



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.

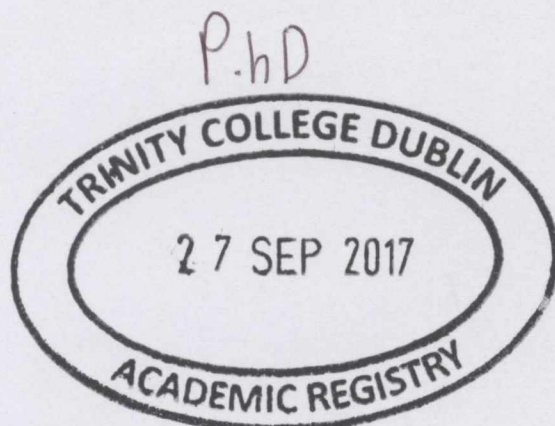
Distributional Design For Bayesian Knowledge Processing In Multiple Participant Systems

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

Department of Electronic and Electrical Engineering,
Trinity College Dublin

July 2017

Shahram Azizi Sazi



School of Engineering

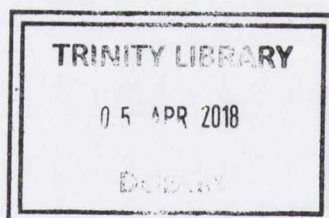
THESIS

11555

50763057



I dedicate this work to my wife, Mina, my daughter, Anita, and my little son, Mohammad Sadra.



Thesis 11555

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other University and it is entirely my own work.

I agree to deposit this thesis in the University’s open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

The copyright belongs jointly to the University of Dublin and Shahram Azizi Sazi.



i Sazi

2017

Acknowledgments

First and foremost, I am very grateful to my supervisor, Dr. Anthony Quinn. Throughout my Ph.D. programme, he has been constantly encouraging, and extremely helpful in indicating which ideas deserved further consideration and which should be forgotten about. Anthony always provided me with quick and insightful comments throughout my research and he generously funded me to travel to conferences, allowing me to learn, and to meet other researchers in this field. Anthony always understood and helped me in other aspects of life and provided me with whatever was possible for him. I could not have asked for a better supervisor than him.

I am also thankful to Sadegh Chegini and Masoud Manafi who provided me with support at different stages of my research.

Finally, I would like to thank my family for their love and support over the years. The financial support of SFI grant 10/CE/11855 (Lero), is gratefully acknowledged.

Abstract

Knowledge processing in an optimal way is of great importance in decision making. In this thesis, we propose a consistent Bayesian mechanism for processing knowledge via the design of the probability distribution. Bayes' rule is exploited as a consistent mechanism when a complete model is available. In this thesis, we examine an important context for this, namely conjugate Bayesian inference for the autoregressive (AR) model. We characterise this as an example of uniquely specified distributional design, and we provide some novel results related to the computation of the associated statistics within fully tractable signal processing algorithms.

In contrast, fully probabilistic design (FPD) is motivated in this thesis for cases where a model of the knowledge is not available. Much of this thesis is concerned with constructing novel FPD solutions of distributional design through processing stochastic knowledge in multiple participant systems (MPS). We propose a centralized deliberator-based scheme where all participants independently report their beliefs (distributions) about some common quantity of interest to a parametric deliberator. The latter processes this knowledge into an optimal distribution via FPD, as it is fully explained in the thesis. Four FPD solutions are derived in Chapters 4-5, addressing various different knowledge specification and ways of processing them. A thorough comparison of these solutions is provided with extensive simulation results. In an extended application in Chapter 6, a selection of these solutions is derived and simulated for processing knowledge in the context of multiple AR nodes.

In Chapter 7, the Bayesian filtering (BF) problem is studied in the case where time-variant states are incompletely modelled. Once again, requiring an optimal distributional design in place of FPD related techniques, stabilized forgetting and logarithmic pooling are derived and simulated. We extend the solution to the distributed knowledge processing for multiple BF nodes. In Chapter 8, we conclude the thesis by summarizing core results of each Chapter, and provide numerous suggestions for further research in this field.

Table of Contents

Acknowledgments	iv
List of Figures	x
List of Tables	xiv
Notational conventions	xv
Acronyms	xvii
Summary	xx
Chapter 1	
The purpose and motivation	1
1.1 Probabilistic model	1
1.1.1 Parametric probabilistic model	3
1.2 Aims of research and proposed methodology	3
1.2.1 Bayesian data processing	6
1.2.2 Stochastic knowledge processing	7
1.2.3 Stochastic knowledge processing in multiple distributed nodes: objectives of thesis	8
1.3 Layout of the thesis	9
1.4 The novel contributions of the thesis	11
Chapter 2	
Distributional decision making: design and approximation	13
2.1 The fundamental decision making (DM)	13
2.2 Loss function for distributional decision making	14
2.2.1 Kullback-Leibler divergence (KLD) and its properties	15
2.3 Distributional design	18
2.3.1 Distributional parameter prior design	22
2.3.1.1 Parameter prior classes and designs	23
2.4 Distributional approximation	24
2.4.1 Distributional stochastic approximation	24
2.4.2 Distributional deterministic approximation	25
2.5 Properties of fully probabilistic design (FPD)	26
2.6 Discussion	27

Chapter 3

Bayesian statistical data processing	29
3.1 Linear regression models	29
3.2 AR model	30
3.3 Time-invariant parametrization of AR model: known model order . . .	31
3.3.1 The p -conditional parameter inference of AR model	32
3.3.2 p -conditional one-step-ahead predictor	36
3.3.3 p -conditional multi-step-ahead predictor	37
3.4 Time-invariant parametrization of the AR model: unknown model order	38
3.5 Classical data processing using the AR model: known model order . . .	41
3.5.1 Technical assumptions for classical strategies	43
3.5.2 p -conditional parameter estimation strategies	44
3.5.3 p -conditional prediction	47
3.5.4 Classical data processing using the AR model: unknown model order	48
3.5.5 Model order estimation strategies	48
3.5.5.1 Principles of model comparison and specialization into AR case	48
3.5.6 Parameter estimation and prediction	54
3.6 Computational load of extended information matrices (EIMs)	54
3.6.1 Computational Load of EIMs: off-line scenario	55
3.6.1.1 Incremental way	56
3.6.1.2 Decremental way	56
3.6.2 Performance in simulation	56
3.6.3 Computational load of EIMs: on-line scenario	58
3.6.3.1 Decremental way	59
3.6.4 Comparison of computational load of EIMs between two different scenarios	60
3.7 Comparative performance evaluation under simulation	61
3.7.1 Comparison between classical approaches and Bayesian method in model order selection	61
3.7.1.1 Comparison between classical approaches and Bayesian method for the estimation of the subset of parameters	63
3.7.1.2 Comparison between classical prediction strategies and Bayesian prediction	67
3.8 Discussion	68

Chapter 4

Multiple participant processing: the moral deliberator	71
4.1 DM schemes for merging knowledge in multiple participant system (MPS)	72
4.2 Distributional design in MPS	74
4.2.1 Distributional pools in MPS	75
4.2.1.1 Weights assignment	76
4.2.1.2 Limitations of the distributional pools	78
4.2.2 FPD for distributional design in MPS	79

4.3.2	Functionally tractable FPD: exponential family	87
4.3.2.1	An illustrative example: the normal case	89
4.4	Comparative Performance Evaluation under Simulation	90
4.4.1	The hierarchical property of the FPD-optimal design (4.15)	91
4.4.2	Optimal design of the Kuhn-Tucker multipliers, λ	92
4.5	Discussion	93

Chapter 5

Multiple participant processing: external predictors		97
5.1	Knowledge definition and modelling	98
5.2	FPD-optimal merging with augmentation (FPD-A)	100
5.2.1	FPD-A of the deliberator's predictor	102
5.3	FPD-sequential merging (FPD-S)	104
5.4	FPD-optimal merging with restriction (FPD-R)	107
5.5	Relationships among FPD-A, FPD-S and FPD-R	111
5.6	An illustrative example: normal case	112
5.6.1	FPD-A: normal case	113
5.6.2	FPD-R: normal case	114
5.7	Comparative performance evaluation under simulation	116
5.8	Discussion	118

Chapter 6

Processing multiple autoregressive nodes		121
6.1	FPD solutions for AR deliberator: known model order	122
6.1.1	FPD-optimal solution of deliberator's AR parameter prior: Application FPD-A (5.16) and FPD-R (5.34)	123
6.2	FPD solutions for deliberator's AR parameter prior: unknown participants' model order	126
6.3	FPD-optimal solutions of the AR deliberator: unknown model order	127
6.4	Comparative performance evaluation under simulation	129
6.4.1	Simulation results for $\mathcal{I}_0$'s predictors with known model order $\mathbf{p}$. . .	129
6.4.2	Simulation results for $\mathcal{I}_0$'s predictors with unknown model order $\mathbf{p}$. .	132
6.5	discussion	132

Chapter 7

BF using logarithmic pooling design		134
7.1	Bayesian filtering (BF) under first order Markov property	135
7.2	Exact BF	139
7.2.1	Kalman Filter (KF)	140
7.2.2	Mixture Kalman Filter (MKF)	142
7.3	A review of Bayesian methods beyond KF	144
7.4	Merging state predictors for designing the time step of BF	145
7.4.1	SF-based design for merging alternative pd	145
7.5	Tractable SF for Exponential Family	146
7.5.1	SF solution in exponential family	147

7.4.1	SF-based design for merging alternative pd	145
7.5	Tractable SF for Exponential Family	146
7.5.1	SF solution in exponential family	147
7.5.2	SF solution for mixture components of exponential family . . .	148
7.6	Illustrative examples and comparative performance under simulation . .	150
7.6.1	Simulation study: evaluation of the example 7.6.1	151
7.6.2	Simulation study for evaluation of example 7.6.2	153
7.7	Optimal tuning of the forgetting factor (FF), λ_n	153
7.7.1	Application: SF-based design for BF in a multiple signal-dependent noise problem	155
7.7.2	Simulation study: multiple single-dependent noise problem . . .	156
7.8	The logarithmic pooling design for incomplete models in MP	159
7.9	Discussion	160

Chapter 8

Conclusions and Future Work		162
8.1	Conclusion of Chapter 3	162
8.1.1	Possible extensions and future directions of Chapter 3	162
8.2	Conclusion for Chapter 4	163
8.2.1	Possible extensions and future directions of Chapter 4	163
8.3	Conclusion for Chapter 5	164
8.3.1	Possible extensions and future directions of Chapter 5	165
8.4	Conclusion for Chapter 6	166
8.4.1	Possible extensions and future directions of Chapter 6	167
8.5	Conclusion for Chapter 7	167
8.5.1	Possible extensions and future directions of Chapter 7	167
8.6	Conclusion	168

Appendix A

Probability Distributions		172
A.1	Exponential Family	172
A.2	Conjugate Exponential Family (CEF)	173
A.3	Non-Exponential Family	173
A.4	The proof of p -conditional OSA predictor (3.34)	175
A.5	Proof of Lemma 3.6.1	176
A.6	Proof of Lemma 3.6.2	177
A.7	The proof of the time step (7.27) and data step (7.30) of KF	177

Bibliography	180
---------------------	------------

List of Figures

1.1	The schematic diagram for the taxonomy of knowledge, K , the model of K and the proposed methodologies	4
2.1	The schematic diagram for the taxonomy of distributional DM	15
2.2	The schematic diagram for processing K via FPD (2.14).	18
2.3	The schematic diagram for KLD-based optimal distributional designs (2.14), (2.24) and (2.25).	28
2.4	The schematic diagram for KLD-based distributional deterministic approximation (2.36) and (2.37).	28
3.1	Comparison of different ways of computation EIMs using recursion in order (3.108), $n = 1000$	57
3.2	Performance of different ways in accuracy of model order inference using recursion in order (3.108) with $N = 6 \times 10^7$. Data are simulated from $AR(4)$ (LHS) and $AR(8)$ (RHS).	57
3.3	Schematic diagram for computation of the EIM using full recursion (3.112) for decremental way in on-line case	60
3.4	Comparison of running time of decremental way in off-line case versus on-line case, $n = 1000$	61
3.5	Comparison of different criteria of model order selection in terms of p_u with $AR(4)$, $n = 1000$	62
3.6	Comparison of different criteria of model order selection in terms of true model order with $n = 1000$, $p_u = 20$	63
3.7	Comparison of different criteria of model order selection in terms of true model order with $n = 10000$, $p_u = 20$	64
3.8	Comparison of different criteria of model order selection in terms of number of observations designed with $AR(4)$ (LHS), $AR(16)$ (RHS) and $p_u = 20$	64
3.9	Comparison of different criteria for estimation of an arbitrary subset in terms of the number of available observations n , with $ \delta = 3$, and simulated from $AR(10)$ (in the above) and $AR(14)$ (in the below).	65
3.10	Comparison of different criteria for estimation of an arbitrary subset in terms of number of observations, n , with $ \delta = 3$, $p_u = 10$ and simulated from $AR(4)$	66
3.11	Comparison of different criteria in one-step-ahead prediction in terms of true model order (known) with $n = 500$	66
3.12	Comparison of different criteria in OSA prediction in terms of time simulated from $AR(6)$ and $r = 1$	67

3.12	Comparison of different criteria in OSA prediction in terms of time simulated from $AR(6)$ and $r = 1$	67
3.13	Comparison of different criteria in OSA prediction in terms of data, n , with $p_u = 20$, and simulated from $AR(6)$ and $r = 4$	68
3.14	Comparison of different criteria in MSA prediction with $n = 100$ and simulated from $AR(8)$ and $r = 2$	69
3.15	Comparison of different criteria in 6-step-ahead prediction with $r = 2, n = 100$	69
4.1	Schematic diagram for different structures; Left-hand side is the structure of a centralized deliberator-based scheme and Right-hand side is the structure of a decentralized scheme.	73
4.2	Schematic diagram for the inverse of an arithmetic mean of KLDs.	77
4.3	Schematic diagram for the inverse of of maximum KLDs.	78
4.4	Schematic diagram for processing K (Definition 4.3.1) into the FPD-optimal design of the moral deliberator (4.15), then deducing their optimal predictor. Here, $\int_{\Theta}$ denotes the marginalization operator.	85
4.5	$A^o(a K)$ for $s = 20$ with $MC = 1000$ runs. Normal case with $r = 1$	91
4.6	$\mathcal{I}_0$'s predictor, $G_0^o(x K)$ (4.40), compared with the participants' pds, $G_l(x)$, for $s = 5$, with shaping parameters averaged across $MC = 100$ runs in all cases. Normal case with $r = 1$	91
4.7	Comparison of $\mathcal{I}_0$'s predictor, $G_0^o(x K)$ (4.40), and the logarithmic pool, $G_0^g(x)$ (4.4), in terms of the participants' expected loss (i.e. $D[G_l G_0^o]$ and $D[G_l G_0^g]$). The $s = 20$ participants are in the order of increasing β_l . Normal case with $r = 1$	92
4.8	Comparison of $\mathcal{I}_0$'s $G_0^o(x \lambda, K)$ (4.40) and $G_0^g(x)$ (4.4) in terms of weighted KLD (4.41), as a function of the number of participants, s . Normal case with $r = 1$ and $w_l = \frac{1}{s}$	93
4.9	Comparison of $\mathcal{I}_0$'s $G_0^o(x \lambda, K)$ (4.40) and $G_0^g(x)$ (4.4) in terms of weighted KLD, as a function of the number of participants, s . Normal case with $r = 1$ and $w_l = \frac{\frac{1}{\beta_l}}{\sum_{j=1}^s \frac{1}{\beta_j}}$	94
5.1	Schematic diagram for processing K (Definition 5.1.1) into the FPD-A (5.16), then deducing their optimal predictor. Here, $\int_{\Theta}$ denotes the marginalization operator.	105
5.2	Schematic diagram for sequential processing of K (Definition 5.1.1) into the FPD-sequential merging (5.22), then deducing their sequential predictor. Here, $\int_{\Theta}$ denotes the marginalization operator.	107
5.3	Schematic diagram for processing K (Definition 5.1.1) into the FPD-R (5.34), then deducing their optimal predictor. Here, $\int_{\Theta}$ denotes the marginalization operator.	111
5.4	Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (5.53), lin pool-FPD weights (5.54), FPD-R (5.61), FPD-moral deliberator (5.62), FPD-S (5.64) against $\mathcal{I}_0$'s μ for $r = 1, n = 500, s = 2$	117

5.6	Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (5.53), lin pool-FPD weights (5.54), FPD-R (5.61), FPD-moral deliberator (5.62), FPD-S (5.64) against $\mathcal{I}_0$'s n for $\mu = 4, r = 1, s = 2$	118
5.7	Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (5.53), lin pool-FPD weights (5.54), FPD-R (5.61), FPD-moral deliberator (5.62), FPD-S (5.64) against $\mathcal{I}_0$'s ν_I for $\mu = 4, r = 1, n = 500$	118
6.1	Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.17), lin pool-FPD weights (6.21), and FPD-R (6.22) for $p_l = 8, s = 10$. (a) $n = 10$, (b) $n = 50$, (c) $n = 500$	130
6.2	In LHS, comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.17), lin pool-FPD weights (6.21), and FPD-R (6.22) for $p_l = 5, l = 1, \dots, s$ and $n = 100$ in terms of s . In RHS, MSE for FPD-A (6.17) against lin pool-FPD weights (6.21), with the same operating factor.	130
6.3	Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.17), lin pool-FPD weights (6.21), and FPD-R (6.22) for $p_l = 5, p_2 = 10, s = 2$ and $n = 100$ in terms of $\mathcal{I}_0$'s variance, r	131
6.4	Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.33), lin pool-FPD weights (6.34), and FPD-R (6.35) for $p_l = 5, s = 10, n = 100$ and $r = 1$ and unknown order p , in terms of p_u	131
6.5	comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.33), lin pool-FPD weights (6.34), and FPD-R (6.35) for $p_l = 5, s = 10$, and $r = 1$ with unknown order p , in terms of n	132
7.1	Graphical model of unknowns under assumptions (7.7) - (7.10)	137
7.2	Graphical model for MKF	142
7.3	State reconstruction using SF versus KF in normal case $r_x = 1, r_\psi = 1$	152
7.4	State reconstruction using MKF vs MSF (normal case).	152
7.5	State reconstruction using SF for the exponential observation model in the slowly-varying state case.	154
7.6	State reconstruction using SF for the exponential observation model in the fast-varying state case.	154
7.7	Comparison of scaled mean squared error for fixed choices of λ versus SF based on VB design: slowly-varying case.	156
7.8	SF-based reconstruction of the state using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.8$: slow-varying case.	157
7.9	Comparison of scaled mean squared error for fixed choices of λ versus SF based on VB design: fast-varying case.	157
7.10	SF-based reconstruction of the state using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.2$: fast-varying case.	158
7.11	SF-based reconstruction of u_n using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.8$: slow-varying case.	158
7.12	SF-based reconstruction of u_n using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.2$: fast-varying case.	159

7.12	SF-based reconstructon of u_n using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.2$: fast-varying case.	159
7.13	Schematic boundary of different filters (solutions) denoting with A-D. For each filter, the associated class of observation models is expressed in [observation model].	160

List of Tables

8.1 A summary of proposed FPD solutions in the thesis with their properties . . 169

Notational conventions

General notational convention	
X	the random variable
x	the realization of a random variable
θ	the unknown parameter
F, G	distributions with known and specified parametric forms (i.e. fixed distributions)
A, M, P	distributions whose parametric form is unspecified (i.e. variational distributions)
$E_{F(x)}[g(X)]$	the expectation of $g(X)$ with respect to $F(x)$
$\mathbb{X}, \mathbb{F}, \mathbb{A}$, etc.	The set of possible values of the quantity shown with Latin letter
Θ	The set of possible values of the quantity shown with Greek letter
$ \mathbb{X} $	the cardinality of the set
$\emptyset$	the empty set
$\mathbf{x}$	multivariate quantity
$\text{supp}(A)$	The support of A
The superscript, o e.g. $A^o(x)$	the FPD-optimal objects (i.e. distributions, parameters, etc.)
The subscript $_I$, e.g. $A_I(x)$	the ideal specifications (i.e. distributions, parameters, etc.)
The subscript $_P$, e.g. $A_P(x)$	the deterministic choice of prior distribution on x in the <i>minimum cross entropy</i> (MXE) context
The superscript, $\tilde{}$ e.g. $\tilde{A}(x)$	the optimal distributional approximation
The superscript, $\check{}$ e.g. $\check{A}(x)$	the approximating distribution
The upper-case math calligraphy $\mathcal{I}$	the decision maker, Bayesian modeller/observer, data or knowledge processor
A upper-case symbol K	the knowledge
$\delta(x)$	delta-type function. It is determined by the type of the argument, x . For a continuous x , then $\delta(x)$ is the Dirac's delta function. For the discrete x , then $\delta(x)$ is the Kronecker function.

More detailed symbols and decorations will be used throughout the thesis which are listed as follows.

Specific notational convention in chapter 3	
p	the model order of AR
i	the running variable for time
j, l	the running variables for order
p_u	the maximum possible order
n	the number of observations in data-batch (realizations)
$\mathbf{x}_n$	the data-batch with n samples/realizations
$\mathbf{a}^{(p)} = [a_1, a_2, \dots, a_p]^T$	the unknown vector of AR coefficients with the model order p
r	unknown innovation variance
$\psi_i^{(p)} = [x_{i-1}, \dots, x_{i-p}]^T, i = p + 1, \dots, n,$	the regressor vectors with the model order p
$g_{n,j} = [x_{p-j-1}, \dots, x_{n-j}]^T, j = 1, \dots, p$	the basis functions of AR matrix with the model order p
$\mathbf{\Psi}_n^{(p)} = [g_{n,1}, \dots, g_{n,p}]^T$	the AR matrix with the model order p .
$= [\psi_{p+1}^{(p)} \dots, \psi_n^{(p)}]$	
$\varphi_i^{(p)T} = [x_i, \psi_i^{(p)T}]$	the extended regressors with the model order p
$\mathbf{\Phi}_n^{(p)} = [\mathbf{x}_n^{(p)}, \mathbf{\Psi}_n^{(p)}],$	the extended AR matrix with the model order p
$\mathbf{s}_n^{(p)}$	sufficient statistics
$\nu_n^{(p)}$	the counter of outer products
$\mathbf{V}_n^{(p)}$	the data accumulator
$\Gamma(\cdot)$	the Gamma function
$I_{n \times m} \in \mathbb{R}^{n \times n}$	square identity matrix
$\mathbf{1}_{m \times n}, \mathbf{0}_{m \times n}$	matrix of size $m \times n$ with all elements equal to one, zero, respectively
$tr(\mathbf{V})$	trace of matrix $\mathbf{V}$
Specific notational convention in chapter 4 and 5	
$\mathcal{I}_0$	the parametric Bayesian deliberator
$\mathcal{I}_l, l = 1, \dots, s$	isolated participants
s	the number of isolated participants
Subscript l	the label assigned to the ordered/ranked participants.
$\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_s]^T$	The Kuhn-Tucker multipliers which are in the role of associated weights to participants in chapter 4
$\mathbf{w} = [w_1, \dots, w_s]^T$	the associated weights to participants in chapter 5

Acronyms

AR	AutoRegressive (model)
AIC	Akaike's Information Criterion
BF	Bayesian Filtering
BIC	Bayesian Information Criterion
BPO	Bayesian Posterior Order
CE	Certainty Equivalent
CEF	Conjugate Exponential Family
DEF	Dynamic Exponential Family
DM	Decision Making
EF	Exponential Family
EIM	Extended Information Matrix
FPD	Fully Probabilistic Design
FPD-A	Fully Probabilistic Design with Augmentation
FPD-R	Fully Probabilistic Design with Restriction
FPD-S	Fully Probabilistic Design Sequential
HQ	Hannan and Quinn
KF	Kalman Filter
KLD	Kullback Leibler Divergence
KT	Kuhn Tucker
LDL	Lower Diagonal Lower decomposition
LS	Least Square
MDL	Minimum Description Length
ME	Maximum Entropy
MKF	Mixture Kalman Filter
ML	Maximum Likelihood
MML	Minimum Message Length
MPS	Multiple Participant Systems
MR	Minimum Risk
MS	Minimum Square
MSA	Multi Step Ahead
MSE	Mean Squared Error

MSF	Mixture of Stabilized Forgetting
MXE	Minimum Cross Entropy
SF	Stabilized Forgetting
OSA	One Step Ahead
VB	Variational Bayes
WSS	Wide Sense Stationary

Summary

This thesis is concerned with designing an appropriate distribution through processing knowledge in multiple participant systems (MPS). A consistent Bayesian mechanism is provided for processing knowledge underlying both complete and incomplete models of knowledge. In particular, we aim to derive functionally tractable and optimal distributional designs. Processing knowledge in a complete model context can be handled via Bayes' rule, yielding a unique design. However, Bayes' rule is not available for knowledge processing in an incomplete model context. Instead, we employ *fully probabilistic design* (FPD) to accomplish this task.

In **Chapter 2**, we review the basics of distributional decision making and the most common methods of distributional design. Also, some important techniques for the related problem of distributional approximation are reviewed. FPD methodology is chosen as our primary method for accomplishing the task of distributional design and its properties are explained. The main contributions of the thesis are presented in Chapters 3-7, as follows:

In **Chapter 3**, we uniquely process conventional (crisp) data into probability distributions via Bayes' rule. As an important context for this, we model dynamic data using an autoregressive (AR) model. We contribute to this area by exploring both inferential and computational tasks. In particular, Bayesian conjugate and tractable inference of model order and unknown parameters are derived. Also, Bayesian one-step-ahead and multi-step-ahead predictors are computed. Efficient recursions (in time and order) for the *extended information matrix* (EIM) are proposed, and the associated computational loads are audited.

Throughout Chapters 4-6, we employ FPD within a centralized deliberator-based scheme for processing stochastic knowledge in multiple participant systems (MPS). FPD solutions using this centralized scheme are not currently available when processing the continuous distributions, and so the key aim of the thesis is to derive one.

In **Chapter 4**, a centralized deliberator-based scheme is adopted to accomplish distributional design using FPD by merging multiple stochastic participants. This scheme is parametric and

moral; the knowledge-constrained set involves bounding the expectation of the expected loss imposed on each participant. The proposed FPD-optimal solution is functionally tractable for a wide class of continuous and discrete parametric deliberators and participants. It quantifies uncertainty in the deliberator’s parametric model, while also yielding the optimal predictor. In **Chapter 5**, the deliberator processes the knowledge in different way, i.e. the deliberator no longer bounds the expectation of expected loss. Once again, the centralized deliberator-based scheme is adopted to process the participants’ predictor in three alternative ways. Alternative ideal distributions are specified in each case, leading to alternative FPD solutions. In particular, a tractable FPD-optimal prior is derived and adopted for which the optimal predictor is deduced. A thorough comparison of FPD solutions is presented.

In **Chapter 6**, the FPD solutions—derived in Chapter 5— are implemented in order to merge multiple AR nodes, for the purpose of deducing the deliberator’s FPD-optimal prior and predictor. The AR participants process their local data into their local predictive distribution using conjugate Bayesian inference—presented in Chapter 3. The AR deliberator processes these predictors along with their own parametric observation model into an optimal AR parameter prior. The optimized prior is used for prediction in both cases of known and unknown model order.

In **Chapter 7**, the context of Bayesian filtering (BF) is considered, i.e. the case where the observation model is sequentially parametrized with *time-variant* states. The task of BF is to infer these states sequentially and tractably. The exact solution for continuous states is only available in the Kalman filter case. We extend this by considering the dynamic model of states that are missing. In place of FPD which can not be applied in this context, we adopt *stabilized forgetting* (SF) to design the time step of BF. This solution is extended to a rich class of observation models and validated in simulation. Furthermore, the SF solution for an incompletely modelled BF is shown to be a special case of the logarithmic pooling design for merging multiple alternative state predictors. This connects these results to the pooling-based solutions for distributional knowledge processing presented in Chapter 4. We mention how FPD solutions for this BF task might be developed in the future.

