directional Statistics*. Chichester, England: John Wiley and Sons, 2000.
- [119] T. D. Downs, "Orientational statistics," *Biometrika*, vol. 59, pp. 665–676, 1972.
- [120] R. W. Butler and T. A. Wood, "Laplace approximation of Bessel function of matrix argument," *Journal of Computational and Applied Mathematics*, vol. 155, pp. 359–382, 2003.