Chapter 8 considers this point in the little more detail. In addition, conclusions and future directions related to all technical achievements in the thesis are presented.

Chapter 1 —

The purpose and motivation

Generally, data are defined as a simple observations e.g. numbers, texts, and messages and symbols [1–3]. The data is referred as *raw data* and it does not have meaning in and of themselves [3]. Information is defined as the data with meaning, relevance and purpose [1]. On the other hand, knowledge is defined to be valuable information from the human mind which can be experience, values, insights, truths, beliefs, perspectives, and contextual information [4].

The representation of knowledge is understood as the way in which we express reason and encode knowledge, via thinking rather than acting. In particular, the formulation of knowledge can be considered as the mathematical representation of knowledge. In this thesis, the knowledge is represented either in the form of crisp knowledge which we refer to as *data*, or in the form of probabilities/distributions which we refer to as the *stochastic knowledge*. We aim to process these knowledges in the Bayesian paradigm which is firmly based on axiomatic foundations by providing a unified and consistent framework [5] for knowledge quantification. This paradigm is fundamentally studied in seminal literature such as [6–8]. Empirical facts, beliefs, observations/values and constraints can be considered as the knowledge in the Bayesian paradigm. In the rest of this Chapter, we set up a fundamental probabilistic model of knowledge. Then we introduce a consistent Bayesian mechanism through which we process this knowledge. We present the main objectives of the thesis and then our proposed methodology.

1.1 Probabilistic model

Data can be seen as the realizations of X which might be heterogeneous [9], or multivariate [10]. A core element of any Bayesian paradigm is the specification of a *probability distribution*

(pd) (i.e. *probability model*) which quantifies and manipulates knowledge via the probability. We wish to specify our pd (i.e. to quantify our belief) in the proposition of $x \in \mathbb{S}_x \subset \mathbb{X}$, for a measurable set $\mathbb{S}_x$. In this thesis, we will assume that this requirement is consistently and fully accomplished by specifying a pd, i.e. $F(x)$, for $x \in \mathbb{X}$, defined with respect to (w.r.t) the counting measure when $\mathbb{X}$ is a countable set, and Lebesgue measure, otherwise. In both cases, we will define,

$$P\{x \in \mathbb{S}_x\} \equiv \int_{\mathbb{S}_x} F(x)dx, \quad \text{i.e. } F(x) \text{ is a pd understood via integration.} \quad (1.1)$$

Formally, when $\mathbb{X}$ is countable e.g., $\mathbb{X} \equiv \{x_1, x_2, \dots\}$, $F(x)$ may be expressed as:

$$F(x) \equiv \sum_i p_i \delta(x - x_i), \quad (1.2)$$

where, $\delta(x)$ is the Dirac's δ -function [11], i.e. $F(x)$ is a linear combination of the Dirac's δ -function. Notably,

$$\int_{\mathbb{X}} \delta(x - x_0)g(x)dx = g(x_0), \quad (1.3)$$

if $x \in \mathbb{X}$ is a continuous variable. In the discrete case, $F(x)$ is completely specified by a probability mass function (pmf), i.e. $p_i \equiv \Pr(X = x_i), i = 1, 2, \dots$

The expected value of a function $g(X)$ of the continuous random variable X , is then defined as:

$$E_{F(x)}[g(X)] \equiv \int_{\mathbb{X}} g(x)F(x)dx. \quad (1.4)$$

For the discrete case ($\mathbb{X}$ is countable), (1.4) is specialized via the sifting property [12], as follows:

$$E_{F(x)}[g(X)] \equiv \sum_i p_i g(x_i). \quad (1.5)$$

For simplicity, we will use small letter, e.g. x , to show both the random variable and its value except for some necessary cases where it is explicitly mentioned. Also, the notation $F(\cdot)$ can represent either the probability density function for continuous x , or a probability mass function for discrete x . We will refer to $F(\cdot)$ as a pd in both cases.

1.1.1 Parametric probabilistic model

Generally, $F(x)$ is the nonparametric observation model of x where F is seen as the unknown, infinite-dimensional quantity. Specifically, in the parametric setting, the observation model, $F(x)$, is parametrized by the unknown, finite-dimensional parameter θ . In particular, a parametric probabilistic model of x , will model the observation, so that the parameter θ describes the data-generating system. We will refer to this as parametric observation model—denoted by $F(x|\theta)$. As mentioned, the Bayesian paradigm is a unified and consistent framework which completely quantifies all unknown quantities. In this paradigm, current knowledge about the parameter model is quantified via a pd of the parameter, called the “prior pd”—denoted by $F(\theta)$. An important task in Bayesian paradigm is to choose $F(\theta)$ which is essentially a decision making task. We will refer to this task as the *prior design*. Some widely-used methods for prior design will be reviewed in Chapter 2. In the following, we express the main aims of this research. Then, we propose our methodology to handle the objectives.

1.2 Aims of research and proposed methodology

Generally, we aim to process the knowledge K , in particular, probabilistic form of knowledge in the context of the Bayesian decision making. For this reason, we provide a consistent Bayesian framework for processing knowledge, K , in the optimal way. The knowledge, K , refers to anything that can serve the designer/decision maker towards model construction. As mentioned in the opening Section, K is divided into data and stochastic knowledge. In this thesis we focus on knowledge processing in parametric setting where we wish to design the prior pd, $F(\theta)$. For this reason, we can use the conditional probability (as a Bayesian mechanism) for processing K in order to update $F(\theta)$ to $F(\theta|K)$. This mechanism is symbolically represented as follows:

$$F(\theta) \xrightarrow{K} F(\theta|K). \quad (1.6)$$

We will show that the mechanism (1.6) can be either uniquely handled or widely specified, depending on the following taxonomy. This important taxonomy is the distinction between

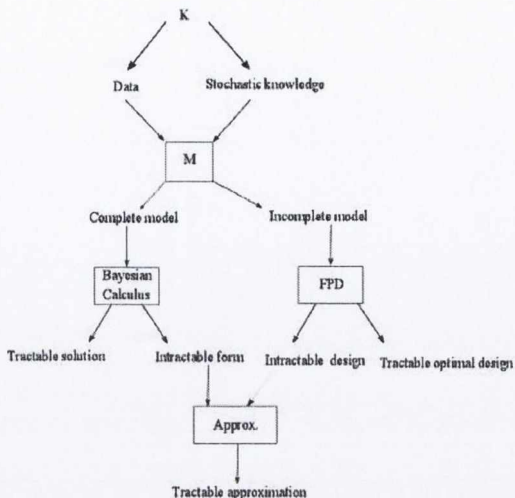


Figure 1.1. The schematic diagram for the taxonomy of knowledge, K , the model of K and the proposed methodologies

the *complete* and an *incomplete* model. We refer to a model as a complete model if the model of knowledge, K , i.e. $M(\theta, K)$, is explicitly available. In particular, K might be a function of some unknown e.g. $K = g(x)$, where the joint model of all unknowns, i.e. $M(x, \theta)$, is required. On the other hand, we refer to the model as an *incomplete* model if the joint model of unknowns is not explicitly available. In Sections 1.2.1 and 1.2.2, we will propose appropriate methodologies to optimally handle the mechanism (1.6) through processing K , depending on whether the model is complete or incomplete. Figure 1.1 fully represents the taxonomy of knowledge, K , the model of K and the proposed methodologies.

Furthermore, *tractability* will be an important property in this thesis. Therefore, it is important to distinguish between the concepts that are *functionally* and *computationally* tractable. These concepts are defined as follows.

Definition 1.2.1. (Functionally tractable). A term is defined to be functionally tractable if it is mathematically derived in the closed form. Notice that this term can be in the form of the standard tabulated function e.g. Gamma function, error function, etc. Notably, most terms presented in this thesis are pds. We consider all standard pds [13] as functionally tractable terms.

For illustration, let

$$\kappa(s) = \int g(x|s)f(x|s)dx, \quad (1.7)$$

then $\kappa(s)$ is functionally tractable if the marginalization (1.7) is computed in an explicit form. For instance, if

$$g(x|s) = 1$$

$$f(x|s) \equiv f(x|\alpha_1, \alpha_2) = x^{\alpha_1-1}(1-x)^{\alpha_2-1}, \quad x \in (0, 1), \quad \alpha_1, \alpha_2 \in \mathbb{R}^+,$$

then,

$$\kappa(s) \equiv \kappa(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)},$$

where $\Gamma(\cdot)$ denotes Gamma function and α_1 and α_2 are known scalars. Notice that the values for $\kappa(s)$ are available via table. On the other hand, if

$$g(x|s) = \exp\{(x-s)^2\},$$

$$f(x|s) = \frac{1}{1 + (x-s)^2}, \quad x, s \in \mathbb{R},$$

then, $\kappa(s)$ cannot be derived in the closed form. In this case, $\kappa(s)$ is not functionally tractable.

Definition 1.2.2. (Computationally tractable). A term is defined to be computationally tractable if it can be computationally evaluated without manipulation and approximation, i.e. is available in number or is directly computed using a table.

For instance, let

$$\kappa(\alpha_1, \alpha_2) \equiv \Gamma(\alpha_1 + \alpha_2) + g(\alpha_1, \alpha_2),$$

where α_1, α_2 are unknown and $g(\cdot)$ is a known function. In this case, $\kappa(\cdot)$ is functionally tractable, however to solve $\kappa(\alpha_1, \alpha_2) \equiv 0$, α_1, α_2 cannot be inverted in a closed form. In particular, inverting the Gamma function involves some manipulations, e.g. the grid search, to specify α_1, α_2 . Therefore, we consider this function as computationally intractable. Recall that most pds have a Gamma function in their normalizing constant. We refer to these pds as computationally tractable pd if the normalizing constant is available in tabulated form. Notice that standard pds are typically functionally tractable. However, they might not be computationally tractable. Some examples will be illustrated in Chapter

4 and 5. The tractability issue will typically happen when a marginalization operator is applied. We will show that in most cases marginalization is not functionally tractable where an approximation needs to be used. Also, the output of marginalization might be in the closed form, however in many cases, its evaluation requires numerical manipulations.

1.2.1 Bayesian data processing

One of the important objectives here is Bayesian data (signal) processing in parametric setting. The signal is typically a discrete time sequence i.e. $\mathbf{x}_n = \{x_1, x_2, \dots, x_n\}$. In signal processing, the signal, $\mathbf{x}_n$, is treated as a stochastic process which mainly emphasizes correlation structure, i.e. the correlation between x_i 's. To be consistent, we refer to the signal as data in the rest of the thesis. Processing data $\mathbf{x}_n$ is quite popular in the task of Bayesian data processing. In Bayesian data processing, observations, $\mathbf{x}_n$, take the role of the knowledge i.e. $K \equiv \mathbf{x}_n$. In data processing, the time-invariant parametric complete model of K is usually available i.e. $F(\mathbf{x}_n|\theta)$, and it is uniquely handled by Bayesian calculus (i.e. Bayes' rule) [8],

$$F(\theta|\mathbf{x}_n) = \frac{F(\mathbf{x}_n|\theta)F(\theta)}{F(\mathbf{x}_n)}. \quad (1.8)$$

Notice that $F(\theta|\mathbf{x}_n)$ in (1.9) is an example of Bayesian conditioning (1.6) where $K = \mathbf{x}_n$. Here, $F(\theta)$ represents our beliefs and uncertainties about an unknown parameter θ , before the data is observed, i.e. prior pd. We refer, $F(\theta|\mathbf{x}_n)$, as the posterior pd of θ , which we read as our updated belief of θ , given that the processed knowledge is $\mathbf{x}_n$. This is a very important example of knowledge processing which allows us to convert the data, $\mathbf{x}_n$, into the pd, $F(\theta|\mathbf{x}_n)$, via Bayesian calculus (1.9). The trivial case of data processing is the *static* case, i.e.

$$x_i \stackrel{iid}{\sim} F(x_i|\theta), \quad i = 1, \dots, n, \quad (1.9)$$

where, *iid*, denotes identically independent distributed. But here we aim to process data using dynamic models. For this reason, we will model the data using the *Autoregressive* (AR) model [14] as one of the flexible models. This AR model is a very important dynamic model for dealing with information bearing signals where there is a regression onto the history of data [15]. In the third Chapter, we will explore Bayesian statistical signal processing as a specific task of Bayesian knowledge processing in the context of the AR

model.

1.2.2 Stochastic knowledge processing

The more advanced task of Bayesian conditioning is related to processing of stochastic knowledge, i.e. $K \equiv F(x)$. The task of processing stochastic knowledge is essentially a Bayesian decision making (DM) task [7, 16]. In this case the parametric model of K , i.e. $M(F(x)|\theta)$ is not available and Bayes' rule is not useful. Therefore, $F(\theta|K)$ is not uniquely available. We note that the knowledge, K , acts as the constraint. In particular, the knowledge constraints the unlimited set of possible pds (i.e. models) to a constrained set which we refer as the *knowledge-constrained* set of possible pds. In Chapter 2, we will show that $F(\theta|K)$ is not uniquely available. Therefore, it becomes a variational quantity that we denote with $A(\theta|K)$, throughout the thesis. Recall that for incomplete models, Bayes' rule does not provide a unique solution of $A(\theta|K)$. This is a core challenge in this thesis and we aim to set up a decision making (DM) scheme by choosing a pd (i.e. model) from the set of candidates, $A(\theta|K)$, through processing K . The fundamental DM task will be fully presented in the second Chapter.

In this thesis, the main methodology is *fully probabilistic design* (FPD) [17–19], where the design is to be understood as choosing a pd. Indeed, this is a DM for the task of the *distributional design*. Some other methods of distributional design have also been explored e.g. *minimum cross-entropy* (MXE) [20] and the *maximum entropy* (ME) [21–23] principles. We are fundamentally interested in the FPD methodology because it is formally justified as a consistent and axiomatic framework of the distributional design [18]. FPD is also consistent with Bernardo's theory of utility-based ranking for decision making in distribution. [24]. FPD is recently justified as the extension of the MXE and the ME principles [19]. We note that in the case where the pd (the member of knowledge-constrained set), is degenerate at crisp data, the FPD-optimal pd is the same as the output of the mechanism of Bayes' rule [25, 26]. The FPD methodology and related principles of distributional design will be fully explained in the second Chapter.

1.2.3 Stochastic knowledge processing in multiple distributed nodes: objectives of thesis

In the recent years, there has been increased interest in determining the optimal merging (i.e. cooperation, interaction, sharing, combination, itegration, fusion, etc.) of the knowledge distributed among multiple knowledge holders which we refer as *nodes*. Nodes are also referred as sources [27], sensors [28, 29], opinions [30] etc.. Recall that the task of merging knowledge is essentially a DM task. There are huge literature in this area i.e. [31–33]. Specifically, the DM task for the distributional design through merging stochastic knowledge in multiple nodes is studied by [34–38]. This task can be accomplished either by one node (i.e. an external node), or can be performed among the nodes themselves. The related DM schemes are respectively the *centralized* scheme and the decentralized (i.e. *distributed*) scheme. Each of these DM schemes is widely studied. For example, the centralized DM scheme is studied in i.e. [28, 39–42]. Also, the decentralized DM scheme is explored in i.e. [30–32, 43]. A comprehensive review of DM schemes and their advantages and disadvantages are presented in the fourth Chapter. Despite the extensive literature on the distributional design through merging knowledge of multiple nodes, there are still significant gaps. The main limitations of current approaches are fully discussed in the Sections 4.2.1.2 and 4.2.2.1.

Recall that one of the main aims is to accomplish the task of data processing using Bayes' rule. In a more advanced setting where Bayes' rule does not provide a solution, we wish to process the stochastic knowledge by investigating the following objectives:

- We fundamentally aim to accomplish the task of distributional design through processing different forms of knowledge.
- We provide a consistent and unified Bayesian scheme of merging knowledge in multiple nodes that allows us to process different knowledge in different ways. For this reason, we adopt a parametric model—which quantifies the uncertainty in the decision maker's model. We then employ the centralized DM scheme of distributional design using FPD.
- We formulate different stochastic knowledges which leads to different knowledge-

constrained sets of possible pds. Processing different knowledge yields different FPD solutions that will be presented throughout Chapters 4 and 5.

- The FPD-optimal solution of parameter prior acts as the prior for the decision maker's generative model of x . In particular, we use this optimal parameter prior via the standard Bayesian calculus (i.e. Bayes' rule) to deduce the FPD-optimal predictor i.e. $\hat{F}(x)$. This optimal predictive pd is employed for processing future data.
- A unique, tractable and optimal weights are automatically derived from our FPD structure. The optimal weights are consistently used as the weights in the optimal predictor. Also these weights can be used in distributional linear pooling which will be fully explored in Chapter 5.
- We will implement our centralized deliberator-based scheme for merging multiple AR nodes where both data and stochastic knowledge are processed. In particular, the local AR nodes process their own local data using Bayesian calculus to deduce their predictive pds. They then report their predictive pds to the central AR node in order to accomplish the task of stochastic knowledge processing (i.e. processing of predictive pds). This is a very important context of science and technology i.e. prediction in financial markets [44], applications of signal processing using AR models [45], seismology [46], multiple step ahead prediction i.e. [47].
- We wish to accomplish Bayesian filtering (BF) for incomplete models where the state evolution model does not explicitly exist. So, the time step of BF needs to be designed. For this reason we process the alternative state predictors using the logarithmic pool to derive the optimal state predictor in time step. Also, the data are processed via Bayes' rule to derive filtering pd in data step.

In the following, we specify the layout of the general tasks that will be presented throughout the thesis.

1.3 Layout of the thesis

This work consists of eight Chapters briefly summarized as follows.

- In the second Chapter, we present the basics of and requirements for distributional decision making. This essentially involves distributional design and distributional approximation. We focus on distributional design and indicate our methodology and its properties.
- In the third Chapter, we fully address the task of Bayesian data processing via the mechanism of Bayes' rule. In particular, we parametrize data using the time-invariant AR model, as one of the flexible models, and perform Bayesian conjugate framework to achieve a tractable inference of unknown parameters, model order inference and future data prediction.
- In the fourth Chapter, we formulate the task of processing the stochastic knowledge in multiple participant systems (MPS). This knowledge includes participant's predictive pds of the common x , and is constrained to bounded expectation of Kullback-Leibler divergence (KLD). For this reason, we will adopt a parametric centralized DM scheme and apply FPD methodology for designing the optimal pd. Furthermore, we will examine optimal solutions within the exponential family.
- In the fifth Chapter, we rearrange the knowledge by relaxing the bounded expectation of KLDs (i.e. the imposed constraint in Chapter 4). Then we process this knowledge (i.e. external predictors of the common x)— provided by multiple participants— in three different ways. We will employ FPD methodology and derive optimal prior and predictive pds underlying each way. Finally different FPD solutions are compared and simulation results are provided.
- In the sixth Chapter, we implement FPD solutions—derived Chapter 5— for processing multiple AR nodes which involves both data and stochastic knowledge. For this reason, AR participants adopt the Bayes' rule mechanism presented in Chapter 3, to process their local data. Then they report the shaping parameters of the predictive pds to the parametric AR deliberator. The AR deliberator processes these predictive pds and their own knowledge to derive the optimal prior and predictor.
- In the seventh Chapter, we process the knowledge in the context of Bayesian filtering (BF) which involves the time step and data step. We note that the time step generally

is intractable and needs to be approximated or designed. We study distributional design of the time step of BF where the model is incomplete (i.e. no state evolution model). Then using the designed time step and Bayes' rule, the data is processed.

- In the eighth Chapter, we conclude the main results of each chapter and suggest the possible implications for further research in this field.

1.4 The novel contributions of the thesis

The main contributions of the thesis are presented in Chapters 3-7 as follows.

In Chapter 3, we propose a novel perspective of the *extended information matrix* (EIM) implied by a formal conjugate Bayesian inference. In particular, we uniquely and tractably process data into the shaping parameters of the posterior pd and predictor in AR model. These shaping parameters are fixed-finite dimensional statistics (i.e. sufficient statistics)—which involve EIM. We contribute the context by exploring both inferential and computational tasks using the prism of EIM for AR model. In particular, we propose a Bayesian inference in a subset of unknowns, Bayesian one-step-ahead (OSA) and multi-step-ahead (MSA) predictors and a new criterion of model order selection. Also, novel recursive algorithms are provided for computing EIM efficiently and the computational load of each recursion is audited. This work is yet to be published.

In Chapter 4, knowing that FPD solution through processing stochastic knowledge in MPS using the centralized scheme is currently available for discrete pd only. In this Chapter, we extend the context by providing two novel contributions. First, the deliberator's unknown pd is parametric on a continuous support. Second, the scheme adopts a hierarchical model for the deliberator, optimizing the deliberator's parameter prior. This allows quantification of the deliberator's uncertainty in their model choice. We formulate conditions under which a conjugate and functionally tractable design is achieved. Some parts of this Chapter are published in [48] and presented in [49, 50].

The limitation of the work in Chapter 4: the main drawback of the proposed FPD solution here is that it might not be generally evaluated and therefore cumbersome numerical manipulations are required. This issue comes from the defined knowledge, K , which

involves the bounded expectation of KLDs. Notably this constraint directly affects the functional form of the FPD solution. Therefore, in the next Chapter we relax this constraint. **In Chapter 5**, we provide two novel contributions. (i) we restructure the knowledge, K , which involve the predictive pds of the same x without the bounded expectation of KLDs. This restructuring of knowledge allows us to derive the FPD-optimal solution with principled and tractable weights. This solution can be derived functionally and in a computationally tractable for a rich class of parametric model of x . The designed weights are used for deducing the optimal predictor of the future data via Bayes' rule. (ii) Three different ways are proposed to merge this knowledge. Each way provides a different FPD solution. A thorough comparison among these FPD solutions is provided. Notably one way of processing this knowledge yields the same functional form of FPD solution presented in Chapter 4. Some parts of this work is presented in [51]. This work is yet to be published. **In Chapter 6**, the main contribution here is to implement the proposed FPD solutions of Chapter 5 for merging multiple AR nodes. For this reason, AR participants adopt conjugate Bayesian inference to deduce their predictors. Also, the AR deliberator employs the proposed FPD solutions in Chapter 5 to deduce their optimal prior and predictor. The FPD-optimal prior and predictor are derived for different cases of AR model with known and unknown model order and an extensive simulation results are provided. This work is yet to be published.

In Chapter 7, The novel contributions of this Chapter are as follows: (i) the task of BF is generally defined. (ii) the time step of BF is designed using *stabilized forgetting* (SF) through merging the alternative state predictor. Then Bayes' rule is employed to derive data step. (iii) A tractable SF is investigated for a rich class of parametric observation model. (iv) the SF solution is extended to general solution for merging multiple alternative state predictors where the logarithmic pooling design is exploited. We perform extensive simulation results for various examples for different operating factors. Some part of this work are published in [52, 53].

Chapter 2 —

Distributional decision making: design and approximation

The decision making (DM) task involves making a choice from a set specification of algorithms, policies, strategies etc. [54,55]. More specifically, the task of choosing a pd from the set of pds can be divided into distributional design and distributional approximation. Recall that processing stochastic knowledge, K , for incomplete model cannot be handled via Bayes' rule. Therefore, we need to accomplish this task using distributional design in order to select the best pd (i.e. model) compatible with our knowledge, K . To do this, We employ FPD [18] as our methodology for designing the optimal pd. Furthermore, in this Chapter the hierarchical extension of FPD is discussed. Some other important principles of distributional design and approximation are reviewed in the sequel. Finally, the promising aspects of FPD are presented.

2.1 The fundamental decision making (DM)

In this section we briefly review the fundamental task of DM. DM is studied in many seminal literatures i.e. [7,8,56]. Generally, the aim of the decision maker ($\mathcal{I}_0$) is to choose the best *decision* (i.e. action)—denoted by d^o — from the set of possible decisions constrained to our knowledge, i.e. $\mathbb{D}_K$. We refer to $\mathbb{D}_K$ as the knowledge-constrained set of decisions. Under Savage's requirements [56], $\mathcal{I}_0$ need to specify a loss to every decision in $\mathbb{D}_K$. The loss function associated with the decision d when the true decision is d_t is defined as follows:

$$L(d, d_t), \text{ where, } L : \mathbb{D}_K \rightarrow \mathbb{R} \quad (2.1)$$

We note that the negative loss function is referred as to the utility function which is more common in economics i.e. [57]. If $\mathcal{I}_0$ is certain about the value of the loss resulted from each decision, it is easy to choose the best one which has the lowest loss value. Yet, the loss resulting from an decision depends on features beyond $\mathcal{I}_0$'s control and they ¹ are uncertain about features. Therefore, they need to take expectation of the associated loss w.r.t. the probability of x as follows:

$$R(d, d_t) \equiv E_{F(x)}[L(d(x), d_t)]. \quad (2.2)$$

The expected loss (2.2) is well known as *risk* i.e. [7, 56]. Furthermore, the expected loss (2.2) ranks all possible decisions to the true decision, d_t . For this reason, it is referred to as the *ranking rule* in the practical contexts i.e. [58] and [59]. Using (2.2), we choose the best decision, d^o , as follows:

$$d^o = \arg \min_{d \in \mathbb{D}_K} R(d, d_t), \quad (2.3)$$

Notably, $\mathbb{D}_K$ is defined as the set of possible decisions constrained to our knowledge.

2.2 Loss function for distributional decision making

Recall that the main aim of this thesis is to select the optimal pd from the possible set of pds. We refer to this task as the distributional DM which involves both distributional design and approximation. The difference between distributional design and approximation can be understood via the existence of the true (zero-expected loss) pd. In distributional design, the true pd does not exist and it needs to be designed. On the other hand, in distributional approximation the true pd is available but it requires to be approximated due to intractability, high dimensionality etc. Figure 2.1 shows the key distinction between the distributional design and approximation. Following the Bernardo's theory [24], the loss function (2.1) for distributional DM is defined as follows:

$$L(F_1(x), F_2(x)) = \ln \frac{F_1(x)}{F_2(x)}, \quad (2.4)$$

¹Throughout this thesis, we adopt the gender-neutral possessive subject/adjective, "they/their", in place of *he/his* or *she/her*.

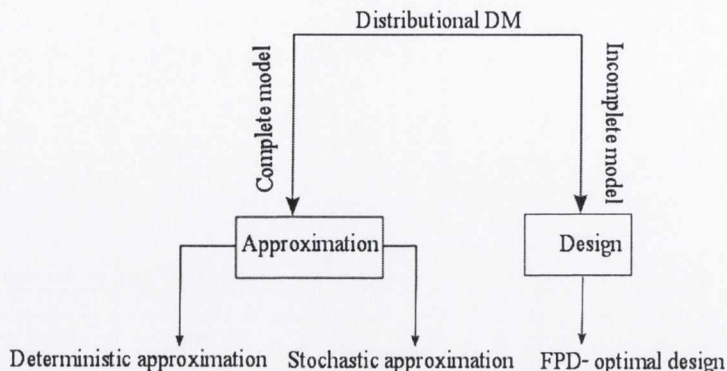


Figure 2.1. The schematic diagram for the taxonomy of distributional DM

This is the loss evaluated at a particular realization of x . In distributional approximation [24], $L(F_1(x), F_2(x))$ (2.4) measures the loss associated with choosing each possible $F_2(\cdot)$ (i.e. the candidate pd to approximate $F_1(\cdot)$), when the true pd is $F_1(\cdot)$. On the other hand, distributional design is related to the case where the true pd does not exist. In this case, we follow Kárný and Kroupa's theory [18] where the FPD methodology is applied. FPD axiomatically requires an ideal pd to rank all pds in the knowledge-constrained set i.e. $F_2 \equiv F_I$.

2.2.1 Kullback-Leibler divergence (KLD) and its properties

In this Section, we aim to specialize the expected loss (2.2) for distributional DM. The expectation of the loss function (2.2) w.r.t. the first argument is the core objective function—which is well known as Kullback-Leibler divergence (KLD) [60]—defined as follows.

Definition 2.2.1 (KLD). Let $F_1 \in \mathbb{F}_1 \neq \emptyset$, $F_2 \in \mathbb{F}_2 \neq \emptyset$, and $D : \mathbb{F}_1 \times \mathbb{F}_2 \rightarrow \mathbb{R}^+$, then using the loss function (2.4), the KLD for a continuous x is defined as:

$$\begin{aligned}
 D[F_1||F_2] &\equiv R[F_1, F_2] \\
 &= E_{F_1} \left[\ln \frac{F_1(x)}{F_2(x)} \right] \\
 &= \int_x \left(\ln \frac{F_1(x)}{F_2(x)} \right) F_1(x) dx,
 \end{aligned} \tag{2.5}$$

being KLD [60] from $F_1(\cdot)$ to $F_2(\cdot)$.

KLD is also well-known as cross-entropy i.e. [20]. KLD is widely used for various purposes in different contexts i.e. for distributional design in signal processing [61], for optimal strategies in control design [62], for information fusion in wireless networks [63], and for forecast aggregation in economic [64], to name a few. Recently KLD has been greatly attractive due to its properties. These properties have been reviewed in many literatures i.e. [60] and [65]. We here summarize the key properties of KLD as follows.

- **KLD-value:** The nature of $\mathbb{F}_1$ and $\mathbb{F}_2$, leads to different values of $D[.||]$ (2.5) as follows:

$$D[F_1||F_2] = \begin{cases} < \infty, & \text{supp}(\mathbb{F}_1) \subseteq \text{supp}(\mathbb{F}_2) \\ = \text{does not exist}, & \text{supp}(\mathbb{F}_2) \subset \text{supp}(\mathbb{F}_1) \end{cases}$$

- **Non-negativity:** KLD is non-negative, i.e.

$$D[F_1||F_2] \geq 0. \quad (2.6)$$

- **Self-similarity:**

$$D[F_1||F_2] = 0, \Leftrightarrow F_1 = F_2 \quad \text{a.s.}, \quad (2.7)$$

where ‘a.s.’ denotes ” *almost surely*”. Recall that, in probability theory, an event happens almost surely if it happens with probability one. This implies that the set of possible exceptions may be non-empty, but it has probability zero.

- **Asymmetry:**

$$D[F_1||F_2] \neq D[F_2||F_1]. \quad (2.8)$$

- **Non-triangle inequality:** for three pds F_1 , F_2 , and F_3 ,

$$D[F_1||F_3] \not\leq D[F_1||F_2] + D[F_2||F_3]. \quad (2.9)$$

The equation (2.9) asserts that the KLD does not obey the triangle inequality.

- **Convexity in the first argument:** the KLD obeys the Jensen's inequalities which is related to convexity. Let $\mathcal{F}$ be a convex function, then

$$\mathcal{F}(D[F_1||F_2]) \leq D[\mathcal{F}(F_1)||F_2], \quad (2.10)$$

where $\mathcal{F}(F_1)$ is point-wisely defined. Notice that KLD is not necessarily convex in the second argument.

Remark 2.2.1. (*The relationship between KLD, Kerridge inaccuracy and Shannon entropy*). KLD (2.5) can be decomposed to the Kerridge inaccuracy and the Shannon entropy as follows.

$$D[F_1||F_2] = \text{Kerr}[F_1||F_2] - H[F_1], \quad (2.11)$$

where,

$$\text{Kerr}[F_1||F_2] \equiv E_{F_1} \left[\ln \frac{1}{F_2(x)} \right], \quad (2.12)$$

is the Kerridge inaccuracy [66] and

$$H[F_1] \equiv -E_{F_1} [\ln F_1(x)], \quad (2.13)$$

is the Shannon entropy [67]. It immediately follows that the solution of minimum KLD equals to the solution of the maximum entropy if and only if Kerridge inaccuracy is zero. An obvious choice of holding this equality is that $F_2(\cdot)$ be the standard uniform pd presented in Appendix A.3.

Definition 2.2.2. (*Closed-form KLD*). Using Definition 1.2.2, the KLD, $D[F_1||F_2]$, is available in the closed form if the integration (2.5) is solved functionally tractable i.e. no approximation is required.

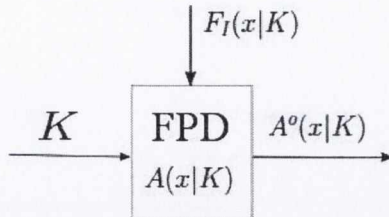


Figure 2.2. The schematic diagram for processing K via FPD (2.14).

2.3 Distributional design

We here aim to review well-known principles of distributional design. Most of these principles are based on KLD (2.5). We note that in distributional design the zero-loss pd does not exist. Therefore, it is required to specify a pd (i.e. fixed-ideal pd) to rank all possible true (unreduced) pds with that. One of the most important principles is the *fully probabilistic design* (FPD) [17, 62]. FPD is formally justified as an axiomatic framework [18] for the design of the optimal pd. FPD chooses the optimal choice of the true pd by minimizing KLD (2.5) with respect to its first argument as follows.

Definition 2.3.1 (FPD principle). *Let $F_I(\cdot)$ be the fixed ideal pd of x , then the FPD-optimal design of $A(\cdot)$ — denoted by $A^o(\cdot)$, is*

$$A^o(x|K) \equiv A^o \equiv \arg \min_{A \in \mathbb{A}_K} D[A||F_I], \quad (2.14)$$

where, $\mathbb{A}_K$ denotes the constrained set of all possible true pds to our knowledge, K .

Recall that FPD (2.14) is achieved using KLD (2.5) replacing $F_1(\cdot)$ with $A(\cdot)$ and $F_2(\cdot)$ with $F_I(\cdot)$. The schematic diagram of processing K via FPD is depicted in Figure 2.2.

The axioms that are accepted as requirements for an uncertainty measure are informally presented in [20]. Also the requirements of Bayesian DM under uncertainty are listed in [68]. Using [18, 20], we can summarize axioms of FPD as uniqueness, consistency and completeness. To show consistency, let $\mathbb{L}$ be the set of loss functions associated with the same ideal pd $F_I(x)$, $\forall L \in \mathbb{L}$. Then the logarithmic loss function (2.4) is the represen-

tative loss function of $\mathbb{L}$ whose values do not depend on the realization of X . Also, the completeness is referred to a complete ordering of the possible pd, i.e. A , in order to select the optimal pd i.e. A° , as the most preferred one in terms of this ordering. This ordering is quantified via a local functional E acting on $\mathbb{L}$, i.e. $E(L)$. This quantified functional is equivalent to $D[.||]$ in (2.14).

We note that FPD [18] and Bernardo's prescription [24] are consistent, since the logarithmic loss function (2.4) is used for both methods. However they solve different tasks of distributional DM.

Remark 2.3.1 (Ideal specification). *The role of $F_I(x)$ in FPD is to quantify preferences and aims of the decision maker— which can be considered as the origin of the knowledge-constrained set, $\mathbb{A}_K$, i.e. $D[F_I||F_I] = 0$, a.s.. The $F_I(\cdot)$ can be interpreted as the ideal choice where no knowledge is applied. We note that the ideal pd need not be in the knowledge-constrained set, $\mathbb{A}_K$. Under Kárný and Kroupa's requirements [18], it is not prescribed by FPD that whether F_I is a member of $\mathbb{A}_K$ or not. But in the case where $F_I \in \mathbb{A}_K$, FPD solution is trivial i.e. $A^\circ \equiv F_I$ a.s.. Furthermore, K forces $F_I(\cdot)$ be specified outside $\mathbb{A}_K$ and we wish to design the optimal pd such that it has the smallest KLD with $F_I(x)$ using (2.14). Therefore, in the rest of this thesis, we specify the ideal pd outside the knowledge-constrained set.*

Notice that processing, K , through FPD (2.14) implies that the whole knowledge is firstly gathered and then processed— which we refer as *one-batch* (i.e. *off-line*) processing. Another way of processing K , is the case where the task of processing interleaves with updating knowledge. In this case, K is processed in a sequence so that one FPD is applied for each step. We refer to this way of processing K as *sequential* (*on-line*) processing. For better sense, we define the sequence of knowledge being processed in each step as follows:

$$K_{1:j} = \{K_1, \dots, K_j\}, \quad j \in \{1, \dots, s\}. \quad (2.15)$$

Notice that (2.15) represents $K_{1:1} \equiv K_1$ and $K_{1:s} \equiv K$. The FPD principle for sequential processing of K is defined as follows.

Definition 2.3.2 (FPD sequential processing). *Using (2.14), the FPD-sequential processing of the sequence of knowledge (2.15) is:*

$$A^o(x|K_{1:j}) \equiv A_j^o \equiv \arg \min_{A \in \mathbb{A}_{K_{1:j}}} D[A||A_{j-1}^o], \quad j = 1, \dots, s \quad (2.16)$$

where A_j^o denotes the output of j^{th} step, and $A_0^o \equiv F_I$.

The hierarchical extension of FPD (2.14) is proposed by [19], where the decision maker's uncertainty model for x is defined hierarchically. The hierarchical setting is as follows. Let G be an unknown, infinite-dimensional pd and so we model:

$$\begin{aligned} x|G, K &\sim G(x|K), \quad x \in \mathbb{X}, \quad G \in \mathbb{G}, \\ G|S, K &\sim S(G|K), \quad S \in \mathbb{S}. \end{aligned} \quad (2.17)$$

Defining the decision maker's knowledge, K , as:

- The unknown parameter, G , is specialized to a pd of x ,
- x is conditionally independent of S given G .

Then the hierarchical model of x is defined as follows:

$$\begin{aligned} M(x, G|S, K) &\equiv M(x|G, S, K)M(G|S, K) \\ &\equiv G(x|K)S(G|K), \quad G \in \mathbb{G}, \quad S \in \mathbb{S}. \end{aligned} \quad (2.18)$$

Notice that $S(G|K)$ is not explicitly available. Therefore, it needs to be designed. Using the principle (2.14), the non-parametric hierarchical FPD principle [19] is defined as follows.

Definition 2.3.3 (Non-parametric hierarchical FPD). *Using (2.18), the FPD-optimal design for the variational quantity S in the hierarchical model is:*

$$S^o(G|K) \equiv \arg \min_{S \in \mathbb{S}_K} D[M||M_I], \quad (2.19)$$

where the ideal hierarchical model of x is denoted by $M_I \equiv M_I(x, G|S, K)$, $Supp(M) \equiv Supp(M_I)$.

The non-parametric hierarchical setting (2.18) can be specialized to standard parametric model in two different ways as follows:

Remark 2.3.2. (*Parametric hierarchical FPD*).

1. ***G is parametric:*** in this case, x is parametrized with the finite-dimensional unknown parameter, θ , having the same pd given either by G or θ i.e.,

$$x|G \stackrel{d}{=} x|\theta, \stackrel{d}{=} (\text{denoting the same pd}), G \in \mathbb{G}. \quad (2.20)$$

This asserts that $\mathcal{I}_0$'s model for x is determined by finite-dimensional unknown parameters, θ . In this case, the hierarchical model (2.18) becomes:

$$M(x, \theta|G, A, K) \equiv G(x|\theta, K)A(\theta|K). \quad (2.21)$$

where A denotes the parameter pd which needs to be designed. In (2.21), θ fulfils the conventional roles of a hyperparameter, hidden field, missing data, etc. , depending on the context. We will widely adopt this parametric setting for merging distributed knowledge in multiple participant context throughout Chapters 4-6.

2. ***A is parametric with unknown hyperparameters:*** This case provides a parametric relaxation of the previous case such that:

$$\begin{aligned} x|G &\stackrel{d}{=} x|\theta, \\ \theta|A &\stackrel{d}{=} \theta|\eta, \end{aligned} \quad (2.22)$$

where η is the unknown hyperparameter. In this case, the hierarchical model (2.18) becomes:

$$M(x, \theta, \eta|G, A, S, K) \equiv G(x|\theta, K)A(\theta|\eta, K)S(\eta|K). \quad (2.23)$$

Notice that θ is conditionally independent of S given A .

In the following we define other important principles for distributional design.

Definition 2.3.4 (minimum cross entropy (MXE) principle for distributional design). Let $F_P(x)$ be the prior pd of x , then the MXE-optimal design of prior pd [20]—denoted by $A_{MXE}(\cdot)$, is

$$A_{MXE}(x) \equiv A_{MXE}^o \equiv \arg \min_{A \in \mathbb{A}_K} D[A||F_P], \quad (2.24)$$

It is shown [69] that the MXE principle and Bayesian conditioning are usually consistent.

Definition 2.3.5 (maximum entropy (ME) principle for distributional design). Let $\mathbb{A}$ be the constrained-knowledge set of possible pds. then using (2.13) the ME-optimal design of $A(\cdot)$ —denoted by $A_{ME}(\cdot)$, is

$$A_{ME}(x) \equiv A_{ME} \equiv \arg \max_{A \in \mathbb{A}_K} H[A], \quad (2.25)$$

Remark 2.3.3. (Special cases of FPD). FPD is justified i.e. [19] as the extension of other established principles i.e. minimum cross entropy (MXE) and maximum entropy (ME) principles of distributional design.

- MXE principle i.e. [20]: This principle is a special case of FPD where $F_I(\cdot)$ is explicitly assumed to be a prior pd.
- ME principle i.e. [22], [21]: This principle is a special case of FPD where $A_I(\cdot)$ is explicitly assumed to be a uniform pd.

2.3.1 Distributional parameter prior design

Recall that our proposed distributional design will aim to process quite extensive structures of knowledge, K , and this knowledge need not be sequentially followed by data. In parametric setting, we specifically process K which involves the parametric observation model, $F(x|\theta)$, then one task of distributional design will be the task of parameter prior design. Recall that for processing data modelled with time-invariant parameter, the Bayes' rule (1.9)—as a consistent mechanism in (1.6)—yields the unique solution for $F(\theta|K)$.

2.3.1.1 Parameter prior classes and designs

In this section, we specialize distributional design to the context of prior design. For this reason, we review the most popular classes of parameter prior design as follows.

Definition 2.3.6 (Conjugate prior). *The pd $F(\theta|\mathbf{s}_0)$ is defined to be conjugate prior to the parametric observation model, $F(x|\theta)$, if the form of posterior pd i.e. $F(\theta|\mathbf{s}_n)$, is identical to the form of prior, $F(\theta|\mathbf{s}_0)$, i.e. the inference is functionally invariant w.r.t. Bayes' rule,*

$$F(\theta|\mathbf{s}_n) \propto F(\mathbf{x}_n|\theta)F(\theta|\mathbf{s}_0), \quad (2.26)$$

where $\mathbf{s}_n$ is fixed finite-dimensional shaping parameters of posterior pd and it will be updated as

$$\mathbf{s}_n = s(\mathbf{x}_n, \mathbf{s}_0), \quad s \in \mathbb{R}^q, \quad q < \infty. \quad (2.27)$$

One important class of conjugate prior pd is *conjugate exponential family* (CEF) (A.3) in Section A.2.

Definition 2.3.7 (Jeffreys prior). *Let $F(x|\theta)$ be the parametric observation pd, with univariate parameter θ . Jeffreys prior is directly derived from, $F(x|\theta)$ via information matrix, $I(\theta)$, as follows:*

$$F(\theta) \propto \sqrt{I(\theta)} \quad (2.28)$$

where $I(\theta)$ is defined as:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log F(x|\theta)\right]. \quad (2.29)$$

Notice that Jeffreys prior is the limit of conjugate prior, i.e. conjugate prior converges to Jeffreys prior by diffusing shaping parameters of conjugate prior.

Definition 2.3.8 (Reference prior). *The basic idea of reference priors originates in Bernardo [70]. This is a prescriptive procedure to design the prior where it describes*

approximately inferences with “little prior information”. Here, the designer’s knowledge, K , is the relationship between x and θ via the parametric observation model, $F(x|\theta)$. Following [70] the objective function which needs to be optimized is:

$$I_\theta(K, A_P) = E_{F(x)}[D(F_{Post}||A_P)] \quad (2.30)$$

where $F_{Post} \equiv F_{Post}(\theta|x)$ denotes the fixed posterior pd, $A_P \equiv A_P(\theta)$ is the variational prior pd, $\mathbb{A}_K$ is the knowledge-constrained set and $D[.||]$ is the KLD defined in Definition 2.2.1. The objective function (2.30) is known as the expected information about θ provided by K . Therefore using (2.30), the reference prior A_P is achieved as,

$$A_P = \arg \max_{A_P \in \mathbb{A}_K} I_\theta(K, A_P). \quad (2.31)$$

2.4 Distributional approximation

Distributional approximation is the other important task of distributional DM. This task is to choose the optimal approximation, $\tilde{F}(\theta) \in \check{\mathbb{F}}_K$, from the set of possible approximating pds constrained to our knowledge, $\check{\mathbb{F}}_K$. Notice that in distributional approximation the complete model is available. So using Bayesian calculus $F(\theta|K)$ exists, however the key problem is that $F(\theta|K)$ is not in tractable/standard form and it has to be replaced with a tractable approximation. Distributional approximation is mainly divided into two kinds of techniques: stochastic technique and deterministic technique.

2.4.1 Distributional stochastic approximation

Stochastic distributional approximation maps the true pd $F(\cdot)$ to the randomly-generated approximation, $\tilde{F}(\cdot)$. Stochastic methods allow us generate a random sample,

$$\theta^{(i)} \sim F(\theta|x_n) \quad (2.32)$$

$$\{\theta\}_m = \{\theta^{(1)}, \dots, \theta^{(m)}\} \quad (2.33)$$

Notice that the set of sample is used to construct the *empirical* pd [71],

$$\tilde{F}(\boldsymbol{\theta}|\mathbf{x}_n) = \frac{1}{m} \sum_{i=1}^m \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}), \quad (2.34)$$

where $\delta(\cdot)$ is the Dirac δ -function (1.3) located at $\boldsymbol{\theta}^{(i)}$. By the strong law of large numbers [72], the empirical pd (2.34) will almost surely (a.s.) converge to the true pd $F(\boldsymbol{\theta}|\mathbf{x}_n)$. Notice that achieving the true pd $F(\boldsymbol{\theta}|\mathbf{x}_n)$ is quite difficult in practice due to the high dimensionality of $\boldsymbol{\theta}$ and the increment of the size of $\boldsymbol{\theta}$ over time. In Chapter 8, we will fully explain some popular stochastic methods i.e. Monte Carlo Markov Chain (MCMC) techniques [73] and Sequential Monte Carlo (SMC) [74].

Although distributional stochastic approximation is widely used, in particular for high-dimensional variables, distributional deterministic approximation is of more interest due to providing a trade-off between computational load and accuracy. Therefore we review the scope of distributional deterministic approximation in the following part.

2.4.2 Distributional deterministic approximation

Distributional deterministic approximation involves well-established techniques to approximate intractable pds based on some well-known distance. In the following, we review very popular techniques of deterministic distributional approximation.

- **The certainty equivalent (CE) approximation:** in this approach, the full posterior pd, $F(\boldsymbol{\theta}|\mathbf{x}_n)$ is replaced by Dirac δ -function located at the point estimate, $\hat{\boldsymbol{\theta}}$,

$$F(\boldsymbol{\theta}|\mathbf{x}_n) \approx \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}). \quad (2.35)$$

The approximation (2.35) is known as the *certainty equivalent* principle [75].

In particular, KLD-based deterministic distributional approximation is quite popular in the literature. Some of the important ones are technically reviewed as follows.

- **Minimum risk (MR) approximation:** Following the Bernardo's prescription [24], the MR approximation is as follows:

$$\tilde{F}(\boldsymbol{\theta}) \equiv \arg \min_{\tilde{F} \in \mathbb{F}_K} D[F||\tilde{F}], \quad (2.36)$$

where the optimal approximation, $\tilde{F}(\cdot)$, is chosen from the set of possible approximating pds, $\check{\mathbb{F}}$, and $F(\cdot)$ denotes the fixed true pd. The optimal approximation, $\tilde{F}(\cdot)$ in (2.36) has the minimal risk.

Notably the Bernardo's approximation (2.36) is not FPD-optimal design (2.14). Since in FPD (2.14), the variational argument of the KLD is the first argument which is totally different from the optimal approximation (2.36) where the variational quantity is the second argument of KLD.

- **Variational Bayes (VB) approximation:** the VB approximation is the other KLD-based method of distributional approximation which is widely adopted in the literature i.e. [61] and [76]. The VB approximation is defined as follows:

$$\tilde{F}(\boldsymbol{\theta}) \equiv \operatorname{argmin}_{\check{F} \in \check{\mathbb{F}}} D[\check{F}||F], \quad (2.37)$$

where, $\check{\mathbb{F}}$ is restricted functionally, using the assumption of conditional independent i.e.,

$$\check{\mathbb{F}} \equiv \left\{ \check{F} : \check{F}(\boldsymbol{\theta}) = \check{F}(\boldsymbol{\theta}_1)\check{F}(\boldsymbol{\theta}_2) \right\}.$$

Notice that a necessary condition for applying the VB approximation is that $\boldsymbol{\theta}$ be multivariate [61], e.g. $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T]^T$. Recall that VB approximation (2.37) is achieved using KLD (2.5) replacing $F_1(\cdot)$ with $\check{F}(\cdot)$ and $F_2(\cdot)$ with $F(\cdot)$.

2.5 Properties of fully probabilistic design (FPD)

In this thesis, we are interested in FPD methodology because of promising prospects which are summarized as follows.

- FPD is a consistent probabilistic framework, which processes stochastic knowledge and the aims/constraints and preferences of the decision maker, in a common Bayesian form, yielding a unique solution.

- FPD enables us to quantify and process the $\mathcal{I}_0$'s multiple aims. Also, a multi-modal ideal pd can be processed via FPD i.e. [18]. As a result, it allows us to merge preferences expressed by several ideal pds in the multiple participant context.
- FPD is characterized as an approximate dynamic programming technique [77], with KLD taking the role of the value function, and yielding an explicit and the unique minimizer via solution of a nonlinear integral equation i.e. [78].

Figure 2.3 shows the schematic diagram for KLD-based optimal distributional designs presented in Section 2.3. Also, the schematic diagram for KLD-based distributional deterministic approximations presented in Section 2.4, follows in Figure 2.4.

2.6 Discussion

In this chapter, we have reviewed the basic DM task and then specialized this task into distributional DM. Two main tasks of distributional DM are distributional design and distributional approximation which are presented in Sections (2.3) and (2.4) respectively. We have noted that the difference between design and approximation is related to whether the true pd exists or not. Also, different ordering of the arguments of KLD leads to different methods. The most important methods of distributional design and approximation are discussed.

Notice that in this Chapter, we have not reviewed the current methods for merging knowledge in the context of multiple nodes. Instead, we thoroughly review the competing approaches, popular schemes of merging distributed knowledge in MPS and the related works in Chapter 4.

In Chapters 4, 5 and 6 we will employ FPD (2.14) as the core methodology for accomplishing the task of distributional design. In particular, FPD is exploited for designing the optimal merging of knowledge provided by multiple nodes. For this purpose, we formulate the knowledge of the system into the knowledge-constraint set. Then we apply the fundamental definition of FPD for designing the optimal pd. The proposed FPD-optimal design accommodates all properties of FPD which were highlighted in Section 2.5.

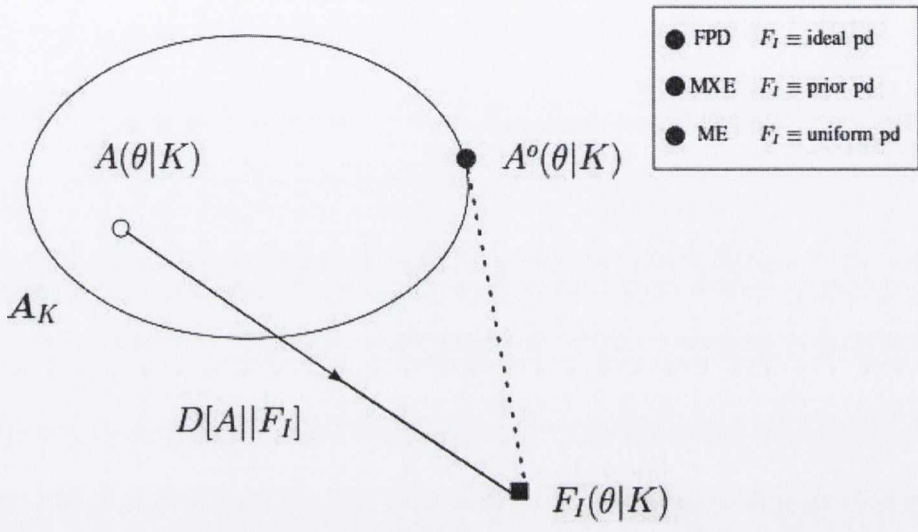


Figure 2.3. The schematic diagram for KLD-based optimal distributional designs (2.14), (2.24) and (2.25).

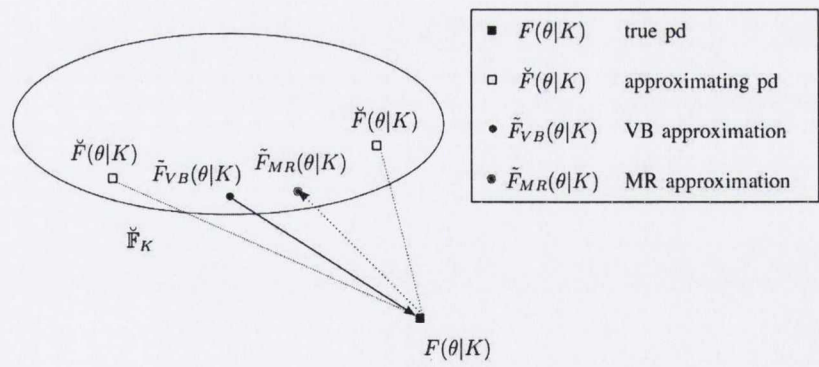


Figure 2.4. The schematic diagram for KLD-based distributional deterministic approximation (2.36) and (2.37).

Chapter 3 —

Bayesian statistical data processing

In this Chapter, we aim to process data (as the non-probabilistic form of knowledge) to infer the posterior pd of the parameters of interest and predictive pd of the quantity of interest, x . For this reason, we model the data with the *autoregressive* (AR) model and adopt the conjugate prior pd presented in Definition 2.3.6. We exploit Bayes' rule to process the data into fixed finite dimensional statistics—which completely specify the posterior pd and predictor. Data processing here involves both inferential tasks—which are performed analytically tractable—and novel algorithms which are computationally efficient. The inferential tasks will be presented in Sections 3.3 and 3.4. Then the computational tasks will be explored in Section 3.6 in sequel.

3.1 Linear regression models

Linear regression model is a special class of linear models which generally has the additive structure as follows:

$$\mathbf{D} = \mathbf{A}\Psi + \mathbf{E} \quad (3.1)$$

where $\mathbf{D} \in \mathbb{R}^{p \times m}$ is the matrix of observed data, $\mathbf{A} \in \mathbb{R}^{p \times q}$ is the parameter (coefficient) matrix, $\Psi \in \mathbb{R}^{q \times m}$ is the regression matrix, $\mathbf{E} \in \mathbb{R}^{p \times m}$ is the matrix of noise. We note that (3.1) is the *off-line* representation of linear regression model where $\mathbf{D}$ has been gathered and saved, at once. Then, the task of data processing will be commenced. Alternatively, linear regression model can be represented in *on-line* scenario where $\mathbf{D}$ are accumulated incrementally with time (i) growing possibly up to infinity. In this scenario, the processing data interleaves with the data acquisition throughout times. The on-line representation of linear regression model is:

$$\mathbf{x}_i = \mathbf{A}\psi_i + \mathbf{e}_i \quad i = 1, 2, \dots \quad (3.2)$$

In regression models, either $\mathbf{A}$ or Ψ is assumed to be known. The aim is to infer the other one to accomplish the task of data processing.

3.2 AR model

The task of data processing using the AR model has been the target of many studies both in classical and Bayesian paradigms. The AR model as a flexible model is widely adapted with all-poles model which has many applications i.e. speech processing [79], [80]; telecommunication networks [61], [81], [82]; data analysis [45]; ecology [83]; economics [84], finance [44]; and wireless communication [85]. The AR model is the specific case of regression model (3.2) where $\mathbf{x}_i$ is regressed on the history of previous data, and ψ_i is assumed to be known, i.e. it may contain any observed variables or their known transformations. Formally,

$$\psi_i = g(\mathbf{D}_{i-1}, \mathbf{y}_i),$$

where auxiliary variable, $\mathbf{y}_i$, may contain any known variables, such as a measured external (exogenous) signal, time variable, i , for time-variant systems, etc. For instance, if the external input, $\mathbf{y}_i$, be a time-variant auxiliary variable, the model (3.2) is specialized into the AR model with exogenous variable (ARX) i.e. [61]. For better understanding we will focus on the univariate AR model in the rest of this chapter. Recall that, the key tasks in the data processing using the AR model are inferential tasks and studying the computational complexity. Inferential tasks mostly involve the inference of unknown parameters, model order estimation and future prediction. Furthermore, the survey on computational complexity involves recursive algorithms to reduce computational load. In particular, the reduction of computational loads is essential for processing real-time data or processing high-dimensional data. Despite extensive research on AR models, surprisingly, there are missing parts yet to be studied. Therefore, it is worthwhile to explore the Bayesian data processing using time-invariant parametrization of the AR model to solve some of the current issues.

3.3 Time-invariant parametrization of AR model: known model order

In this section, we aim to model the data using the AR model in Bayesian paradigm. The Wold representation of the univariate, parametric the AR model with known model order p , is defined as,

$$x_i = \sum_{j=1}^p a_j x_{i-j} + e_i, \quad i = p+1, \dots, n, \quad (3.3)$$

where e_i is the observation error assumed to be normal with zero mean and variance r i.e.,

$$e_i \stackrel{ciid}{\sim} \mathcal{N}_e(0, r), \quad i = p+1, \dots, n, \quad (3.4)$$

where, $\mathcal{N}$, denotes the normal pd and *ciid* denotes that e_i 's are conditionally independent identical distributed. Therefore, the univariate AR observation model [61] is,

$$F(x_i | \mathbf{x}_{i-1}, \mathbf{a}^{(p)}, r, p) \equiv F(x_i | \psi_i^{(p)}, \mathbf{a}^{(p)}, r, p) \equiv \mathcal{N}_x(\psi_i^{(p)T} \mathbf{a}^{(p)}, r), \quad i = p+1, \dots, n, \quad (3.5)$$

where, $\mathbf{a}^{(p)}$ and r are the unknown parameters of the AR model, and $\psi_i^{(p)}$ is the regressor vector. Notice that the observation model (3.5) belongs to *dynamic exponential family* [61]. The Wildcard form of p -conditional joint pd of data and unknowns is,

$$F(\mathbf{x}_n^{(p)}, \mathbf{a}^{(p)}, r | p) = \prod_{i=p+1}^n F(x_i | \psi_i^{(p)}, \mathbf{a}^{(p)}, r, p) F(\mathbf{a}^{(p)}, r | p) \quad (3.6)$$

where, $\mathbf{x}_n^{(p)} = [x_{p+1}, \dots, x_n]^T$, the first term in right hand side (3.6) is the observation model (3.5) from $p+1$ to n , and the second term is the prior pd of unknowns.

Remark 3.3.1. (*Compact Wold representation*)

The compact form of Wold representation is,

$$\mathbf{x}_n^{(p)} = \Psi_n^{(p)} \mathbf{a}^{(p)} + \mathbf{e}_n^{(p)}, \quad (3.7)$$

where, $\mathbf{e}_n^{(p)} = [e_{p+1}, \dots, e_n]^T$ and $\Psi_n^{(p)} = [\mathbf{g}_{n,1}, \dots, \mathbf{g}_{n,p}]^T$ are defined in the Section

notational conventions.

3.3.1 The p -conditional parameter inference of AR model

In this section, our aim is the inference of p -conditional unknown parameters of the AR model, $(\mathbf{a}^{(p)}, r)$. In particular, we are interested in deducing the tractable inference of parameters. Tractability is assured when we adopt the conjugate prior pd (Definition 2.3.6) and the form of the posterior pd is identical for all i . Adopting a conjugate prior pd i.e. Normal-inverse Gamma (NiG) [61],

$$\begin{aligned} F(\mathbf{a}^{(p)}, r|p) &\equiv \text{NiG}_{\mathbf{a}^{(p)}, r}(\mathbf{V}_p^{(p)}, \nu_p^{(p)}) \\ &= \frac{r^{\frac{-\nu_p^{(p)}}{2}}}{\kappa_{\mathbf{a}, r}(\mathbf{V}_p^{(p)}, \nu_p^{(p)})} \exp \left\{ -\frac{1}{2r} [-1, \mathbf{a}^{(p)T}] \mathbf{V}_p^{(p)} [-1, \mathbf{a}^{(p)T}]^T \right\}, \end{aligned} \quad (3.8)$$

where $\nu_p^{(p)}$ is the counter, $\mathbf{V}_p^{(p)}$ is the *extended information matrix* (EIM)— which is partitioned as follows,

$$\mathbf{V}_p^{(p)} = \begin{bmatrix} v_{p,11}^{(p)} & \vdots & \mathbf{v}_{p,a1}^{(p)T} \\ \dots\dots\dots & & \\ \mathbf{v}_{p,a1}^{(p)} & \vdots & \mathbf{V}_{p,aa}^{(p)} \end{bmatrix}. \quad (3.9)$$

The EIM, $\mathbf{V}_p^{(p)} \in \mathbb{R}^{(p+1) \times (p+1)}$, is partitioned into blocks, isolating $v_{p,11}^{(p)}$ the $(1, 1)$ element, $\mathbf{v}_{p,a1}^{(p)}$ is a $(p \times 1)$ vector, and $\mathbf{V}_{p,aa}^{(p)}$ is a $(p \times p)$ matrix. In (3.8), the normalizing constant, $\kappa_{\mathbf{a}, r}(\cdot)$, is computed as follows,

$$\kappa_{\mathbf{a}, r}(\mathbf{V}_p^{(p)}, \nu_p^{(p)}) = \Gamma \left(\frac{\nu_p^{(p)}}{2} \right) (\lambda_p^{(p)})^{-\frac{\nu_p^{(p)}}{2}} |\mathbf{V}_{p,aa}^{(p)}|^{-\frac{1}{2}} (2\pi)^{\frac{1}{2}}, \quad (3.10)$$

where, the reminder, $\lambda_p^{(p)}$, is,

$$\lambda_p^{(p)} = v_{p,11}^{(p)} - \mathbf{v}_{p,a1}^{(p)T} (\mathbf{V}_{p,aa}^{(p)})^{-1} \mathbf{v}_{p,a1}^{(p)}. \quad (3.11)$$

Using Bayes' rule (1.9), the posterior pd is attained via substituting (3.5) and (3.8) into (1.9) as follows:

$$F(\mathbf{a}^{(p)}, r|\mathbf{x}_n, p) \propto F(x_n|\psi_n^{(p)}, \mathbf{a}^{(p)}, r, p) F(\mathbf{a}^{(p)}, r|p) \quad (3.12)$$

$$= NiG_{\mathbf{a}^{(p)},r}(\mathbf{V}_n^{(p)}, \nu_n^{(p)}). \quad (3.13)$$

Notably, the posterior pd (3.13) is uniquely determined by *sufficient statistics*, $\mathbf{s}_n^{(p)} \equiv (\mathbf{V}_n^{(p)}, \nu_n^{(p)})$, and the functional recursion (3.12). Therefore, the posterior pd (3.13) are updated via time recursions on sufficient statistics, i.e.,

$$\mathbf{s}_i^{(p)} = \mathbf{s}(\mathbf{s}_{i-1}^{(p)}, x_i) \quad i = p+1, \dots, n \quad (3.14)$$

In particular, the counter and accumulator are respectively updated in two scenarios as follows,

$$\nu_n^{(p)} = \nu_p^{(p)} + (n - p) \quad (\text{off-line}) \quad (3.15)$$

$$= \nu_{n-1}^{(p)} + 1 \quad (\text{on-line}) \quad (3.16)$$

$$\mathbf{V}_n^{(p)} = \mathbf{V}_p^{(p)} + \sum_{i=p+1}^n \varphi_i^{(p)} \varphi_i^{(p)T} \quad (\text{off-line}) \quad (3.17)$$

$$= \mathbf{V}_{n-1}^{(p)} + \varphi_n^{(p)} \varphi_n^{(p)T}, \quad (\text{on-line}) \quad (3.18)$$

where, $\varphi_n^{(p)}$ is defined in notational conventions. Also (3.15) and (3.17) express off-line recursions and (3.16), (3.18) express on-line recursions [61], respectively. Notice that $\mathbf{V}_n^{(p)}$ has the same partitioning (3.9), i.e.

$$\mathbf{V}_n^{(p)} = \begin{bmatrix} v_{n,11}^{(p)} & \vdots & \mathbf{v}_{n,a1}^{(p)T} \\ \dots\dots\dots & & \\ \mathbf{v}_{n,a1}^{(p)} & \vdots & \mathbf{V}_{n,aa}^{(p)} \end{bmatrix}, \quad (3.19)$$

Remark 3.3.2. (*Diffuse prior*)

Whenever we do not have any preferences, we use non-informative prior pd diffusing shaping parameters e.g. $\mathbf{V}_p^{(p)} = \rho I_{p+1}$, $\nu_p^{(p)} = \rho$ where I_{p+1} is the $(p+1) \times (p+1)$ identity matrix and ρ is a small positive scalar. $\square$

Remark 3.3.3. (*Evidence*)

The normalizing constant, $\kappa_{\mathbf{a},r}(\cdot)$, is a projection of data and obtained by marginalizing $F(\mathbf{a}^{(p)}, r, \mathbf{x}_n|p)$ over global parameters, $(\mathbf{a}^{(p)}, r)$. It is referred as ‘evidence’, ‘marginalized observation model’ or ‘marginal likelihood’ in the literature i.e. [86].

ization over $\mathbf{a}_{\delta'}^{(p)}$, the p -conditional marginal posterior pd of $(\mathbf{a}_{\delta}^{(p)}, r)$, is,

$$F(\mathbf{a}_{\delta}^{(p)}, r | \mathbf{x}_n, p) = \text{NiG}_{\mathbf{a}_{\delta}, r}(\mathbf{V}_{n, \delta}^{(p)}, \nu_n^{(p)}), \quad (3.26)$$

where $\mathbf{V}_{n, \delta}^{(p)}$ is the Bayesian multivariate reminder and it is computed,

$$\mathbf{V}_{n, \delta}^{(p)} \equiv \mathbf{V}_{n, 1\delta}^{(p)} - \mathbf{V}_{n, \delta'\delta}^{(p)T} (\mathbf{V}_{n, \delta'\delta'}^{(p)})^{-1} \mathbf{V}_{n, \delta'\delta}^{(p)}. \quad (3.27)$$

Notably, shaping parameters of marginal NiG (3.26) is a transformation on original shaping parameters, $(\mathbf{V}_n^{(p)}, \nu_n^{(p)})$. Notice that after necessary mathematical operations on $\mathbf{V}_n^{(p)}$, the index, δ , enters shaping parameters of marginal NiG (3.26) via $\mathbf{V}_{n, \delta}^{(p)}$.

- **The inference of a subset of coefficients:** using (3.21) and marginalization over $\mathbf{a}_{\delta'}^{(p)}$, the marginal pd of $\mathbf{a}_{\delta}^{(p)}$ is,

$$F(\mathbf{a}_{\delta}^{(p)} | \mathbf{x}_n, p) = \text{St}_{\mathbf{a}_{\delta}^{(p)}}(\mu_{n, \delta}^{(p)}, \tau_{n, \delta}^{(p)}, \eta_{n, \delta}^{(p)}), \quad (3.28)$$

where shaping parameters are computed based on shaping parameters of marginal NiG in (3.26).

Notably, the shaping parameters of iG is a projection of shaping parameters of marginal NiG in (3.26) which is invariant w.r.t. any choice of δ .

Remark 3.3.5. (Bayesian version of Multivariate LS reminder)

- Noting that $\mathbf{V}_{n, \delta}^{(p)} \equiv \mathbf{\Lambda}_{n, \delta}^{(p)}$ is the Bayesian version of multivariate LS reminder, where $\mathbf{\Lambda}_{n, \delta}^{(p)} \in \mathbb{R}^{(|\delta|+1) \times (|\delta|+1)}$ is the multivariate form of $\lambda_n^{(p)}$ in (3.11) at time n . We note that it corresponds to LS reminder underlying diffuse priors $\nu_p^{(p)} \ll \nu_n^{(p)}$, and $\mathbf{V}_p^{(p)} \ll \mathbf{V}_n^{(p)}$.
- Although $\mathbf{V}_{n, \delta}^{(p)}$ depends on δ , marginalization of $F(\mathbf{a}_{\delta}^{(p)}, r | \mathbf{x}_n, p)$ (3.26) over $\mathbf{a}_{\delta}^{(p)}$, yields $iG_r(\cdot)$ (3.25), for any choice of δ . It is obvious that $iG_r(\cdot)$ (3.25) does not depend on δ which requires for consistency.

Remark 3.3.6. (Bayesian point estimation)

- Bayesian point estimation of $\mathbf{a}^{(p)}$ and its variance-covariance matrix are attained as follows:

$$\hat{\mathbf{a}}_n^{(p)} = E_{F(\mathbf{a}^{(p)} | \mathbf{x}_n, p)}[\mathbf{a}^{(p)} | \mathbf{V}_n^{(p)}, \nu_n^{(p)}] = \mu_n^{(p)} \quad (3.29)$$

$$Cov_{F(\mathbf{a}^{(p)}|\mathbf{x}_n, p)}[\mathbf{a}^{(p)}|\mathbf{V}_n^{(p)}, \nu_n^{(p)}] = \frac{(\nu_n^{(p)} - p + 2)^2}{\nu_n^{(p)} - p} \lambda_n^{(p)} \mathbf{V}_{n,aa}^{(p)}. \quad (3.30)$$

- Bayesian point estimation of r and its variance are attained as follows

$$\hat{r} = E_{F(r|\mathbf{x}_n, p)}[r|\mathbf{V}_n^{(p)}, \nu_n^{(p)}] = \frac{\lambda_n^{(p)}}{\nu_n^{(p)} - p - 3} \quad (3.31)$$

$$Var_{F(r|\mathbf{x}_n, p)}[r|\mathbf{V}_n^{(p)}, \nu_n^{(p)}] = \frac{(\lambda_n^{(p)})^2}{(\nu_n^{(p)} - p - 3)^2(\nu_n^{(p)} - p - 4)}. \quad (3.32)$$

- For any subset $\mathbf{a}_\delta^{(p)}$,

$$Cov_{F(\mathbf{a}^{(p)}, r|\mathbf{x}_n, p)}(\mathbf{a}_\delta^{(p)}, r) = 0.$$

3.3.2 p -conditional one-step-ahead predictor

We here aim to compute the tractable Bayesian p -conditional *one-step-ahead* (OSA) predictor. This is computed as follows:

$$\begin{aligned} F(x_{n+1}|\mathbf{x}_n, p) &= \int_r \int_{\mathbf{a}^{(p)}} F(\mathbf{x}_{n+1}, \mathbf{a}^{(p)}, r|p) d\mathbf{a}^{(p)} dr \\ &= \int_r \int_{\mathbf{a}^{(p)}} F(x_{n+1}|\mathbf{x}_n, \mathbf{a}^{(p)}, r, p) F(\mathbf{a}^{(p)}, r|\mathbf{x}_n, p) d\mathbf{a}^{(p)} dr \\ &= \frac{F(\mathbf{x}_{n+1}|p)}{F(\mathbf{x}_n|p)} \end{aligned} \quad (3.33)$$

Using Remark 3.3.3,

$$\begin{aligned} F(x_{n+1}|\mathbf{x}_n, p) &= \frac{\kappa_{\mathbf{a},r}(\mathbf{V}_{n+1}^{(p)}, \nu_{n+1}^{(p)})}{\sqrt{2\pi} \cdot \kappa_{\mathbf{a},r}(\mathbf{V}_n^{(p)}, \nu_n^{(p)})} \\ &= St\left(\hat{\mathbf{a}}^{(p)T} \psi_{n+1}^{(p)}, \frac{\nu_n^{(p)}}{\lambda_n^{(p)} (1 + \psi_{n+1}^{(p)T} \mathbf{V}_{n,aa}^{(-1)} \psi_{n+1}^{(p)})}, \nu_n^{(p)} + 1\right) \end{aligned} \quad (3.34)$$

The proof of (3.34) is presented in Appendix A.4. Moreover, the expected value and variance of Bayesian OSA predictor are respectively:

$$E(x_{n+1}|\mathbf{x}_n, p) = \hat{\mathbf{a}}^{(p)T} \psi_{n+1}^{(p)}, \quad (3.35)$$

$$Var(x_{n+1}|\mathbf{x}_n, p) = \frac{\lambda_n^{(p)}(1 + \psi_{n+1}^{(p)T} \mathbf{V}_{n,aa}^{(-1)} \psi_{n+1}^{(p)})}{\nu_n^{(p)} - 2} \quad (3.36)$$

3.3.3 p -conditional multi-step-ahead predictor

The aim here is to obtain Bayesian p -conditional multi-step-ahead (MSA) predictor. For this reason, we wish to compute MSA predictor for the next k steps, i.e. $F(x_{n+k}|\mathbf{x}_n, p)$. Let $k > 1$, using OSA predictor (3.34), the block k -step-ahead predictor is the product of the scalar Student's- t pd,

$$F(x_{n+1}, \dots, x_{n+k}|\mathbf{x}_n, p) = \prod_{i=1}^k F(x_{n+i}|\mathbf{x}_{n+i-1}, p) \quad (3.37)$$

We note that (3.37) is not multivariate Student's- t pd. The reason is that the next Student's- t pd is highly non-linear conditioning on the earlier data (which are unknown) and a standard form of $F(x_{n+i}|\mathbf{x}_n, p)$, $i = 2, 3, \dots$ cannot be attained. However, we can exploit the product of Student's- t structure (3.37) in the context of an exact sequential sampling. Therefore we use generated/artificial datum to augment data to obtain shaping parameters of each scalar Student's- t pd (3.37), using (3.34). Once we attain exact k -dimensional sample from structure (3.37), i^{th} - sample ($i = 1, \dots, k$) of the block is associated to i^{th} -step-ahead predictor. Notably, this way proposes a simpler, elegant and fundamental sampling method for k -step ahead prediction. There are several methods available in the literature. For instance, [89] exploits one-step-ahead predicted value as a datum for the next step to built up a multi-step-ahead prediction which is quite different from our proposal (3.37). Also, [90] deduces an approximate k -step-ahead expectation by mixing the empirical approximation of the posterior inference of $\mathbf{a}^{(p)}$ as follows,

$$E[\psi_{n+k}^{(p)}|\mathbf{x}_n] \approx \frac{1}{N} \sum_{i=1}^N E[\psi_{n+k}^{(p)}|(\mathbf{a}^{(p)}, r)^{(i)}, \mathbf{x}_n]. \quad (3.38)$$

The MSA predictor (3.38) [90] is completely different from our proposal which does not construct any pd in favour of constructing the first moment. In against, our approach does not require sampling from $\mathbf{a}^{(p)}$ since we exactly marginalise w.r.t. the parameters $(\mathbf{a}^{(p)}, r)$. Although we then simulate from the exact block k -step-ahead predictor.

Remark 3.3.7. (*Bayesian computation*)

Notice that LDL factorization is widely used for Bayesian computation in literature i.e. [55] and [91]. This decomposition is mainly exploited for computational reason and numerical stability.

3.4 Time-invariant parametrization of the AR model: unknown model order

Here we relax the assumption of known model order allowing to be $p = 1, \dots, p_u$. The discrete non-uniform prior pd of p is adopted as

$$Pr\{L = p\} \equiv \alpha_0^{(p)}, p = 1, \dots, p_u. \quad (3.39)$$

where $\alpha_0^{(p)}$ are known a priori. We note that in the case where $\alpha_0^{(p)}$ is unknown, the Dirichlet hyper prior pd can be adopted [61]. Since the global inference of unknowns $(\mathbf{a}, r)$ is a special case of an arbitrary subset inference of unknowns, we here particularly interested in the inference of a subset of $(\mathbf{a}, r)$ in the case where p is unknown. This involves a posteriori marginalization over p that leads to a finite mixture model in all cases. In general, components wights will be the posterior model order inference which is derived as,

$$\begin{aligned} \alpha_n^{(p)} &= \Pr\{L = p | \mathbf{x}_n\} \\ &= \int_{\mathbf{a}} \int_r F(\mathbf{a}, r, p | \mathbf{x}_n) dr d\mathbf{a} \end{aligned} \quad (3.40)$$

Using Remark 3.3.3, (3.40) becomes,

$$\alpha_n^{(p)} = \frac{\kappa_{\mathbf{a},r}(\mathbf{V}_n^{(p)}, \nu_n^{(p)})}{\sum_{d=1}^{p_u} \kappa_{\mathbf{a},r}(\mathbf{V}_n^{(d)}, \nu_n^{(d)})}, p = 1, \dots, p_u, \quad (3.41)$$

Notably (3.41) is also seen as the normalizing normalizing constant of $NiG(\mathbf{V}_n^{(p)}, \nu_n^{(p)})$.

Remark 3.4.1. (*Model order inference using automatic Ockham's regularization*)

Based on posterior model order (3.41), we can define a criterion i.e. Bayesian posterior order (BPO) as follows,

$$\begin{aligned} BPO(p) &= -2 \log [\alpha_n^{(p)}] \\ &= -2 \log [\kappa_{a,r}(\mathbf{V}_n^{(p)}, \nu_n^{(p)})] + 2 \log \left[\sum_{d=1}^{p_u} \kappa_{a,r}(\mathbf{V}_n^{(d)}, \nu_n^{(d)}) \right] \end{aligned} \quad (3.42)$$

Since the second term in (3.42) is constant, BPO reduces to,

$$\begin{aligned} BPO(p) &= -2 \log [\kappa_{a,r}(\mathbf{V}_n^{(p)}, \nu_n^{(p)})] \\ &= \nu_n^{(p)} \log [\lambda_n^{(p)}] - 2 \log \left[\Gamma \left(\frac{\nu_n^{(p)}}{2} \right) \right] + \log [|\mathbf{V}_{n,aa}^{(p)}|], \end{aligned} \quad (3.43)$$

where, the first term represents data fidelity and the second and third terms together are Ockham's term. We note that BPO engenders Ockham's term automatically due to marginalization over $(\mathbf{a}^{(p)}, r)$ and it does not require any extra operations. $\square$

It will be necessary also to evaluate the conditional model order inference which is derived:

$$\begin{aligned} \beta_n^{(p)} &= Pr(L = p | L \geq p_0, \mathbf{x}_n) \\ &= \frac{\kappa_{a,r}(\mathbf{V}_n^{(p)}, \nu_n^{(p)})}{\sum_{d=p_0}^{p_u} \kappa_{a,r}(\mathbf{V}_n^{(d)}, \nu_n^{(d)})}, \quad p = p_0, \dots, p_u. \end{aligned} \quad (3.44)$$

Recall that we here aim to derive posterior pd of a subset of parameters e.g. $\mathbf{a}_\delta$ where p is unknown. Let $\mathbf{a}_\delta = [a_{k_1}, \dots, a_{k_q}]^T$ be a subset of $\mathbf{a}^{(p_u)}$, with $\delta = \{k_1, \dots, k_q\}$ as a selection of q distinct indexes from $\{1, \dots, p_u\}$ and $p_0 (\leq p_u)$ is the biggest index in δ . We will adopt the following assumptions:

1. **Availability:** For inference of $\mathbf{a}_\delta$, we assume that the indices of all elements of $\mathbf{a}_\delta$ are known.
2. **Nesting:** we will assume $\mathbf{a}^{(p-1)} \subset \mathbf{a}^{(p)}$, where $\mathbf{a}^{(p)} \equiv (\mathbf{a}^{(p-1)}, a_p)$, and $r^{(j)} = r, \forall j = 1, \dots, p_u$. In the other words, we are adopting a nested model.

3. **Consistency:** we assume all propositions involving $\mathbf{a}_\delta$ and r are conditioned on $L \geq p_0$.

Underlying assumptions 1-3 and using (3.26), the conditional joint pd of $(\mathbf{a}_\delta, r)$ is:

$$\begin{aligned} F(\mathbf{a}_\delta, r | L \geq p_0, \mathbf{x}_n) &= \sum_{p=p_0}^{p_u} \beta_n^{(p)} \cdot F(\mathbf{a}_\delta, r | L = p, \mathbf{x}_n) \\ &= \sum_{p=p_0}^{p_u} \beta_n^{(p)} \cdot \text{NiG}_{\mathbf{a}_\delta, r}(\mathbf{V}_{n, \delta}^{(p)}, \nu_n^{(p)}). \end{aligned} \quad (3.45)$$

Notably using (3.26) and (3.44), therefore (3.45) is a finite mixture of marginal NiG with weights, $\beta_n^{(p)}$, given in (3.44).

Remark 3.4.2. (Marginal pds with p unknown)

- The inference of $\mathbf{a}_\delta$ is the mixture of Student's-t pd with components (3.28) and weights $\beta_n^{(p)}$ (3.44).
- Since r is shared by p_u candidate models, we use unconditional model order inference $\alpha_n^{(p)}$ (3.41) for marginal inference of r as follows:

$$F(r | \mathbf{x}_n) = \sum_{p=1}^{p_u} \alpha_n^{(p)} iG_r \left(\frac{\nu_n^{(p)} - p - 2}{2}, \frac{\lambda_n^{(p)}}{2} \right). \quad (3.46)$$

Remark 3.4.3. (Prediction and point estimation with unknown p)

- The OSA predictive pd is:

$$\begin{aligned} F(x_{n+1} | \mathbf{x}_n) &= \sum_{p=1}^{p_u} \alpha_n^{(p)} F(x_{n+1} | \mathbf{x}_n, p) \\ &= \sum_{p=1}^{p_u} \alpha_n^{(p)} \mathcal{St} \left(\hat{\mathbf{a}}^{(p)T} \psi_{n+1}^{(p)}, \frac{\nu_n^{(p)}}{\lambda_n^{(p)} (1 + \psi_{n+1}^{(p)T} \mathbf{V}_{n, aa}^{(-1)} \psi_{n+1}^{(p)})}, \nu_n \right), \end{aligned} \quad (3.47)$$

where $\alpha_n^{(p)}$ is given in (3.41). Also MSA predictor is the mixture of the p -conditional MSA predictor (3.37) with weight $\alpha_n^{(p)}$.

- The moments in the case of unknown p , are a linear combination of p -conditional case with associated weights.

□

3.5 Classical data processing using the AR model: known model order

Here, we review some of the established classical approaches for processing data using the AR model. Notice that we refer to non-Bayesian methods as classical approaches. In these approaches, the existence of prior pd is not explicitly assumed. Classical approaches usually design the statistics non-prescriptively, then they investigate the property of statistics. Here we briefly review some key concepts in classical statistical signal processing. Let x_n be a random process. This random process is called to be *strict sense stationary* (SSS) if its finite order pds are time invariant, i.e.,

$$F(x_{n_1+i}, \dots, x_{n_k+i}) = F(x_{n_1}, \dots, x_{n_k}), \quad \forall k, i, \text{ and } n_1, \dots, n_k \quad (3.48)$$

Also, x_n , is said to be *wide sense stationary* (WSS) if its mean and autocorrelation functions are time invariant, i.e.,

$$E(x_{n+i}) = \mu_n = E(x_n) \quad (3.49)$$

$$\phi(n, n+i) = E(x_n x_{n+i}) = \phi(i) \quad (3.50)$$

More details on properties of moments (3.49) and (3.50) are provided in [14]. Recall that the classical approaches directly process the accumulator matrices which is related to EIM in Bayesian paradigm. We freshly investigate the core matrices of classical approaches and connect them to EIM, in the AR model. For this reason, we firstly need to define the two important variants of the data cross-product matrices.

Outer product matrix: one of the key variants in the classical approach for the AR model is the accumulator of $(n-p)$ outer products, $\mathbf{C}_n^{(p)}$ — which we refer as outer product matrix. This matrix enjoys the same partitioning as $\mathbf{V}_n^{(p)}$ (3.20) with a scalar $c_{n,11}^{(p)}$,

$$c_{n,11}^{(p)} = \mathbf{x}_n^{(p)T} \mathbf{x}_n^{(p)}$$

$$= \sum_{i=p+1}^n x_i^2, \quad (3.51)$$

and $\mathbf{c}_{n,a1}^{(p)} = \Psi_n^{(p)T} \mathbf{x}_n^{(p)}$ is a $(p \times 1)$ vector with elements,

$$(\mathbf{c}_{n,a1}^{(p)})_l = \sum_{i=p+1}^n x_i x_{i-l} \quad l = 1, \dots, p \quad (3.52)$$

and $\mathbf{C}_{n,aa}^{(p)} = \Psi_n^{(p)T} \Psi_n^{(p)}$ is a $(p \times p)$ full-rank, symmetric and invertible matrix with elements

$$\begin{aligned} (\mathbf{C}_{n,aa}^{(p)})_{kl} &= (\Psi_n^{(p)T} \Psi_n^{(p)})_{kl} \\ &= \mathbf{g}_{n,k}^T \mathbf{g}_{n,l} \\ &= \sum_{i=p+1}^n x_{i-k} x_{i-l} \quad k, l = 1 \dots p \end{aligned} \quad (3.53)$$

Sample auto-correlation matrix: the other important variant is being sample auto-correlation matrix, with the same partitioning as $\mathbf{V}_n^{(p)}$ (3.20), as follows,

$$\mathbf{D}_n^{(p)} = \begin{bmatrix} d_0 & d_1 & \cdots & d_p \\ d_1 & d_0 & \cdots & d_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ d_p & d_{p-1} & \cdots & d_0 \end{bmatrix}, \quad (3.54)$$

which is a deterministic, Toeplitz-symmetric matrix with descending diagonal from left to right, with elements,

$$\begin{aligned} d_l &= x_l * x_{-l} \\ &= \sum_{i=l+1}^n x_i x_{i-l}, \quad l = 0, \dots, p. \end{aligned} \quad (3.55)$$

Remark 3.5.1. (*Relationship between classical variants with EIM*)

- Based on Bernstein-von Mises theorem [92, 93], and under diffuse prior i.e. $\mathbf{V}_p^{(p)} \ll$

$V_n^{(p)}$, EIM (3.20) becomes $C_n^{(p)}$,

$$V_n^{(p)} = C_n^{(p)}. \quad (3.56)$$

Furthermore, under ergodicity assumption and $n \gg p$,

$$V_n^{(p)} = D_n^{(p)}, \quad (\text{in prob.}) \quad (3.57)$$

where ‘in prob.’ denotes convergence in probability expressed via Law of Large Numbers (LLN).

- In practice, the matrix (3.54) is widely used for computational savings where the Levinson algorithm i.e. [94], [95] is applied.

3.5.1 Technical assumptions for classical strategies

In this section, we aim to provide the technical assumptions which are exploited in the classical methods in order to estimate unknown parameters of the AR model. The most popular classical strategies are *Maximum likelihood*(ML), *least squared*(LS) and *mean squared*(MS) (i.e. Wiener) approaches. These strategies are based on different modelling assumptions as follows.

ML strategy: in this strategy, the parametric observation model (3.4) is adopted to Wold structure (3.3). But no prior pd is explicitly assumed.

LS strategy: this strategy involves a non-parametric model in which observation is structured via (3.3).

MS strategy: this strategy adopts a non-parametric model to structure (3.3) in which e_i is assumed to be real-finite valued, zero-mean, wide-sense stationary (WSS) random process. Underlying MS modelling assumptions, the first two moments can be defined as:

$$E(x_n) = g_1(\mathbf{a}^{(p)}, r) \quad (3.58)$$

$$\begin{aligned} E(x_n x_{n-l}) &= \phi(l), \quad l = 0, \dots, p \\ &= g_2(\mathbf{a}^{(p)}, r) \end{aligned} \quad (3.59)$$

Notice that (3.58) and (3.59) are a function of unknowns and therefore they are unknown as well. Furthermore, using the auto-correlation sequence (3.59), the Toeplitz auto-correlation matrix (unknown) can be defined as:

$$\Phi^{(p)} = \begin{bmatrix} \phi(0) & \phi(1) & \cdots & \phi(p) \\ \phi(1) & \phi(0) & \cdots & \phi(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(p) & \phi(p-1) & \cdots & \phi(0) \end{bmatrix} \quad (3.60)$$

Notably $\Phi^{(p)}$ (3.60) is unknown matrix which enjoys the same partitioning as $V_n^{(p)}$ in (3.20).

3.5.2 p -conditional parameter estimation strategies

In this section, we provide classical estimation strategies for unknown parameters of the AR model in the case where p is unknown. Underlying ML modelling assumptions, the ML estimate of $\mathbf{a}^{(p)}$, is:

$$\mathbf{a}_{ML}^{(p)} = \arg \max_{\mathbf{a}^{(p)}} \mathcal{L}(\mathbf{a}^{(p)}, r | \mathbf{x}_n, p), \quad (3.61)$$

where $\mathcal{L}(\cdot)$ is the likelihood function and it equals to observation model (3.5). Notice that the computation of $\mathbf{a}_{ML}^{(p)}$ for the AR model is not feasible in practice [96], since it requires the first p data. A practical alternative is LS approach in which $\mathbf{a}^{(p)}$ is estimated as,

$$\hat{\mathbf{a}}_{LS}^{(p)} = \underset{\mathbf{a}^{(p)}}{\operatorname{argmin}} \|\mathbf{x}_n^{(p)} - \Psi_n^{(p)T} \mathbf{a}^{(p)}\|^2, \quad (3.62)$$

where $\|\cdot\|$ is Euclidean norm. It immediately follows,

$$\mathbf{C}_{n,aa}^{(p)} \hat{\mathbf{a}}_{n,LS}^{(p)} = \mathbf{C}_{n,a1}^{(p)}, \quad (3.63)$$

$$\hat{\mathbf{a}}_{n,LS}^{(p)} = \left(\mathbf{C}_{n,aa}^{(p)} \right)^{-1} \mathbf{C}_{n,a1}^{(p)}. \quad (3.64)$$

We note that LS estimation for a subset $\mathbf{a}_\delta^{(p)}$ —denoted by $\hat{\mathbf{a}}_{\delta,LS}^{(p)}$ is the associated subset selection of global estimation $\hat{\mathbf{a}}_{LS}^{(p)}$. It is demonstrated that $\hat{\mathbf{a}}_{\delta,LS}^{(p)}$ is consistent i.e. [97].

Underlying MS modelling assumptions, the MS estimate of $\mathbf{a}^{(p)}$ is,

$$\hat{\mathbf{a}}_{n,MS}^{(p)} = \arg \min_{\mathbf{a}^{(p)}} E[(x_i - \psi_i^T \mathbf{a}^{(p)})^2]. \quad (3.65)$$

Therefore, $\hat{\mathbf{a}}_{MS}^{(p)}$ is attained by solving normal equations,

$$\hat{\mathbf{a}}_{MS}^{(p)} = (\Phi_{aa}^{(p)})^{-1} \phi_{a1}^{(p)}, \quad (3.66)$$

We note that, $\hat{\mathbf{a}}_{MS}^{(p)}$, (3.66) is not statistics. Since it is a unknown function of autocorrelation sequence $\phi(\cdot)$, which involves unknown parameters $(\mathbf{a}^{(p)}, r)$. Therefore, using (3.54), the estimate of unknown autocorrelation matrix, $\Phi^{(p)}$, is:

$$\hat{\Phi}^{(p)} = \frac{1}{n-p} D_n^{(p)}, \quad (3.67)$$

substituting (3.67) into (3.66), the MMSE estimate of $\mathbf{a}^{(p)}$ is asserted [98]. Notably, underlying MS modelling assumptions and LLN, for large n ,

$$\hat{\Phi}_n^{(p)} \rightarrow \Phi^{(p)}, \quad (\text{in prob.}) \quad (3.68)$$

Remark 3.5.2. (Classical estimate of r)

- *ML estimate of r —conditioning on accessibility of the first p data— can be attained by substituting $\hat{\mathbf{a}}_{ML}^{(p)}$ into likelihood function, in step-wise sense, as follows,*

$$\begin{aligned} \hat{r}_{n,ML}(p) &= \arg \max_r \mathcal{L}(r | \hat{\mathbf{a}}_{n,ML}^{(p)}, \mathbf{x}_n, p) \\ &= \frac{1}{n-p} [-1, \hat{\mathbf{a}}_{n,ML}^{(p)}] \left\{ \sum_{i=p+1}^n \varphi_i^{(p)} \varphi_i^{(p)T} \right\} [-1, \hat{\mathbf{a}}_{n,ML}^{(p)}]^T \end{aligned} \quad (3.69)$$

- *Although, observation variance, r , is not available in LS and MS approaches, but related problem provides a formula using LS approach and MS criteria are respectively as follows,*

$$\begin{aligned}
\hat{r}_{n,LS}(p) &\equiv \|\hat{e}_n^{(p)}\|_{LS}^2 \\
&= \|\mathbf{x}_n^{(p)} - \Psi_n^{(p)T} \hat{\mathbf{a}}_{LS}^{(p)}\|^2
\end{aligned} \tag{3.70}$$

$$\hat{r}_{n,MS}(p) \equiv \hat{\phi}_n(0) - \sum_{j=1}^p (\hat{\mathbf{a}}_{n,MS}^{(p)})_j \hat{\phi}_n(j), \tag{3.71}$$

where $(\hat{\mathbf{a}}_{n,MS}^{(p)})_j$ is the j^{th} - element of $\hat{\mathbf{a}}_{n,MS}^{(p)}$.

In the following, we technically summarize the interrelationships between Bayesian point estimates (Remark 3.3.6) and the above classical criteria.

Remark 3.5.3. (*Relationship with Bayesian point estimates*)

- *LS estimate: underlying Bernstein-von Mises theorem, adopting a diffuse prior NiG for $(\mathbf{a}^{(p)}, r)$, i.e. $\nu_p^{(p)} \ll \nu_n^{(p)}$, and $V_p^{(p)} \ll V_n^{(p)}$,*

$$\hat{\mathbf{a}}_{n,LS}^{(p)} = \hat{\mathbf{a}}_n^{(p)} \tag{3.72}$$

$$\hat{r}_{n,LS} = \lambda_n^{(p)}. \tag{3.73}$$

- *MS estimate: underlying MS modelling assumption, and using LLN for large n ,*

$$\hat{\mathbf{a}}_{n,MS}^{(p)} = \hat{\mathbf{a}}_n^{(p)} \quad (\text{in prob.}) \tag{3.74}$$

$$\hat{r}_{n,MS}(p) = \lambda_n^{(p)} \quad (\text{in prob.}).$$

- *Under regularity conditions [99], posterior pd (3.13) converges in probability to the Dirac's delta function, $\delta[\cdot]$, as follows,*

$$\lim_{n \rightarrow \infty} F(\mathbf{a}^{(p)}, r | \mathbf{x}_n, p) = \delta[(\mathbf{a}^{(p)}, r) - (\hat{\mathbf{a}}^{(p)}, \hat{r})_\infty] \tag{3.75}$$

where $(\hat{\mathbf{a}}^{(p)}, \hat{r})_\infty = \lim_{n \rightarrow \infty} (\hat{\mathbf{a}}_n^{(p)}, \hat{r}_n)$. Notice that $\hat{\mathbf{a}}_n^{(p)}$ and $\hat{r}_n$ can be substituted by related LS and MS estimates.

3.5.3 p -conditional prediction

In this section, a technical review of classical prediction strategies from the Bayesian perspective is provided. For this reason, let,

$$\begin{aligned} h_n \equiv h(\mathbf{x}_n) &= g(\hat{\mathbf{a}}_n^{(p)}, \hat{r}_n) \\ &= \hat{\mathbf{a}}_n^{(p)T} \psi_{n+1}^{(p)}, \end{aligned} \quad (3.76)$$

the classical OSA prediction, x_{n+1} , is computed,

$$\begin{aligned} F(x_{n+1}|\mathbf{x}_n, p) &\approx F(x_{n+1}|h_n, p) \\ &= \delta[x_{n+1} - \hat{\mathbf{a}}^{(p)T} \psi_{n+1}^{(p)}]. \end{aligned} \quad (3.77)$$

This strategy exploits *certainty equivalent* (CE) technique (2.35), in which the uncertainty about $(\mathbf{a}^{(p)}, r)$ is replaced by their estimates. Furthermore, k -step ahead prediction in the classical approach i.e. [89] is attained sequentially as follows,

$$\hat{x}_{n+k} = \hat{\mathbf{a}}_{n+k-1}^{(p)T} \hat{\psi}_{n+k}^{(p)}, \quad (3.78)$$

where $\hat{\psi}_{n+k}^{(p)} = [x_{n+k-p}, \dots, x_n, \hat{x}_{n+1}^{(p)}, \dots, \hat{x}_{n+k-1}^{(p)}]^T, p \geq k$.

More details of classical MSA prediction are provided in [47].

Remark 3.5.4. (*Drawbacks of classical prediction*)

- The main drawback of classical OSA predictor (3.77) is that the actual uncertainty about $(\mathbf{a}^{(p)}, r)$ is not accounted for, i.e. it is adopted the following approximation,

$$F(\mathbf{a}^{(p)}, r|\mathbf{x}_n, p) \approx \delta[(\mathbf{a}^{(p)}, r) - (\hat{\mathbf{a}}_n^{(p)}, \hat{r}_n)], \quad n < \infty. \quad (3.79)$$

Equation (3.79) is true if and only if h_n is sufficient statistics for inference of $(\mathbf{a}^{(p)}, r)$.

- The Bayesian OSA predictor (3.33) and the classical prediction strategies (3.77) converges only as n increases.

3.5.4 Classical data processing using the AR model: unknown model order

In this section, we study classical estimation and prediction strategies in the AR model with unknown p . For this reason, we focus on LS strategy. We note that in the case where p is unknown, this strategy usually exploits CE technique in a stage wise sense. In particular, it involves model order estimation in order to estimate the unknown parameters. We note that LS strategy (i.e. $\hat{r}_{LS}(p)$ (3.70)) does not have any term to control complexity which leads to the over-fitting problem. Therefore, it requires an external term—which we refer as *Ockham's term* (i.e. penalty term) to solve this issue. For this reason, various criteria have been proposed which will be shortly presented.

3.5.5 Model order estimation strategies

Beyond the Bayesian solution (3.43), various methods of model order estimation are proposed to estimate p in the literature i.e. [100]. Universally they are an objective function of order p with an additive structure as follows,

$$C(p) = C_d(p) + C_o(p), \quad (3.80)$$

where $C_d(p) \equiv \hat{r}_{LS}(p)$, is a non-increasing function of p (i.e. data fidelity term) and $C_o(p)$ is an non-decreasing function of p (i.e. Ockham's term). The Ockham's term has to be performed to avoid this problem in CE-based estimation strategies. In the following, we review the principle beyond different methods and specialize them to our AR case.

3.5.5.1 Principles of model comparison and specialization into AR case

We here consider the most popular criteria of model order estimate involving Akaike's Information Criterion (AIC) [101], Bayesian Information Criterion (BIC) [102], Hannan and Quinn (HQ) [103], Minimum Description Length (MDL) [104] and Minimum Message Length (MML) [96]. Notice that these follows the additive structure (3.80) criteria. These principles are reviewed for the general case as follows:

BIC (Oliver and Baxter version, 1994): in BIC, the prior pd on all unknowns $(\boldsymbol{\theta}, p)$, is adopted. Using the Bayes' rule:

$$\begin{aligned} F(\boldsymbol{\theta}^{(p)}, p | \mathbf{x}_n) &= \frac{F(p)F(\boldsymbol{\theta}^{(p)}|p)F(\mathbf{x}_n|\boldsymbol{\theta}^{(p)}, p)}{\kappa(\mathbf{x}_n)} \\ &= \frac{F(p)F(\boldsymbol{\theta}^{(p)}|p)\mathcal{L}(\boldsymbol{\theta}^{(p)}|p, \mathbf{x}_n)}{\kappa(\mathbf{x}_n)}, \end{aligned} \quad (3.81)$$

where p here is the number of free parameters in the model. BIC adopts the discrete uniform prior of p which classifies the unknown space $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ into p_u local parameter spaces, $\boldsymbol{\theta}^{(p)} \in \boldsymbol{\Theta}^{(p)}$, $p = 1, \dots, p_u$. It also adopts a uniform prior on $\boldsymbol{\theta}^{(p)}$ i.e. $F(\boldsymbol{\theta}^{(p)}|p) = 1$. Adopting these priors to (3.81), eliminating constant terms w.r.t. p and taking a second-order Taylor series expansion of $\mathcal{L}(\boldsymbol{\theta}^{(p)}|p, \mathbf{x}_n)$ about $\boldsymbol{\theta}_{MAP}^{(p)}$,

$$F(\boldsymbol{\theta}^{(p)}, p | \mathbf{x}_n) \approx \mathcal{L}(\boldsymbol{\theta}_{MAP}^{(p)}|p, \mathbf{x}_n) \exp \left\{ -\frac{1}{2}(\boldsymbol{\theta}^{(p)} - \boldsymbol{\theta}_{MAP}^{(p)})^T [nI(\mathbf{x}_n|\boldsymbol{\theta}_{MAP}^{(p)}, p)](\boldsymbol{\theta}^{(p)} - \boldsymbol{\theta}_{MAP}^{(p)}) \right\} \quad (3.82)$$

where $I(\cdot)$ denotes the *Fisher information matrix*. Then marginalization of (3.82) w.r.t $\boldsymbol{\theta}^{(p)}$ yields,

$$F(p | \mathbf{x}_n) \approx \mathcal{L}(\boldsymbol{\theta}_{MAP}^{(p)}|p, \mathbf{x}_n) \left(\frac{2\pi}{n}\right)^{\frac{p}{2}} |I(\mathbf{x}_n|\boldsymbol{\theta}_{MAP}^{(p)}, p)|^{-1/2} \quad (3.83)$$

Recall that the Ockham's term is emerged in (3.83) via marginalization. Ignoring terms that are bounded as n increases to infinity, the BIC is derived as,

$$-2 \log F(p | \mathbf{x}_n) \approx -2 \log [\mathcal{L}(\boldsymbol{\theta}_{MAP}^{(p)} | \mathbf{x}_n, p)] + p \log(n) = BIC(p) \quad (3.84)$$

Notably, (3.84) is only valid for large n .

BIC for AR case: substituting $\boldsymbol{\theta}^{(p)} \equiv (\mathbf{a}^{(p)}, r)$ (where the number of free parameters are $p + 1$), and using (3.5), BIC (3.84) is specialized to AR case as follows,

$$BIC(p) = n \log \left[\hat{r}_n^{(p)}(\hat{\mathbf{a}}_n^{(p)}) \right] + (p + 1) \log(n) \quad (3.85)$$

where $\hat{\mathbf{a}}_n^{(p)}$ and $\hat{r}_n^{(p)}$ are respectively the estimate of $\mathbf{a}^{(p)}$ and r . $\square$

Remark 3.5.5. (*BIC versus BPO*)

Although the prior pd of unknowns are assumed in both criteria, the BPO (3.43) is the only consistent and completely tractable criterion for model order inference. In against, BIC (3.85) is obviously achieved with approximations and remained valid for large n . Furthermore, for small n , BPO outperforms other BIC and indeed computational load is small and could not be an obstacle. However, BIC might be popular for large n , in terms of complexity. $\square$

AIC: it is a ML-based principle and no prior for unknowns is adopted. The AIC involves the KLD as follows,

$$D \left[F_{\theta_0^{(p)}} || F_{\hat{\theta}_{ML}^{(p)}} \right] \quad (3.86)$$

where $F_{\theta_0^{(p)}} \equiv F(\mathbf{x}_n | \theta_0^{(p)}, p)$, $F_{\hat{\theta}_{ML}^{(p)}} \equiv F(\mathbf{x}_n | \hat{\theta}_{ML}^{(p)}, p)$ and $\theta_0^{(p)}$ is a target value of parameter. Underlying certain regularity conditions, KLD (3.86) is approximated as,

$$\begin{aligned} D \left[F_{\theta_0^{(p)}} || F_{\hat{\theta}_{ML}^{(p)}} \right] &\equiv E_{F_{\theta_0^{(p)}}(\mathbf{x}_n)} \left[\log \left(\frac{F(\mathbf{x}_n | \theta_0^{(p)}, p)}{F(\mathbf{x}_n | \hat{\theta}_{ML}^{(p)}, p)} \right) \right] \\ &\approx \frac{1}{n} \left[\log(F(\mathbf{x}_n | \theta_0^{(p)}, p)) - \log(F(\mathbf{x}_n | \hat{\theta}_{ML}^{(p)}, p)) \right] \end{aligned} \quad (3.87)$$

Furthermore, using (3.87) the AIC principle of the model comparison is,

$$AIC(p) = 2 \left[\log \left(F(\mathbf{x}_n | \theta_0^{(p)}, p) \right) - \log \left(F(\mathbf{x}_n | \hat{\theta}_{ML}^{(p)}, p) \right) \right] + 2c_o(p), \quad (3.88)$$

where $c_o(p)$ is the number of free parameters (i.e. the dimension of $\theta^{(p)}$). Since, the first term in RHS (3.88) does not relate to the ML estimate, $\hat{\theta}_{ML}^{(p)}$, therefore,

$$AIC(p) \equiv -2 \log \left(F(\mathbf{x}_n | \hat{\theta}_{ML}^{(p)}, p) \right) + 2c_o(p). \quad (3.89)$$

Remark 3.5.6. (AIC for AR case:) substituting $\theta^{(p)} = (\mathbf{a}^{(p)}, r)$ and using (3.5), AIC

(3.89) is specialized to AR case as follows,

$$AIC(p) = n \log \left(\hat{r}_n^{(p)}(\hat{\mathbf{a}}_n^{(p)}) \right) + 2(p+1) \quad (3.90)$$

Remark 3.5.7. (*AIC versus BPO*.) the AIC is completely different from the BPO. Since in AIC, no prior is adopted. Also it involves an approximated version of KLD which usually leads to a poor estimate of p .

Remark 3.5.8. (*HQ*.) this criterion is specifically introduced to adjust AIC criterion in AR case. Since the AIC criterion in the AR case overestimates p , therefore [103] proposed the HQ as follows,

$$HQ(p) = n \log(\hat{r}_n^{(p)}(\hat{\mathbf{a}}_n^{(p)})) + 2(p+1) \log(\log n)$$

MDL and MML are based on the idea of optimal coding theory [104–106] which minimizes total code length,

$$-\log [F(\mathbf{x}_n, \boldsymbol{\theta}, p)] \quad (3.91)$$

MDL: In this criterion, $\boldsymbol{\theta}$ in (3.91) is quantized (due to Shannon theory) to discrete values, i.e. $\boldsymbol{\theta}_d$, therefore (3.91) becomes:

$$\begin{aligned} -\log [F(\mathbf{x}_n, \boldsymbol{\theta}, p)] &= -\log [\epsilon(\mathbf{r}) F(\boldsymbol{\theta}_d | p)] - \log [F(\mathbf{x}_n | \boldsymbol{\theta}_d, p)] \\ &= -\log [\epsilon(\mathbf{r})] - \log [F(\mathbf{x}_n, \boldsymbol{\theta}_d | p)] \end{aligned} \quad (3.92)$$

where $\epsilon(\cdot)$ is volumetric space and $\mathbf{r}$ is,

$$\mathbf{r} = -(\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}_{MAP}^{(k)})^T [n \mathbf{I}(\boldsymbol{\theta}_{MAP}^{(k)} | k, \mathbf{x}_n)] (\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}_{MAP}^{(k)}). \quad (3.93)$$

By definition the discretization allows $F(\mathbf{x}_n, \boldsymbol{\theta} | p)$ to be projected around the discrete value $\boldsymbol{\theta}_d$, for each p , where the Ockham's term is created. The second term of right hand-side (3.92) is approximated using the second-order Taylor expansion around the $\hat{\boldsymbol{\theta}}_{MAP}$. Then,

the expected value of resulting approximation is,

$$-\log [F(\mathbf{x}_n, \boldsymbol{\theta}, p)] \approx -\log [\epsilon(\mathbf{r})] - \log [F(\mathbf{x}_n, \hat{\boldsymbol{\theta}}_{MAP}|p)] + \frac{\mathbf{r}}{2} \quad (3.94)$$

Minimizing $\mathbf{r}$ (i.e. $\hat{\mathbf{r}}$) is revealed by setting the first derivative of right-hand side (3.94) w.r.t $\mathbf{r}$ equals to zero. Then replacing $\hat{\mathbf{r}}$ into (3.94) yields the MDL criterion of p , as follows,

$$MDL(p) = -\log [\epsilon(\hat{\mathbf{r}})] - \log [F(\mathbf{x}_n, \hat{\boldsymbol{\theta}}_{MAP}|p)] + \frac{\hat{\mathbf{r}}}{2} \quad (3.95)$$

Remark 3.5.9. (MDL for AR case:) Let $\boldsymbol{\theta} = (\mathbf{a}, r)$, Rissanen [105] compromised the universal MDL principle and exploited the measure of prediction performance in order to estimate model order in AR case as follows,

$$MDL(p) = \hat{r}_n^{(p)}(\hat{\mathbf{a}}_n^{(p)}) \left[1 + \frac{p+1}{n} \log(n) \right] \quad (3.96)$$

MML (Wallace-Freeman version): Slightly different from MDL which is a criterion just for p , MML is a estimation strategy for estimating both unknown parameter $\boldsymbol{\theta}$ and p . For simplicity, let θ be univariate. Similar to MDL, continuous parameter space is quantized to particles, $\delta_\epsilon(\theta_d) = [\theta_d - \frac{\epsilon}{2}, \theta_d + \frac{\epsilon}{2}]$, with particle's precision ϵ . Recall that,

$$-\log [F(\mathbf{x}_n, \theta, p)] = -\log [F(\theta, p)] - \log [F(\mathbf{x}_n|\theta, p)] \quad (3.97)$$

The p -conditional prior pd of θ is adopted and approximated as follows,

$$\begin{aligned} F(\theta|p) &\approx Pr(\theta \in \delta_\epsilon(\theta_d)|p) \\ &= Pr(\theta_d - \frac{\epsilon}{2} \leq \theta \leq \theta_d + \frac{\epsilon}{2}|p) \\ &\approx \epsilon F(\theta_d|p) \end{aligned} \quad (3.98)$$

Using the second-order Taylor expansion of $\log [F(\mathbf{x}_n|\theta_d, p)]$ around the target estimate of θ (i.e. θ'), and substituting (3.98) into (3.92),

$$-\log [F(\mathbf{x}_n, \theta, p)] \approx -\log(\epsilon) - \log [F(\theta'|p)] - \log [F(\mathbf{x}_n|\theta', p)] - \frac{\epsilon^2}{24} \frac{\partial^2}{\partial \theta'^2} \log [F(\mathbf{x}_n|\theta', p)]$$

(3.99)

The minimizing ϵ (i.e. $\hat{\epsilon}$) is obtained by setting the first derivative of (3.99) w.r.t. ϵ to zero. Substituting $\hat{\epsilon}$ into (3.99), it yields a function of θ' and p as follows,

$$\begin{aligned} -\log [F(\mathbf{x}_n, \theta, p)] &\approx -\log [F(\theta'|p)] + \frac{1}{2} \log \left[\frac{I(\mathbf{x}_n|\theta', p)}{12} \right] - \log [F(\mathbf{x}_n|\theta', p)] + \frac{1}{2} \\ &= -\log [F(\mathbf{x}_n, \theta'|p)] + \frac{1}{2} \log \left[\frac{I(\mathbf{x}_n|\theta', p)}{12} \right] + \frac{1}{2} \end{aligned} \quad (3.100)$$

where $I(\mathbf{x}_n|\theta'^{(p)}) \equiv I(\mathbf{x}_n|\theta', p) = -\frac{\partial^2}{\partial \theta'^2} \log [F(\mathbf{x}_n|\theta', p)]$ is Fisher information matrix. Using (3.100), the p -conditional MML estimate of $\theta'^{(p)}$ is attained,

$$\hat{\theta}_{MML}^{(p)} \equiv \underset{\theta'^{(p)}}{\operatorname{argmin}} \left\{ -\log [F(\mathbf{x}_n, \theta'^{(p)}|p)] + \frac{1}{2} \log \left[\frac{I(\mathbf{x}_n|\theta'^{(p)})}{12} \right] + \frac{1}{2} \right\} \quad (3.101)$$

Recall that exploiting CE technique and replacing MML estimate (3.101) into (3.100), the MML criterion of p is derived as,

$$MML(p) \equiv -\log [F(\mathbf{x}_n, \hat{\theta}_{MML}^{(p)}|p)] + \frac{1}{2} \log \left[\frac{I(\mathbf{x}_n|\hat{\theta}_{MML}^{(p)})}{12} \right] + \frac{1}{2} \quad (3.102)$$

A multivariate version of MML principle (3.100) is achieved after some manipulations [100] as follows,

$$MML(\theta'^{(p)}, p) \equiv -\log [F(\mathbf{x}_n, \theta'|p)] + \frac{p}{2} [1 + \log(R_p)] + \frac{1}{2} \log [\det \{I(\mathbf{x}_n|\theta'^{(p)})\}] \quad (3.103)$$

where R_p is the lattice constants [106]. Recall that MML criterion (3.103) is known as MML87 in literature i.e. [106], [107]. Notice that MML estimate of θ and p can be computed with the same procedure as the univariate case. In the following, we deduce a MML criterion of p , based on EIM (3.20) as follows.

Remark 3.5.10. (MML for AR case:)

- Let $\theta = (\mathbf{a}, r)$, using the submartix $V_{n,aa}^{(p)}$, we can define a MML criterion of p as

follows,

$$MML(p) = n \log [\hat{r}_n^{(p)}(\hat{\mathbf{a}}_n^{(p)})] + \frac{p}{2}[1 + \log(R_p)] + \frac{1}{2} \log |V_{n,aa}^{(p)} - V_{p,aa}^{(p)}| \quad (3.104)$$

Remark 3.5.11. (*(MDL and MML versus BPO:)* MDL and MML are both pseudo-Bayesian approach in which a prior pd on only θ is considered but not for p , instead CE technique is employed. They involves many manipulation and approximations. In against, the BPO is simple and tractable which automatically engenders the Ockham's term without any approximation.

Remark 3.5.12. (*Bayesian view of the classical model order estimate:*)

From Bayesian perspective, classical order estimate of p , is equivalent to degeneration of posterior model order (3.41) in $\hat{p}$, i.e.,

$$\Pr\{L = p | \mathbf{x}_n\} \approx \delta(p - \hat{p}), \quad (3.105)$$

which is effectively a CE strategy.

3.5.6 Parameter estimation and prediction

When p is unknown, the classical estimation and prediction strategies replaces the model order estimate, $\hat{p}$ (3.105), in p -conditional estimations e.g. $\hat{\mathbf{a}}_{n,LS}^{(p)}$ (3.64) and $\hat{r}_{n,LS}^{(p)}$ (3.70). Also, the classical prediction is achieved by substituting $\hat{p}$ (3.105) into OSA prediction (3.77) and MSA prediction (3.78).

3.6 Computational load of extended information matrices (EIMs)

In this section, recursive algorithms are presented to reduce Bayesian computational loads. For this reason, we provide efficient recursions on EIM—which is the core object in Bayesian computation. We note that the data can be processed in *off-line scenario* (in n -data batch),

3.6.1.1 Incremental way

In this way, the EIM is augmented in order (i.e. from $L = p - 1$ to $L = p$), using (3.108) in off-line scenario. This augmentation requires $N = (n - p)(p + 1) + \frac{p(p+1)}{2}$. Therefore, the total multiplications is,

$$N_{I-o} = 3(n - 1) + \sum_{p=2}^{p_u} [(n - p)(p + 1) + \frac{p(p + 1)}{2}] \quad (3.109)$$

where N_{I-o} denotes the multiplications required incremental order via (3.108).

3.6.1.2 Decremental way

In this way, EIM is computed by extracting the lower sub-matrix from the maximum order (i.e. $L = p_u$) to minimum order ($L = 1$) using (3.108). In particular, eliding the first row and column of EIM at order $L = p$ and adding the matrix $\varphi_n^{(p-1)}\varphi_n^{(p-1)T}$ to sub-matrix $V_{n,aa}^{(p)}$ yields EIM at order $L = p - 1$. This requires only $N = \frac{p(p+1)}{2}$. Therefore, the total computational load for all orders in this way is,

$$N_{d-o} = (n - p_u)[(p_u + 1) + \frac{p_u(p_u + 1)}{2}] + \frac{1}{2} \sum_{p=2}^{p_u} p(p + 1), \quad (3.110)$$

where N_{d-o} denotes the multiplications required for applying the decremental way of the order using (3.108). The following lemma confirms that the decremental way is more efficient in terms of computational loads exploiting (3.108).

Lemma 3.6.1. *The computational load of exploiting the decremental way i.e. (3.110) is smaller than that of the incremental way i.e. (3.109) using (3.108) for $p_u \geq 8$. The proof is given in Appendix A.5.* □

3.6.2 Performance in simulation

In this section, we provide the extensive simulation result using the comparison of different ways for computing EIM in off-line scenario. Our PC processor is Intel Core(TM) i7 – 4770 CPU @3.40 GHz (8CPUs), 3.4 GHz. In Figure 3.1, we compare computational load (3.110) with (3.109) and (3.106) by plotting running time(s) in terms of maximum order p_u with

$n = 1000$. This figure shows that the decremental way is the fastest way to compute EIMs using the recursion in order (3.108) for different $p_u = \{10, 20, 30, 40\}$.

Moreover, Figure 3.2 shows the performance of different ways in accuracy of BPO (3.43)

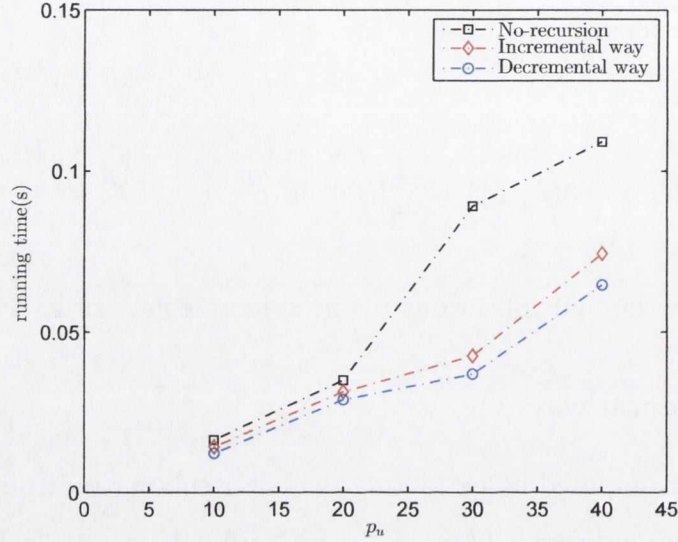


Figure 3.1. Comparison of different ways of computation EIMs using recursion in order (3.108), $n = 1000$

to estimate model order in terms of different p_u . Here data are generated from $AR(4)$ (in LHS) and $AR(8)$ (in RHS). We here fix computational load to $N = 6 \times 10^7$ at the level of EIM for $p_u = \{10, 14, 18, 20\}$, e.g. for $p_u = 10$, the number of available data for no-recursion, recursion in order with incremental and decremental ways are respectively $n = \{245, 960, 1050\}$.

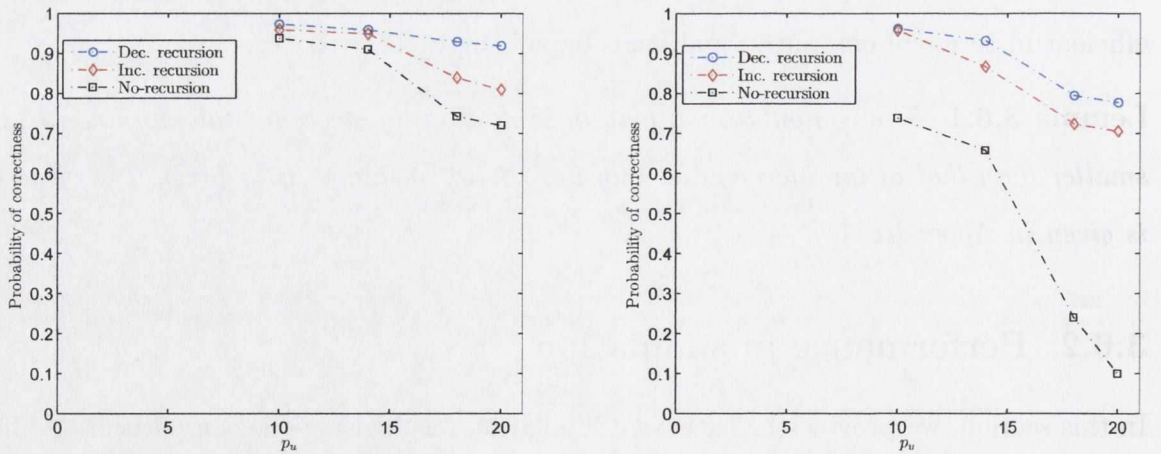


Figure 3.2. Performance of different ways in accuracy of model order inference using recursion in order (3.108) with $N = 6 \times 10^7$. Data are simulated from $AR(4)$ (LHS) and $AR(8)$ (RHS).

3.6.3 Computational load of EIMs: on-line scenario

Recall that the on-line scenario allows us to interleave updating data with computing EIM [61]. The exploitation of recursion in time and also a full-recursion (i.e. recursion in both time and order) will be worthwhile in on-line scenario. Notice that the recursion in time is available in literature i.e. [108] and [61]. However, its computational load is not audited. The recursion in time on EIMs [61] is:

$$\begin{aligned} V_i^{(p)} &= g^{(p)}(V_{i-1}^{(p)}, \varphi_i^{(p)}) \\ &= V_{i-1}^{(p)} + \varphi_i^{(p)} \varphi_i^{(p)T}; \quad i = p+1, \dots, n \end{aligned} \quad (3.111)$$

where $g^{(p)}(.)$ is the transformation function on EIM and is invariant w.r.t time. Using (3.111) $N = \frac{(p+1)(p+2)}{2}$ is required to update EIM from time i to $i + 1$ at order p . Therefore, updating EIM from $i = p + 1$ to $i = n$ at order p requires,

$$N = (n - p) \frac{(p + 1)(p + 2)}{2}.$$

Moreover, we exploit (3.108) and (3.111) to introduce full-recursion,

[illegible]

where $g(\cdot)$ is transformation function on EIMs and is invariant w.r.t time and order. Notice that the full recursion (3.112) enables us to update EIMs in both time and order with $N = i(p + 1) + p(p + 1)$ multiplications. Notice that the construction of EIMs for all possible order in a specific time i in the on-line scenario can be seen as the construction of EIMs with i -batch data in off-line scenario. Also from Lemma 3.6.1, the decremental way in (3.6.1.2) is more efficient. Therefore, we investigate computational load of EIMs in decremental way in on-line scenario.

3.6.3.1 Decremental way

Exploiting decremental way (3.6.1.2) at fixed time i and then updating time using (3.111) yields an efficient algorithm to compute EIMs as follows.

Algorithm (Full-recursion in the decremental way:)

1. Compute EIM at time $i = 2$ (since $p \geq 1$) in order $L = 1$. It requires $N = 3$.
2. Exploit (3.111) to update the EIM from time i to $i + 1$ ($2 \leq i \leq p_u$), and use (3.108) to extract lower submatrix from order $p = i - 1$ to one at time $i + 1$ ($3 \leq i \leq p_u$).
The total number of multiplications required is

$$\begin{aligned}
 N &= \sum_{p=1}^{p_u-1} \frac{p(p+1)}{2} + \sum_{p=2}^{p_u} (p+1)\left(1 + \frac{p}{2}\right) \\
 &+ \sum_{p=3}^{p_u} [(p+1)^3 - 3(p+1)^2 + 2(p+1) - 1] \\
 &= \frac{1}{4}p_u^4 + p_u^3 + \frac{1}{12}p_u^2 - \frac{1}{3}p_u - 4
 \end{aligned} \tag{3.113}$$

3. The $(n - p_u - 1)$ time update of EIMs from $i = p_u + 1$ to $i = n$ at order $L = p_u$ requires $N = (n - p_u - 1) \frac{(p_u+1)(p_u+2)}{2}$.
4. Apply the decremental way to draw lower sub-matrix from the order p_u to one at time i ($p_u + 2 \leq i \leq n$) requires $N = \sum_{p=2}^{p_u} (n - p_u - 1) \frac{p(p+1)}{2}$. $\square$

Therefore, the total computational load of the Algorithm 3.6.3.1 is

$$\begin{aligned}
 N_{d-f} &= \frac{1}{4}p_u^4 + p_u^3 + \frac{1}{12}p_u^2 - \frac{1}{3}p_u - 4 + (n - p_u - 1) \frac{(p_u+1)(p_u+2)}{2} \\
 &+ (n - p_u - 1) \sum_{p=2}^{p_u} \frac{p(p+1)}{2}
 \end{aligned} \tag{3.114}$$

where N_{d-f} is the number of multiplications needed for decremental way using full-recursion (3.112). Figure 3.3 shows the schematic diagram of computation EIMs for Algorithm 3.6.3.1.

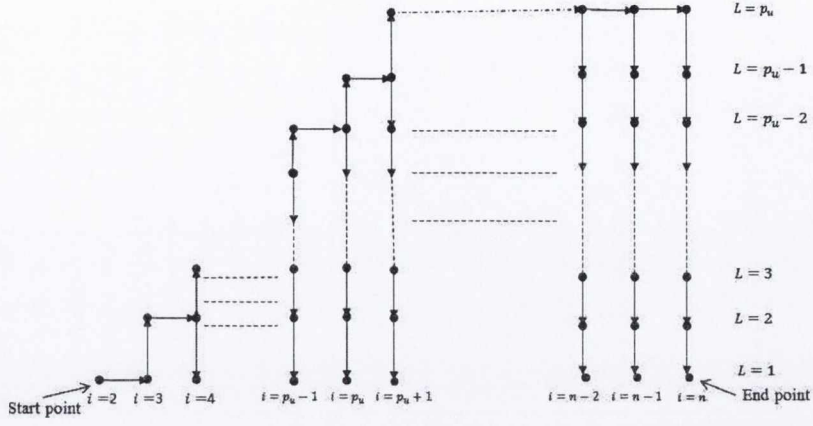


Figure 3.3. Schematic diagram for computation of the EIM using full recursion (3.112) for decremental way in on-line case

3.6.4 Comparison of computational load of EIMs between two different scenarios

In this part, we aim to compare computational load of EIM in the off-line scenario versus the on-line scenario both in decremental way. Notice that the computation of EIM in off-line scenario can also be performed sequentially through time at different orders via the on-line scenario. This analogy is displayed in the following lemma.

Lemma 3.6.2. *Computation of EIMs using the decremental way in off-line case requires smaller number of multiplications compared to the on-line scenario. The proof is presented in Appendix A.6.* $\square$

We here provide a simulation-based evidence for Lemma 3.6.2. In Figure 3.4, we audit the running time for exploiting the decremental way for different scenario in terms of p_u . This figure shows that decremental way in off-line case is faster than that of on-line case in order, with $n = 1000$ for different $p_u = \{10, 20, 30, 40\}$. Also this difference is more distinctive as p_u grows.

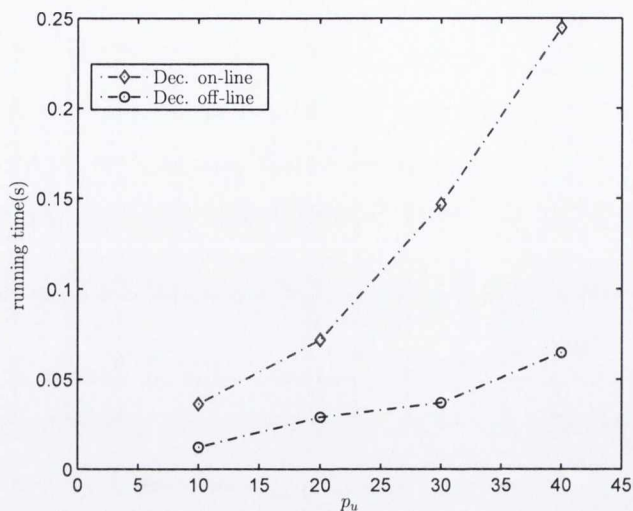


Figure 3.4. Comparison of running time of decremental way in off-line case versus on-line case, $n = 1000$

3.7 Comparative performance evaluation under simulation

In this section, we aim to compare different criteria of the model order selection in extensive simulation. Then point estimates of a subset of AR coefficients of different methods are compared. Also, simulation results for OSA and MSA predictors using different approaches are provided for the AR model.

3.7.1 Comparison between classical approaches and Bayesian method in model order selection

In this part, we examine a simulation result to compare the performance of different criteria for model order selection in term of some operating factors i.e. maximum model order (p_u), number of observations (n), and true model order (p). Notice that we also exploit LS method (3.63) to estimate AR coefficients in classical estimation strategies. Recall that we design simulation with halving poles randomly in the annular,

$$\{(r, \theta); a < r < b \ (0 < a < b < 1), \ -\pi < \theta < \pi\}, \quad (3.115)$$

where, r and θ ¹ respectively indicate the radius and angle of a pole in the z -plane which is distributed as,

$$f(\theta, r) = f(\theta) \cdot f(r) \tag{3.116}$$

where

$$\begin{aligned} f(\theta) &= \frac{1}{2\pi} & -\pi < \theta < \pi \\ f(r) &= \frac{2r}{b^2 - a^2} & a < r < b. \end{aligned} \tag{3.117}$$

The rest of poles are their conjugate complex pairs. In order to have a desirable performances, let $a = 0.7$ and $b = 0.9$, and MC denotes the number of Monte Carlo runs. Figure 3.5 shows the performance of the different criteria of model order selection by assessing the probability of correct identification of true order in terms of maximum possible order p_u . Observations here are simulated from $AR(4)$ with $n = 1000$. From this Figure, the identification of the model order in all criteria does not depend on the choice of p_u . Notice that BPO generally outperforms other criteria, however it is competitive with BIC, MDL and MML. The reason can be followed from the discussion in 3.5.5.1. In

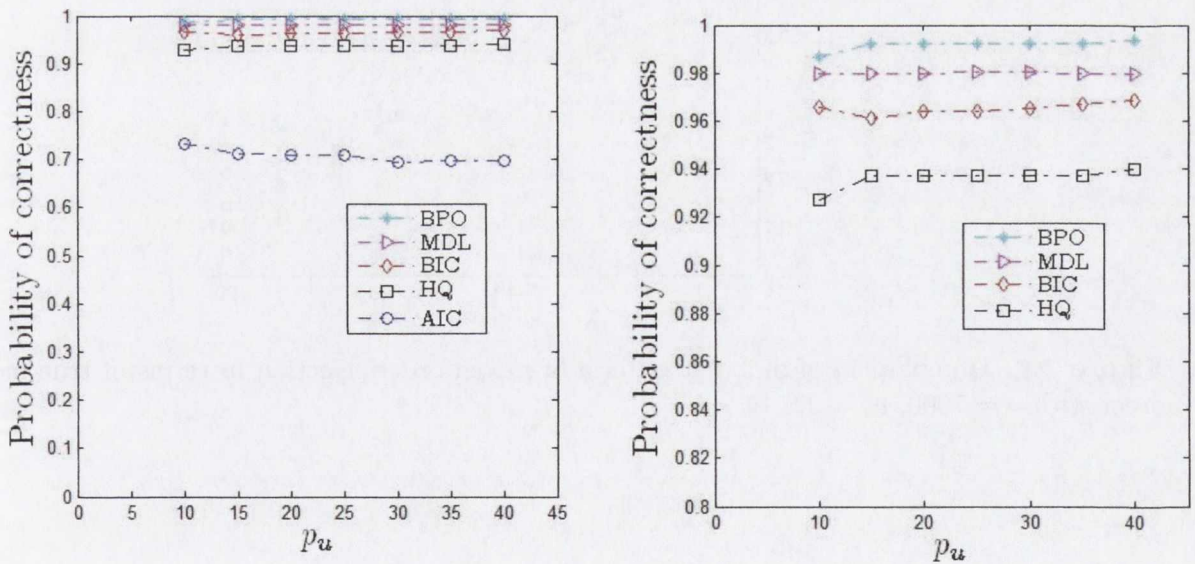


Figure 3.5. Comparison of different criteria of model order selection in terms of p_u with $AR(4)$, $n = 1000$.

Figure 3.6, different criteria are compared in terms of the true model order. In this figure,

¹Here θ is the old-fashioned symbol for the angle not parameter.

the simulated data are designed with $p_u = 20$, $n = 1000$ and $MC = 1000$. In general, this figure shows that BPO outperforms others. Once again BIC, MDL and MML are competitive with BPO for $p \leq 10$, in the identification of the true model order. Notice that the ability of identifying the true order in all criteria decreases with the increase of the true order. To see the reason, we repeat simulation with $MC = 10000$ in the Figure 3.7. A comparison between the Figures 3.6 and 3.7 shows that the performance usually collapse for large p . In particular, this collapse happens for larger p by increasing n , which is related to the small value of ratio, n/p . In Figure 3.8, we examine the different criteria of model order selection in terms of observed data (i.e. $n = \{100, 500, 1000\}$). Here the simulation is designed with $p_u = 20$, $AR(4)$ (LHS) and $AR(16)$ (RHS). It obviously shows that BPO outperforms for both designs. In particular, it is more distinctive where the model order is increased from $p = 4$ to $p = 16$.

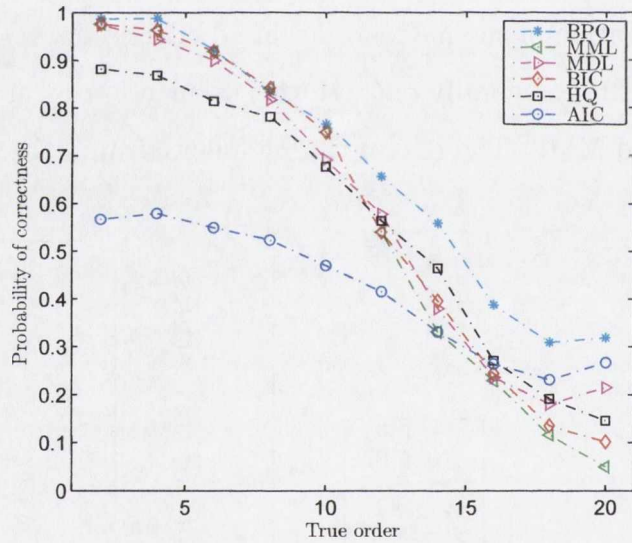


Figure 3.6. Comparison of different criteria of model order selection in terms of true model order with $n = 1000$, $p_u = 20$.

3.7.1.1 Comparison between classical approaches and Bayesian method for the estimation of the subset of parameters

In this part, we aim to compare the performance of our Bayesian approach with LS estimation strategy in order to estimate the subset, $\mathbf{a}_\delta^{(p)}$. We first design the simulation

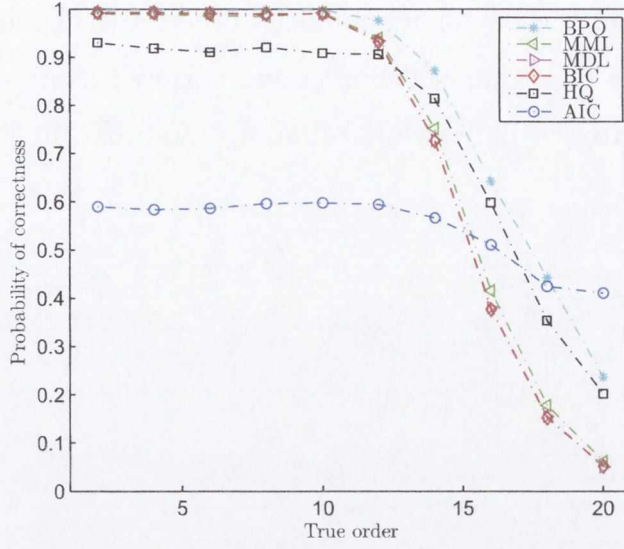


Figure 3.7. Comparison of different criteria of model order selection in terms of true model order with $n = 10000$, $p_u = 20$.

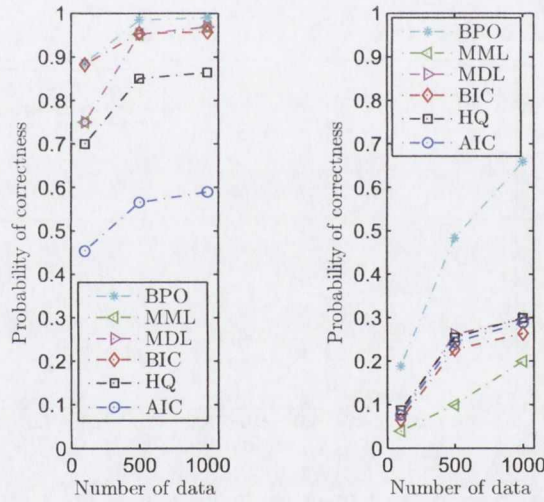


Figure 3.8. Comparison of different criteria of model order selection in terms of number of observations designed with $AR(4)$ (LHS), $AR(16)$ (RHS) and $p_u = 20$.

for known p . In this case, we simulate data from $AR(10)$ (Figure 3.9 above) and $AR(14)$ (Figure 3.9 below) for different choices of the n (i.e. $n = \{50, 100, 500, 1000\}$). Also, $\mathbf{a}_\delta^{(p)}$ for $|\delta| = 3$, is randomly chosen from $\mathbf{a}^{(p)}$. For simplicity, we compare these criteria using the evaluation of the Euclidean norm (EN) as follows:

$$EN(n, p) = \|\mathbf{a}_\delta^{(p)} - \hat{\mathbf{a}}_{n,\delta}^{(p)}\|. \quad (3.118)$$

Figure 3.9 shows that the Bayesian approach outperforms classic estimation to estimate $\mathbf{a}_3^{(p)}$ for different choices of true model order and n as well. Notice that the performance of different criteria converges by growing n from $n = 50$ to $n = 1000$.

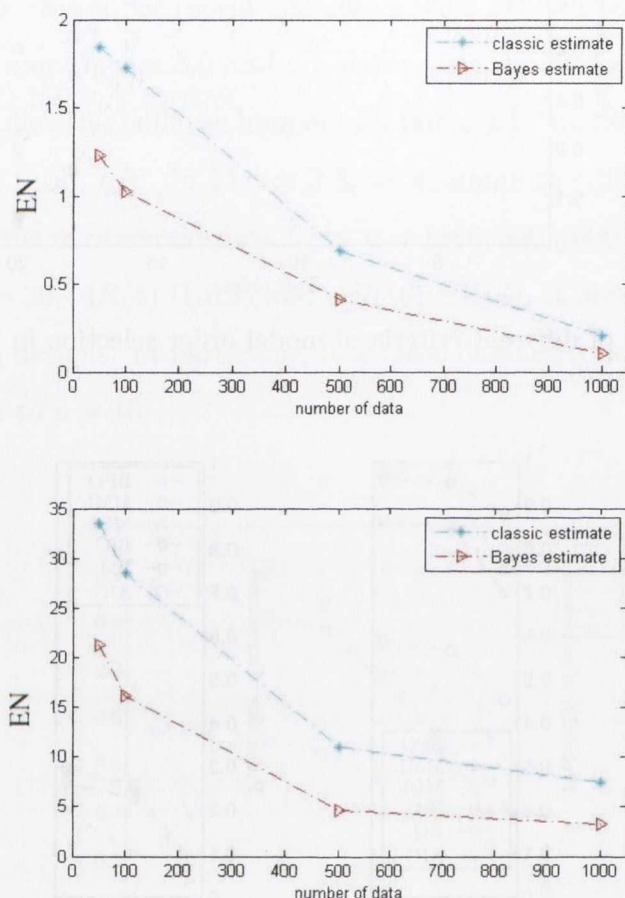


Figure 3.9. Comparison of different criteria for estimation of an arbitrary subset in terms of the number of available observations n , with $|\delta| = 3$, and simulated from $AR(10)$ (in the above) and $AR(14)$ (in the below).

Next we compare the performance of BPO approach versus to CE-based criteria in subset estimation for the case where p is unknown. For this reason, let $\mathbf{a}_\delta = (a_1, a_2, a_4)$, and $|\delta| = 3$. Recall that the CE-based criteria usually exploit (3.90)-(3.96) and then select the corresponding estimates of the subset, i.e. $\hat{\mathbf{a}}^{(\hat{p})}$, where $\hat{p}$ is estimated order, in the case $\hat{p} \geq 4$. Notice that in order to compute $\hat{\mathbf{a}}^{(\hat{p})}$, we use LS estimate of $\mathbf{a}^{(\hat{p})}$ for CE-based criteria except for MML where this method provides the estimate of parameters too. In contrast, our Bayesian proposal is the linear combination of the first shaping parameters of

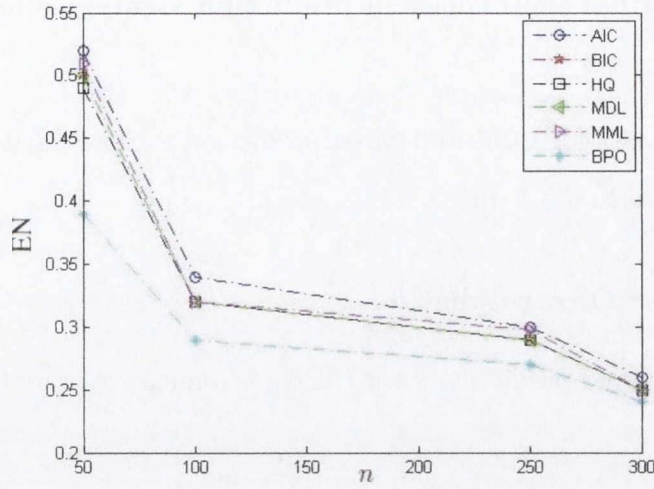


Figure 3.10. Comparison of different criteria for estimation of an arbitrary subset in terms of number of observations, n , with $|\delta| = 3$, $p_u = 10$ and simulated from $AR(4)$.

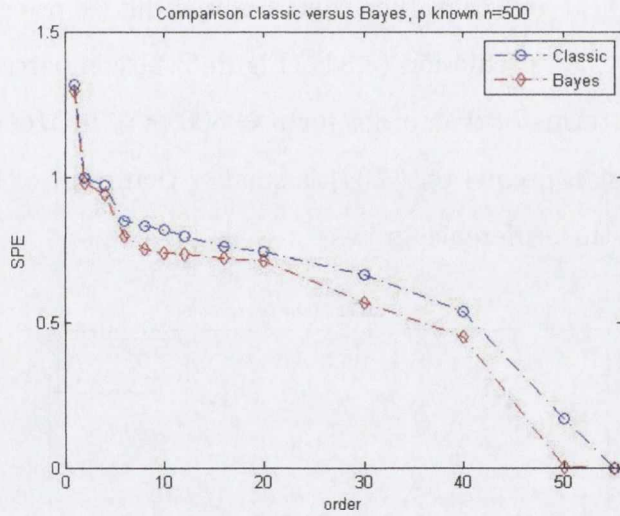


Figure 3.11. Comparison of different criteria in one-step-ahead prediction in terms of true model order (known) with $n = 500$.

Student's- t pd with weights $\beta_n^{(p)}$ (3.44).

Figure 3.10 shows the performance of different criteria assessing the norm between the simulated with estimated value of $\mathbf{a}_\delta$, in terms of n for $|\delta| = 3$, $p_u = 10$. Here the data are simulated from $AR(4)$ for $MC = 1000$. It is obvious that our Bayesian approach outperforms CE-based criteria. In particular, all criteria converge by growing n . Notice that for consistency reason, we consider $L \geq 4$.

3.7.1.2 Comparison between classical prediction strategies and Bayesian prediction

Here the simulation results for OSA and MSA prediction are provided for the AR model discussed in Sections 3.3.2 , 3.3.3 and 3.5.4.

Simulation results for OSA prediction

We here consider the squared prediction error (SPE) as a measure of performance as follows

$$SPE(n, p) = \frac{1}{MC} \sum_{i=1}^{MC} (e_{n+1}^{(p)})_{(i)}^2 \quad (3.119)$$

Notice that we here use Yule-walker approach [98] for computing OSA prediction (3.77) for LS strategy as a classical approach. For this reason, a diffuse prior, $V_p^{(p)} = (0.001)I_p$ is adopted to compute Bayesian prediction (3.34). Figure 3.11 compares (3.34) versus (3.77) by evaluating (3.119) in terms of true order for $n = 500$, $r = 1$, $MC = 1000$. This figure shows that SPE of Bayesian prediction (3.34) is smaller than that of LS prediction (3.77) and by growing order p , this difference grows.

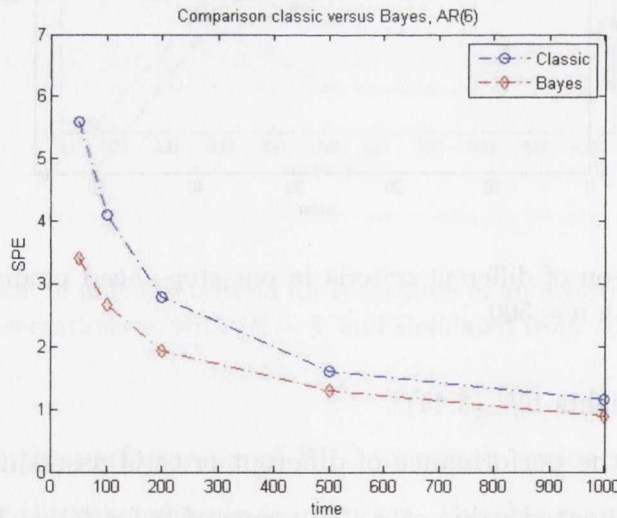


Figure 3.12. Comparison of different criteria in OSA prediction in terms of time simulated from $AR(6)$ and $r = 1$.

Figure 3.12 shows the evaluation of (3.119) for different criteria in terms of n . It is shown that SPE of Bayesian prediction (3.34) is smaller than that of LS prediction (3.77), they converge by growing n . Notice that increasing n leads to the tendency of (3.119) to $r = 1$

which is considered as the variance of simulated data. In other words, for large n , the estimate of AR parameters are more accurate and squared prediction error tends to be r .

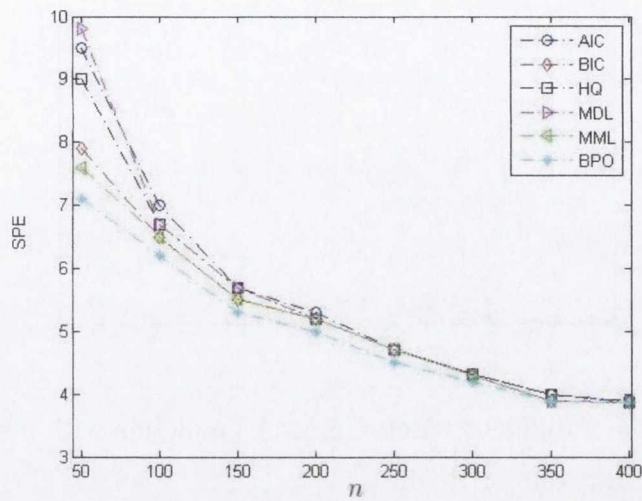


Figure 3.13. Comparison of different criteria in OSA prediction in terms of data, n , with $p_u = 20$, and simulated from $AR(6)$ and $r = 4$.

Figure 3.13 shows the evaluation of (3.119) in terms of time for $AR(6)$ and $r = 4$, $p_u = 20$. This also shows that the Bayesian approach outperforms other approaches for small n . Furthermore SPE (3.119) for all criteria converge by growing n .

Simulation for k -step ahead prediction:

The different criteria of MSA prediction (3.37), (3.38) and (3.78) are here compared in simulation. Figure 3.14 shows that despite of deterioration performance of different criteria (i.e. increasing trend in SPE) for large steps, the rate of increasing SPE of (3.37) is less than other approaches for $AR(8), r = 2, n = 100$. Moreover, Figure 3.15 shows the SPE for 6-step ahead of different criteria in terms of different true model order. This figure shows that Bayesian approach (3.37) outperforms for all true model order, for different value of p .

3.8 Discussion

We have provided a unique, consistent and tractable Bayesian inference of model order, and any subset of global set of parameters of the AR model. Notably, we no longer need

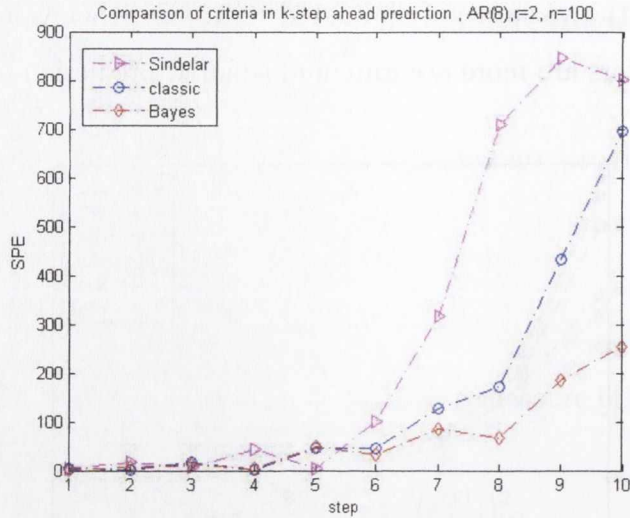


Figure 3.14. Comparison of different criteria in MSA prediction with $n = 100$ and simulated from $AR(8)$ and $r = 2$

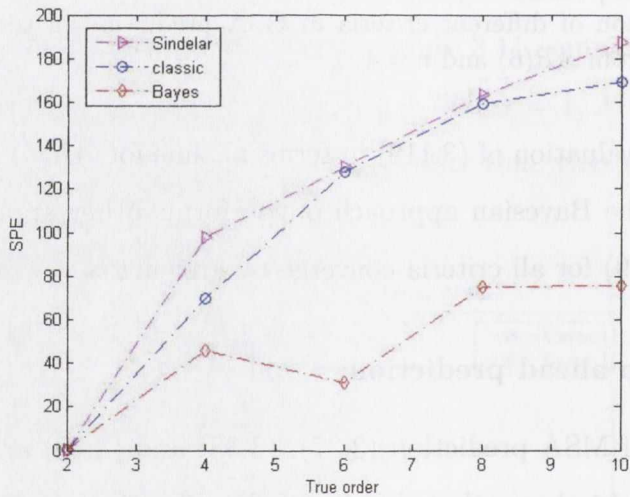


Figure 3.15. Comparison of different criteria in 6-step-ahead prediction with $r = 2, n = 100$

to exploit CE technique to compute OSA and MSA prediction. Instead, we have applied Bayesian predictors discussed in Sections 3.3.2 and 3.3.3 which are indeed simple and elegant. LDL decomposition is popular for Bayesian computations in the current Bayesian literature [55,91]. Recursions introduced in Section 3.6 provide us efficient tools to compute Bayesian tasks without a need of LDL decompositions.

We have discussed on widely used criteria of model order estimation and specialized them into the AR case. Moreover, automatic Ockham's regularization of BPO has been mathematically justified and a comprehensive explanation of emerging Ockham's term in

estimation strategies of model order in AR case is provided. Despite a massive literature on MDL and MML principles, yet they are not fully explored in AR case in Bayesian context. Some possible explorations will be discussed in Chapter 8.

The proposed Bayesian inference can be exploited in multiple participant systems (MPS) [48]. In particular, when each participant (i.e. an automate node [109]) models x with the AR model and processes the data set into a predictive pd (i.e. Student's- t pd). Then it is quite important to process these predictive pds in order to reach a global design. A promising context for applying this idea is available in [48]. This task will be explored in Chapter 6.

Chapter 4 —

Multiple participant processing: the moral deliberator

In this Chapter, we address the fundamental problem of optimal processing of stochastic knowledge distributed among nodes in a knowledge-sharing network. One of the nodes—called the deliberator¹—mediates between all the other nodes in order to reach an optimal compromise, and each of these other nodes interacts only with the deliberator. We will refer to the contributing nodes as participants and knowledge-sharing network as a system. The deliberator is a Bayesian moral (i.e. empathetic) decision maker, who bound their own expectation of the expected loss (KLD) suffered by each participant, were they to accept the merged pd. This is fully explained in Section 4.3. Therefore, we shorten the title of this Chapter to *multiple participants processing; the moral deliberator*.

Processing stochastic knowledge by the deliberator is essentially a distributional DM. In particular, this is the specialization of distributional design presented in Section 2.3 where the knowledge K is defined in a particular way and consequently the ideal pd is specified in a particular form. Now, we can implement the fundamental FPD (2.14) to the interesting problem in engineering where processing knowledge of distributed MP arises [30, 43, 110, 111]—which is fully expressed in Section 4.2.2. We start to review the schemes of DM for merging knowledge distributed in MPS. The merits and demerits of each scheme are addressed. Then, different approaches of the distributional design within centralized scheme in MPS is explored. Next, we define the knowledge and formulate the task of DM. More importantly, our proposed scheme and methodology are adopted for merging multiple pds of the same x .

¹Throughout this thesis, we adopt the gender-neutral possessive adjective, “*their*”, in place of “*his/her*”, and we adopt the gender-neutral pronoun, “*they*”.

4.1 DM schemes for merging knowledge in multiple participant system (MPS)

The task of merging knowledge in MPS is very often referred as the *consensus* of knowledge. It seems that the consensus knowledge of participants originated in *management science* and *statistics* i.e. [112, 113] and references therein. Then some Bayesian approaches have been proposed to merge expert judgements i.e. [114]. Latter, they have been implemented in the wireless network [32, 115].

In this part, we review different DM schemes for merging the distributed knowledge in MPS. Two major DM schemes are widely adopted in the literature, namely the *decentralized* (i.e. distributed, flat) DM scheme i.e. [29, 38, 111] and the *centralized* (i.e. non-flat) DM scheme [36, 41, 48]. In the following, each scheme is expressed with its advantages and disadvantages.

- **DM centralized scheme:** In this scheme, the local nodes i.e. participants, report their data/pds to the central node, i.e. deliberator. The deliberator is also referred as the facilitator [39]). The deliberator processes the provided knowledge to reach a global and optimal decision. In this scheme, all DM efforts are happened at a single node (deliberator) where a global decision is reached. We note that *hard detection technique* for the purpose of target detection in the fusion of multiple sensors, i.e. [29], has some similarity with this scheme. some other published applications of the centralized DM scheme include a dynamic load balancing scenario for wireless networks [40], the classification of geographic data from multiple sensors [116], reliability estimation among multiple sources [28], optimized traffic management [42, 117], and airspace resource management [41].
- **DM decentralized scheme:** In this scheme, the deliberator does not exist and DM efforts are distributed among participants. In particular, all nodes have the capability of making a decision. They process their own knowledge and the knowledge provided by their *neighbouring* participants [118] to reach a local decision. Notice that neighbouring participants of a specific participant i.e. $\mathcal{I}_1$ are referred to participants who provide knowledge for $\mathcal{I}_1$, i.e. $\mathcal{I}_2$ and $\mathcal{I}_3$ in Figure 4.1. In the decentralized

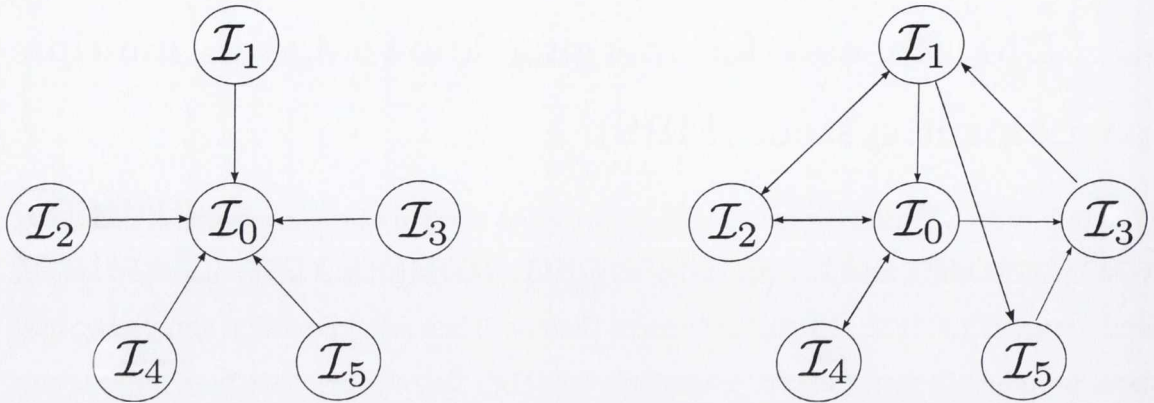


Figure 4.1. Schematic diagram for different structures; Left-hand side is the structure of a centralized deliberator-based scheme and Right-hand side is the structure of a decentralized scheme.

scheme, there is no a central node for providing the global decision. The decentralized DM scheme has been applied in fault diagnosis for telecommunication networks [30], feature extraction for big data sets [31], analysis of consensus algorithms in networked systems [32], and optimization of rewards in on-line learning [43], for example.

The schematic diagram for different DM schemes is presented in Figure 4.1.

Remark 4.1.1. (*Advantages and disadvantages of each scheme*).

- *The centralized scheme allows us to provide a global decision reached by the deliberator. This global decision is acceptable to be deserved as the optimal decision for the whole system. However, in the traditional centralized scheme, the deliberation (i.e. decision-making) itself is confined to the deliberator. This may lead to scalability problems [29] in cases where the number of participants, s , is large. The vulnerability of failure of the overall system to deliberator node might happen [109]. In particular, it is not easy to transmit the big data set to the deliberator, in the system where each participant involves a large number of local state and control variables i.e. [119], or in highly dynamic networks i.e. [110]. Furthermore, handling a centralized scheme for large-scale systems with a big amount of data creates time delays and even failure in the knowledge transmission [120].*
- *In the current research, the decentralized scheme receives more attraction due to low-complexity [111] and scalability [29]. However, this scheme has its own drawbacks. For*

instance, it is unable to propose a global solution, and only provides local decisions—which increase conflict among participants. Also local decisions are expected to serve to the whole system, while they might not be reliable due to limited knowledge/resources of participants. Furthermore, designing an appropriate and formal way of communication (interconnection) among participants, in particular, for a large-scale system is the other challenge. These properties reduce the real applications of a decentralized DM scheme in practice i.e. [41].

Therefore, it is worthwhile to explore a centralized deliberator-based scheme for merging the distributed knowledge in order to solve the current issues. In our proposed scheme, no data are transmitted from the participants to the deliberator. Instead, the participants report only pds. In particular, reporting the shaping parameter of pds is adequate. Therefore, the issue of transferring data and time delays are hugely improved. The deliberator’s task is to process the shaping parameters for designing the optimal pd. This scheme leads to the following advantages.

- **Low computational complexity:** The proposed scheme requires low computational load. Because, the large scale of raw data are converted to shaping parameters in each participant. Then the shaping parameters are reported to the deliberator. This leads to a significant computational savings in the deliberator.
- **The reduction of time delays:** The transmission of the whole data-batch from the distributed participants to the deliberator, requires many bandwidths and creates long-time delays. But in our proposed scheme, only few bandwidths are needed to transfer shaping parameters.

4.2 Distributional design in MPS

The task of distributional design through merging multiple pds are widely studied. For instance, [121] comments to infer the posterior pd of the parameter through merging knowledge partially shared among multiple participants. It is assumed that participants are dependent because of shared knowledge. Notice that this assumption leads to an incomplete model of knowledge. References [114] and [122] are the other related works proposed under an incomplete model of knowledge.

One of desirable approach for merging multiple pds is copula-based approach i.e. [123]. Following the Sklar's theorem, the construction of joint pds of $(x_1, \dots, x_n)$ can be written as the copula function, $C(\cdot)$, of marginals:

$$G_0(x_1, \dots, x_s) = C[G_1(x_1), \dots, G_s(x_s)] \quad (4.1)$$

Notice that (4.1) is general formula for copula functions where $G_l, l = 1, \dots, s$, are cumulative distributions. For differentiable G_l and $C(\cdot)$, the joint density is also available i.e. [123]. Also for the purpose of sensor fusion and feature selection, [124] constructs the joint pd of sufficient statistics through merging marginal pds using hypothesis test. Then [125] studies the construction of the joint pd of features within exponential family. The constructed pd can be seen as the MXE-optimal design (2.24) (the special case of FPD) where the knowledge-constrained set is the set of moment constraints.

The other popular methods are distributional pools. Reference [126] suggest the pooling of different forecasts with unequal weights. Pooling design has been advocated as a established approach for combining expert judgements i.e. [127, 128].

In the rest, we focus on distributional design in MPS, where all nodes construct pd of same x of the same functional form with differently chosen hyperparameters. This is not the requirement, but we use it in what follows for ease of interpretation. In particular, we are interested in investigating distributional pools and FPD approaches as very popular approaches for merging pds in MPS. In the following, each approach is fully addressed.

4.2.1 Distributional pools in MPS

Distributional pools can be seen as a construction of an additive or multiplicative mixture across participants. Recall that the task of distributional pool is to merge s pds in order to reach the optimal pd. The distributional pools are also referred to opinion pools i.e. [129]. We here specifically focus on the optimal design of predictive pd using pooling approaches. The most popular distributional pools are linear and logarithmic pools.

- **Linear pooling prediction:** The pioneering works of linear pool are [112, 130]. Latter, the linear pool has been quite popular to combine multiple prior pds of the same parameter i.e. [35, 131, 132]. It is also widely used for merging predictive

pds of same x [34]. This pooling approach involves a weighted linear mixture of participant-indexed predictions, G_l s, as follows:

$$G_0^a \equiv G_0^a(x) = \sum_{l=1}^s w_l^a G_l(x), \quad (4.2)$$

where G_0^a denotes the linear pooling prediction and the w_l^a is the weight of l -th participant which belongs to $(s-1)$ —probability simplex Δ_{s-1} (i.e. $w_l^a \in \Delta_{s-1}$). This pooling approach is justified [34] as the solution of,

$$G_0^a \equiv \arg \min_{G \in \mathbf{G}_K} \sum_{l=1}^s w_l^a D[G_l || G], \quad (4.3)$$

where G is the variational pd as a member of knowledge-constrained set of predictive pds, $\mathbf{G}_K$, and $D[.||.]$ is defined in Definition 2.2.1.

- **Logarithmic pooling prediction:** This pool is a weighted geometric mixture of participant-indexed prediction, with weights w_l^g as follows:

$$G_0^g \equiv G_0^g(x) = \frac{1}{\kappa_{G^g}} \prod_{l=1}^s [G_l(x)]^{w_l^g}, \quad (4.4)$$

where $\kappa_{G^g} = \int \prod_{l=1}^s [G_l(x)]^{w_l^g} dx$, G_0^g denotes the logarithmic pooling prediction and by definition, $w_l^g \in \Delta_{s-1}$. This pooling approach is justified [34,36] as the solution of,

$$G_0^g \equiv \arg \min_{G \in \mathbf{G}_K} \sum_{l=1}^s w_l^g D[G || G_l], \quad (4.5)$$

4.2.1.1 Weights assignment

Recall that one of the main challenges in distributional pools (4.2), (4.4) is the optimal design of the weights. There are many approaches for the design of (unequal) weights i.e [64,133–135] in linear pool, and [36,136] in the logarithmic pool. We here review some empirical approaches for the design of weights.

- **The inverse of arithmetic mean of KLDs:** this empirical design i.e. [48], proposes

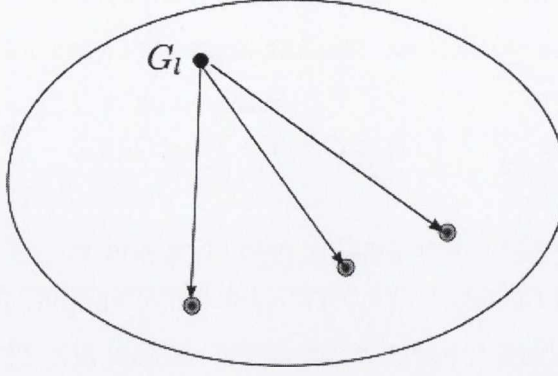


Figure 4.2. Schematic diagram for the inverse of an arithmetic mean of KLDs.

the average KLD from G_l to all the other participant pds,

$$\beta_l \equiv \frac{1}{s} \sum_{j=1}^s D[G_l || G_j]. \quad (4.6)$$

which provides a measure of outlierness of the $\mathcal{I}_l$ among the ‘cluster’ of participant pds. $\mathcal{I}_0$ can empirically assign the weight to $\mathcal{I}_l$, via $\frac{1}{\beta_l} > 0$, $l = 1, \dots, s$. Normalizing this ratio can be interpreted as the $\mathcal{I}_0$ ’s probability of correctness of the $\mathcal{I}_l$, i.e.

$$w_l = \frac{\frac{1}{\beta_l}}{\sum_{j=1}^s \frac{1}{\beta_j}}, \quad l = 1, \dots, s. \quad (4.7)$$

- **The inverse of maximum KLD:** the other KLD-based empirical design of weight i.e. [134] is proportional to the inverse of KLD from G_l to whose $\mathcal{I}_l$ has the maximum distance in KLD sense, i.e.

$$\beta_l^* \equiv D[G_l || G_l^*], \quad (4.8)$$

where, $G_l^* = \operatorname{argmax}_j D[G_l || G_j]$. Here, G_l^* is pd of the furthest participant. Using (4.8), weights are obtained as:

$$w_l^* \equiv \frac{\frac{1}{\beta_l^*}}{\sum_{j=1}^s \frac{1}{\beta_j^*}}, \quad l = 1, \dots, s \quad (4.9)$$

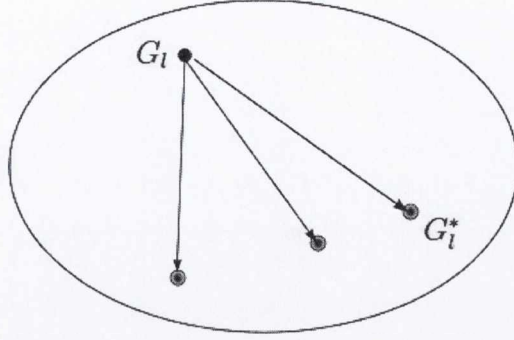


Figure 4.3. Schematic diagram for the inverse of maximum KLDs.

4.2.1.2 Limitations of the distributional pools

Most approaches of the distributional pools focus on (4.2), (4.4) which obviously have different structure. Although, distributional pools enable us process stochastic knowledge in MPS, the following limitations of distributional pools are currently exist:

1. Ambiguity in the structural form of merging the predictive pds, G_{l_i} s, in distributional pools due to the lack of a unified and axiomatic framework.
2. The lack of consistent processing of the generative model of x , (i.e. $G_o(x|\theta)$) available to the deliberator, for the future data. Since in the distributional pools, all pds have the same nature i.e. all are predictors or prior pds. This property of distributional pools does not allow the deliberator share their parametric model of x . Notice that in the literature i.e. [35,36], a parametric model of x is inconsistently adopted for pooling multiple priors of the same parameter. This is inconsistent because the parametric model is externally adopted which does not belong to any node. the deliberator's model does not exist. Instead, $\mathcal{I}_0$'s task is merging pds provided by participants. Therefore, these approaches do not have the capacity of uncertainty, where the processing of future data is impossible due to the lack of the deliberator's parameter model.
3. The lack of a principled design of weights.

We will show that the adaptation of the centralized DM scheme using FPD with a parametric deliberator is going to solve these issues in a successful way.

4.2.2 FPD for distributional design in MPS

Similar to distributional pools, FPD [17] exploits KLD as a ranking rule. However, this is quite different from distributional pools, because one objective function (i.e. one KLD) is minimized in FPD, while in distributional pools, a weighted combination of KLDs with unknown weights is minimized. Also, in FPD, an ideal pd is explicitly and consistently proposed as the origin of the set of parameter pds—which true pds of parameter are ranked with. But there is not such possibility in distributional pools. Recall that a summary of properties of FPD is presented in Section 2.5. FPD is exploited for MPS i.e. [25,37,38,111], [48]. In particular, [37] uses FPD in the centralized DM scheme, designing the deliberator’s parameter prior in the special case of discrete pds (i.e. multinomial pd). Reference [38] proposes a decentralized DM scheme for merging multiple multinomial pds using FPD. In both [37,38], weights assignment is underlying the numerical approximations. Reference [111] adopts a decentralized DM scheme for merging multiple pds using FPD. Reference [25] heuristically adopts FPD for merging the external predictive pds, where again, no formal design of weights is proposed. We note that [25,111] propose a heuristic and informal way of weights assignment.

4.2.2.1 Limitations of available FPD in MPS

The main limitations of the available FPD approaches are as follows.

- available FPD methods do not provide an axiomatic and tractable design of weights in the optimal design of parameter prior.
- FPD has been widely studied in decentralized scheme, however employing FPD yet to be adopted for the centralized scheme is not widely. To authors, the published research on centralized merging via FPD is [37]

We here adopt a parametric deliberator-based model of x —which quantifies the uncertainty in the deliberator node— and employ the centralized DM scheme of distributional design using FPD. The deliberator is a Bayesian moral (i.e. empathetic) decision maker, who bound their own expectation of the expected loss suffered by each participant. This will be fully explained in Section 4.3. The moral deliberator process pds into a FPD-optimal

design of the parameter prior. Recall that in FPD, we employ only one KLD to rank the set of possible parameter priors to an explicit ideal pd of parameter. This provides a consistent and unified structure for merging the stochastic knowledge. We note that this is the extension of the available FPD method with centralized DM scheme i.e. [37] to the continuous case with a finite-dimensional parameter. The FPD-optimal design of parameter prior acts as the prior for the deliberator's generative model of x (i.e. $G_0(x|\boldsymbol{\theta})$). In particular, we employ this optimal parameter prior using Bayesian calculus to deduce the FPD-optimal predictor, $G_0(x|\boldsymbol{\theta})$. This optimal predictive pd is employed for processing future data. The proposed FPD here is the adaptation of the parametric hierarchical FPD item 1 in Remark 2.3.2. Notice that this is the specialization of the (nonparametric) hierarchical FPD optimal design [19] to the parametric deliberator model. Also, this is the first time which we apply this for continuous pds in the centralized scheme for MPS.

4.3 Knowledge definition and modelling

We propose a system involving $s + 1$ nodes ($s \geq 1$), s of them (i.e. $\mathcal{I}_1, \dots, \mathcal{I}_s$) are *participants*. Each participant independently contributes a known pd of the same x , i.e. $G_l(x) \equiv G_l$, $l = 1, \dots, s$. The extra (i.e. $(s + 1)$ st) node is the *deliberator*, $\mathcal{I}_0$ —receive these pds from the s participants, and also contribute their own parametric pd of x , i.e. $G_0(x|\boldsymbol{\theta}, K) \equiv G_0$, conditioned on their aggregate knowledge, K . Here, $\boldsymbol{\theta} \in \Theta$ is $\mathcal{I}_0$'s finite-dimensional, unknown parameter, modelled hierarchically by $\mathcal{I}_0$, as defined next. We note that $x|G_0 \stackrel{d}{=} x|\boldsymbol{\theta}$, where $\stackrel{d}{=}$ denotes equality in pd. $\mathcal{I}_0$ adopt FPD methodology to merge G_l , $l = 1, \dots, s$ and $G_0(x|\boldsymbol{\theta}, K) \equiv G_0$, leading to an optimal design of the $\mathcal{I}_0$'s parameter prior. This prior can then be used to express their uncertainty in the parametric model. Also it will be used to predict x , as explained in the sequel. $\mathcal{I}_0$'s knowledge, K , is specified below. Hereinafter, we will suppress the notational dependence on K of $\mathcal{I}_0$'s model factors, e.g. $G_0(x|\boldsymbol{\theta}, K) \equiv G_0(x|\boldsymbol{\theta})$. In the context of this Chapter, participants are forced to approximate their pd, $G_l(x)$, with $G_0(x|\boldsymbol{\theta})$ imposed by $\mathcal{I}_0$, and so their expected loss is $D[G_l||G_0(\cdot|\boldsymbol{\theta})]$. As a *moral* deliberator, $\mathcal{I}_0$ impose an upper bound on the expected loss related to each participant, which will be presented in item 3 of the knowledge K defined as follows:

Definition 4.3.1. (*Knowledge, K , processed by the parametric moral deliberator, $\mathcal{I}_0$, in FPD*). $\mathcal{I}_0$, and $\mathcal{I}_l, l = 1, \dots, s$, model a common unknown quantity, x . $\mathcal{I}_0$'s knowledge, K , comprises the following elements;

1. a known parametric form, $G_0(x|\boldsymbol{\theta})$, of model for x , conditioned on the unknown, finite-dimensional parameter, $\boldsymbol{\theta}$.
2. known pds, $G_l(x)$, $l = 1, \dots, s$, reported independently to $\mathcal{I}_0$ by the s participants.
3. $\mathcal{I}_0$'s parameter pd, $A(\boldsymbol{\theta}|K)$, confined to the following set:

$$A \in \mathbb{A}_K \equiv \{A(\boldsymbol{\theta}|K) : E_{A(\boldsymbol{\theta}|K)} [D[G_l||G_0(\cdot|\boldsymbol{\theta})]] \leq \beta_l, l = 1, \dots, s\} \neq \emptyset. \quad (4.10)$$

Here, $\beta_l > 0$ bounds $\mathcal{I}_0$'s expectation of the expected loss suffered by the l^{th} -participant when $G_0(x|\boldsymbol{\theta})$ is imposed on them. A lower choice of β_l expresses $\mathcal{I}_0$'s greater confidence in the l^{th} -participant. $\boldsymbol{\beta} \equiv \{\beta_1, \dots, \beta_s\}^T$ is known to $\mathcal{I}_0$, and therefore forms part of their knowledge.

In the following, $\mathcal{I}_0$'s model of $(x, \boldsymbol{\theta})$, conditioning on available K , is defined as follows.

Definition 4.3.2 ($\mathcal{I}_0$'s joint model of unknowns). Let M be $\mathcal{I}_0$'s joint pd of x and $\boldsymbol{\theta}$, and let A be their (hierarchical) parameter prior pd of $\boldsymbol{\theta}$. Then, using the chain rule:

$$\begin{aligned} M(x, \boldsymbol{\theta}|G_0, A, K) &= M(x|\boldsymbol{\theta}, G_0, A, K)M(\boldsymbol{\theta}|A, K) \\ &= G_0(x|\boldsymbol{\theta})A(\boldsymbol{\theta}|K). \end{aligned} \quad (4.11)$$

$\mathcal{I}_0$ ' specify their joint ideal pd of $(x, \boldsymbol{\theta})$ as follows,

$$M_I(x, \boldsymbol{\theta}|G_0, F_I, K) = G_0(x|\boldsymbol{\theta})F_I(\boldsymbol{\theta}|K). \quad (4.12)$$

Notably, $G_0(x|\boldsymbol{\theta})$ is not the variational pd in (4.11) and (4.12). Also, the pd of x in the joint true and ideal models is the same. Therefore, the only variational quantity in the joint model (4.11) is $A(\boldsymbol{\theta}|K)$.

No unique solution for $M(\cdot)$ (4.11), in particular $A(\cdot)$ is available. Since, $M(\cdot)$ is incomplete model. Instead, $\mathcal{I}_0$, ranks all possible parameter pds, $A(\cdot)$, consistent with their

knowledge, K (Definition 4.3.1), to ideal parameter pd, $F_I(\cdot)$, using FPD. The definition of the FPD principle (2.14) for processing K in Definition 4.3.1, into an optimal design of $A(\boldsymbol{\theta}|K)$ —being the only free factor in $\mathcal{I}_0$'s model (4.11)—is therefore as follows:

Definition 4.3.3 (The FPD principle for the parameter pd). *Using (2.14), the FPD-optimal design of $\mathcal{I}_0$'s parameter pd, $A(\boldsymbol{\theta}|K)$, is defined as*

$$A^\circ(\boldsymbol{\theta}|K) \equiv \arg \min_{A \in \mathbb{A}_K} D[A||F_I], \quad (4.13)$$

where F_I is the ideal parameter pd and $\mathbb{A}_K$ is the variational space of $A(\cdot)$, constrained by K (Definition 4.3.1).

$\mathcal{I}_0$'s parametric model (Definition 4.3.2) is hierarchical, via the parameter prior pd, $A(\boldsymbol{\theta}|K)$. However, it is not a fully relaxed, nonparametric hierarchy, in that the deliberator assumes knowledge of the parametric form of their model for x , i.e. of $G_0(x|\boldsymbol{\theta})$ (item 1 of K , in Definition 4.3.1). This allows x to be continuous—extending the contexts in [37,38], while avoiding the technical difficulties of optimally designing a nonparametric prior pd. FPD for a nonparametric prior is addressed in [19], but not in a multiple participant context. The following lemma proposes the FPD-optimal design of the parameter pd, i.e. $A^\circ(\boldsymbol{\theta}|K)$, in Definition 4.3.3, when processing knowledge, K , specified in Definition 4.3.1.

Lemma 4.3.1 ($\mathcal{I}_0$'s FPD-optimal design of the parameter pd). *The solution of (4.13), when processing K (Definition 4.3.1), and underlying the regularity conditions,*

$$D[G_l||G_0(\cdot|\boldsymbol{\theta})] < \infty, \quad l = 1, \dots, s, \quad (4.14)$$

is as follows:

$$A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K) \propto F_I(\boldsymbol{\theta}|K) \exp \left\{ - \sum_{l=1}^s \lambda_l D[G_l||G_0(\cdot|\boldsymbol{\theta})] \right\}. \quad (4.15)$$

$\boldsymbol{\lambda} \equiv \boldsymbol{\lambda}(\boldsymbol{\beta})$ is a vector of non-negative Kuhn-Tucker multipliers, λ_l , $l = 1, \dots, s$.

Proof. $D[A||F_I]$ (4.13) is a convex function of variational A in the convex set, $\mathbb{A}_K$ (4.10). Therefore, the minimizer exists, and it minimizes the following Kuhn-Tucker functional:

$$\begin{aligned}
\mathcal{L}(A, F_I, \boldsymbol{\lambda}) &\equiv \mathcal{L}(A, \boldsymbol{\lambda}) \\
&= \int A(\boldsymbol{\theta}|K) \log \frac{A(\boldsymbol{\theta}|K)}{F_I(\boldsymbol{\theta}|K)} d\boldsymbol{\theta} + \sum_{l=1}^s \lambda_l \{E_{A(\boldsymbol{\theta}|K)} [D[G_l||G_0(\cdot|\boldsymbol{\theta})]] - \beta_l\} \\
&= \int A(\boldsymbol{\theta}|K) \log \left(\frac{A(\boldsymbol{\theta}|K)}{\frac{1}{c(\boldsymbol{\lambda})} F_I(\boldsymbol{\theta}|K) \exp \{-\sum_{l=1}^s \lambda_l D[G_l||G_0(\cdot|\boldsymbol{\theta})]\}} \right) d\boldsymbol{\theta} - \log c(\boldsymbol{\lambda}) - \sum_{l=1}^s \lambda_l \beta_l \\
&= D[A||A^o] - \log(c(\boldsymbol{\lambda})) - \sum_{l=1}^s \lambda_l \beta_l,
\end{aligned} \tag{4.16}$$

where $A^o(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$ is given by (4.15), and

$$c(\boldsymbol{\lambda}) = \int F_I(\boldsymbol{\theta}|K) \exp \left\{ -\sum_{l=1}^s \lambda_l D[G_l||G_0(\cdot|\boldsymbol{\theta})] \right\} d\boldsymbol{\theta} < \infty \tag{4.17}$$

normalizes this pd. Using the properties of KLD, $A^o(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$ is therefore the unique FPD minimizer defined in (4.13). Using the *complementary slackness* [137] property of convex optimization, it follows that

$$\lambda_l C_l(\boldsymbol{\lambda}) = 0, \quad l = 1, \dots, s, \tag{4.18}$$

where $C_l(\boldsymbol{\lambda}) \equiv E_{A^o(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)} [D[G_l||G_0(\cdot|\boldsymbol{\theta})]] - \beta_l$. Complementary slackness guarantees that the Kuhn-Tucker multipliers, $\boldsymbol{\lambda}(\boldsymbol{\beta})$, are unique, with either $\lambda_l(\boldsymbol{\beta}) > 0$, s.th $C_l(\boldsymbol{\lambda}) = 0$ (i.e. those cases where the constraint is active), or else $\lambda_l(\boldsymbol{\beta}) = 0$, s.th $C_l(\boldsymbol{\lambda}) < 0$ (inactive constraint). $\square$

The key computational task implied by Lemma 4.3.1 is the evaluation of the optimal Kuhn-Tucker multipliers. More specifically, the processing of the specified bounds, $\boldsymbol{\beta}$ (item 3 in Definition 4.3.1) via the unique, deterministic mapping, $\boldsymbol{\beta} \rightarrow \boldsymbol{\lambda}$ is the computational challenge. This is highly intractable, but the following remark specifies the general procedure.

Remark 4.3.1. (*Computation of $\boldsymbol{\lambda} \equiv \boldsymbol{\lambda}(\boldsymbol{\beta})$*). Following Definition 4.3.3, the optimal

Kuhn-Tucker multipliers—denoted by λ^o are:

$$\begin{aligned}\lambda^o &\equiv \operatorname{argmin}_{\lambda} D[A^o||F_I] \text{ subject (by (4.10)) to} \\ E_{A^o(\theta|\lambda, K)}[D(G_l||G_0(\cdot|\theta))] &\leq \beta_l, \quad l = 1, \dots, s.\end{aligned}\tag{4.19}$$

Equivalently, this requires the unique solution, in λ , of the following system of non-linear equations:

$$\begin{aligned}\frac{\partial}{\partial \lambda_l} E_{A^o(\theta|\lambda, K)}[\log \frac{A^o(\theta|\lambda, K)}{F_I(\theta|K)} + \sum_{j=1}^s \lambda_j \{D[G_j||G_0(\cdot|\theta)] - \beta_j\}] &= 0, \\ \frac{\partial}{\partial \lambda_l} [-\log(c(\lambda)) - \sum_{j=1}^s \lambda_j \beta_j] &= 0, \quad l = 1, \dots, s.\end{aligned}\tag{4.20}$$

We note that solving (4.20) generally does not lead to an explicit solution of λ^o (except for some special cases). Therefore, numerical techniques is required i.e. [38].

A further challenge is the requirement for $\mathcal{I}_0$ to specify upper bounds, β , on expected KLDs, as part of their knowledge, K (Definition 4.3.1). Recall that (4.6) provides the empirical design of β . Using that, $\mathcal{I}_0$ respect more (smaller β_l) those participants which cluster well among the set of participants. Also (4.7) can be used as an empirical design of λ . Notice that using (4.7), we no longer need to employ Remark 4.3.1.

4.3.1 FPD-optimal design of the deliberator's predictor

In many MPS, the aim is to construct an optimal predictor, $G_0^o(x|K)$, after merging the participants' knowledge [64, 138]. The following remark adopts standard Bayesian calculus to construct this predictor, consistent with the FPD-optimal parameter prior (4.15).

Remark 4.3.2. (*$\mathcal{I}_0$'s FPD-optimal predictive pd*) Adopting $A = A^o$ (4.15) in $\mathcal{I}_0$'s model (4.11), then $\mathcal{I}_0$'s FPD-optimal predictor, G_0^o , is deduced by marginalization over θ . It follows that

$$G_0^o(x|\lambda, K) = E_{A^o(\theta|\lambda, K)}[G_0(x|\theta, K)]\tag{4.21}$$

Figure 4.4 shows the schematic diagram for processing K in Definition 4.3.1. Then, $\mathcal{I}_0$'s optimal predictor is deduced using Bayes' rule and applying marginalization.

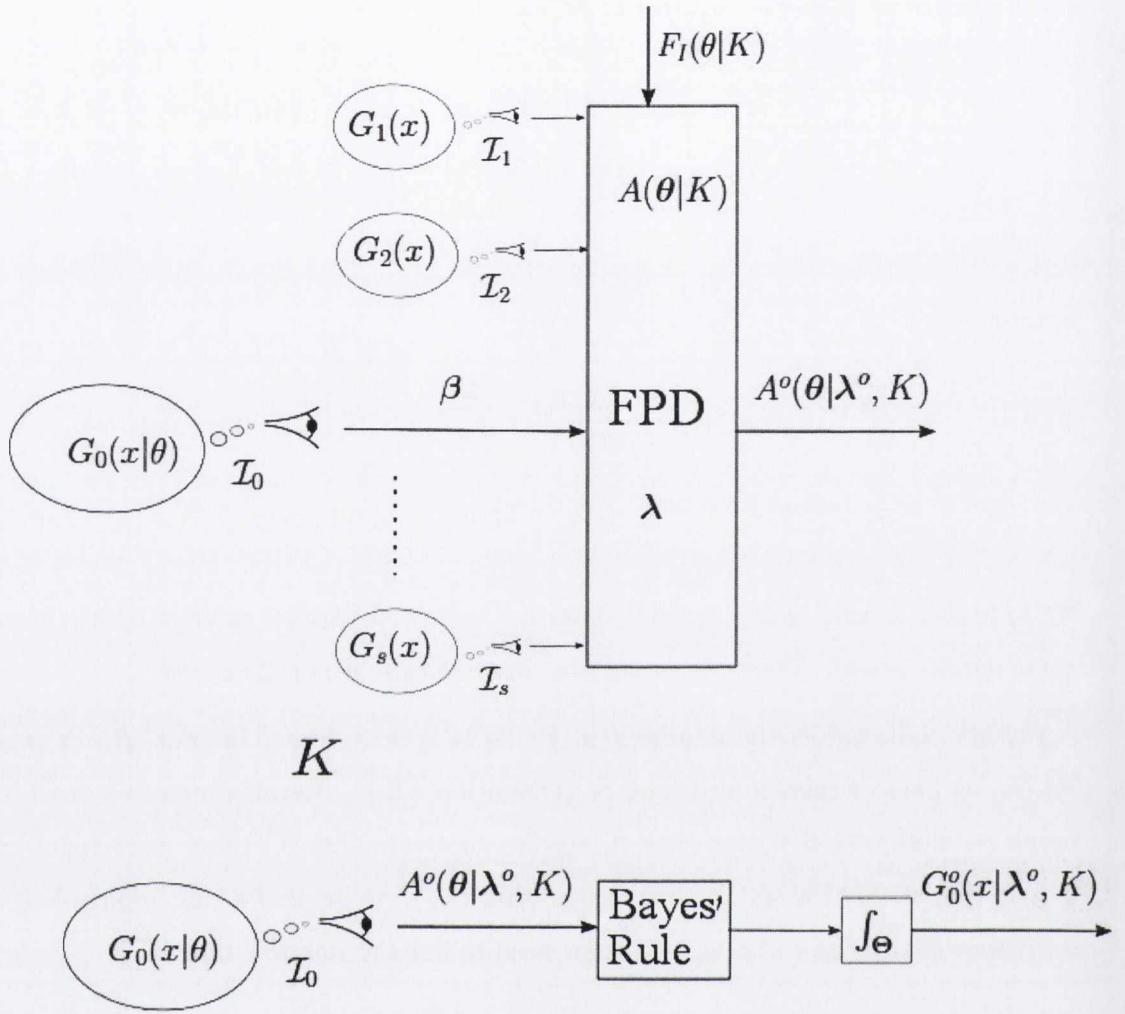


Figure 4.4. Schematic diagram for processing K (Definition 4.3.1) into the FPD-optimal design of the moral delibrator (4.15), then deducing their optimal predictor. Here, $\int_{\Theta}$ denotes the marginalization operator.

There is no uncertainty associated with the functional form of the FPD-optimal predictor (4.21). This is because $\mathcal{I}_0$'s hierarchy (Definition 4.3.2) is parametric, with known functional forms. If an *unknown* nonparametric model (2.18), had instead been adopted *a priori*, the FPD-optimal prior would necessarily be nonparametric, furnishing predictive uncertainty. This will be addressed in future work, using Definition 2.3.3, as an established theory of nonparametric, hierarchical FPD developed in [19]. The FPD-optimal parameter prior pd (4.15) furnishes uncertainty in respect of $\mathcal{I}_0$'s parametric model, $x \sim G_0(x|\theta)$. An optimal choice, $G_0(x|\hat{\theta})$, requires a criterion for estimation of $\theta \sim A^o(\theta|\lambda, K)$ (4.15). The

expectation, i.e.

$$\hat{\boldsymbol{\theta}} \equiv E_{A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)}[\boldsymbol{\theta}], \quad (4.22)$$

is justified as the minimum Bayes' risk estimate under many loss function choices [14, 61]. More importantly, $A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$ (4.15), allows to exploit (Monte Carlo) randomization [61], yielding the random sample, $\boldsymbol{\theta}^{(i)} \sim A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$, $i = 1, 2, \dots$. Then using (2.34), we can approximate (4.22) or (4.21)—to arbitrary accuracy—the integral in (4.21) with a simple sum, and eliminates the need for the (potentially difficult) optimization yielding the estimate, $\hat{\boldsymbol{\theta}}$ or G_0° . Recall that pooling approaches (4.2) and (4.4) do not provide this strategy for evaluation of (4.21). In the following, we wish to compare the structure of our predictor (4.21) to the logarithmic pooling prediction.

Remark 4.3.3. (*Relationship to the logarithmic pool*). *The FPD-optimal design, $A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$ (4.15), involves a geometric mean of participant-indexed Gibbs-type pds [19], with weights, λ_l . It can be seen as an FPD-optimal parametric extension of the logarithmic pool [34] of the participants' contributed pds given in (4.4). FPD replaces this with (4.21), i.e. the geometric mean is replaced by an optimized infinite mixture of parametric observation models.*

A special case of Lemma 4.3.1 arises when x is finite-state (discrete) and $G_0(x|\boldsymbol{\theta})$ (4.11) is multinomial, with $\boldsymbol{\theta} = \mathbf{p}$, as explained in the following Corollary.

Corollary 4.3.1 (of Lemma 4.3.1: quantization of x). *Let $x \in \mathbb{X} \equiv \{v_1, \dots, v_m\}$, then*

$$\begin{aligned} G_0(x|\boldsymbol{\theta}) &\equiv G_0(x|\mathbf{p}) \equiv Mu_x(1, \mathbf{p}) \equiv \mathbf{p}, \\ \mathbf{p} &\equiv \{p_k, \ k = 1, \dots, m\} \in \Delta_{m-1}, \end{aligned} \quad (4.23)$$

where Mu denotes the multinomial probability mass function (pmf) and Δ_{m-1} denotes the open probability simplex of dimension $m - 1$ [139]. Also,

$$G_l \equiv \mathbf{p}_l \equiv \{p_{l,k}, \ k = 1, \dots, m\} \in \Delta_{m-1}, \ l = 1, \dots, s, \quad (4.24)$$

where $\mathbf{p}_l$ is defined on the same set, $\mathbb{X}$, and $p_{l,k} \equiv \Pr_l[x = v_k]$ are known. Adopting these

assignments in (4.15), the FPD-optimal design of $A(\mathbf{p}|\boldsymbol{\lambda}, K)$ is

$$\begin{aligned}
A^\circ(\mathbf{p}|\boldsymbol{\lambda}, K) &\propto F_I(\mathbf{p}|K) \exp \left\{ - \sum_{l=1}^s \lambda_l D[\mathbf{p}_l || \mathbf{p}] \right\} \\
&= F_I(\mathbf{p}|K) \prod_{k=1}^M \exp \left\{ - \sum_{l=1}^s \lambda_l p_{l,k} \log \frac{p_{l,k}}{p_k} \right\} \\
&\propto F_I(\mathbf{p}|K) \prod_{k=1}^M p_k^{\sum_{l=1}^s \lambda_l p_{l,k}}.
\end{aligned} \tag{4.25}$$

$A^\circ(\mathbf{p}|\boldsymbol{\lambda}, K)$ (4.25) is a Dirichlet pd [55,61] if $F_I(\mathbf{p}|K)$ is Dirichlet, a result which was derived in [37] (using Kerridge inaccuracy instead of KLD) and [38]. Consider $\mathcal{I}_o$'s parametric observation model, $G_0(x|\boldsymbol{\theta})$ (4.11), under a quantization operator [139], $Q : \mathbb{R} \rightarrow \mathbb{X}$:

$$G_0(x|\boldsymbol{\theta}) \xrightarrow{Q} Mu_x(1, \mathbf{p}) \tag{4.26}$$

i.e. the pre-image of $\mathbb{X}$ under Q is a partition of $\mathbb{R}$. Therefore (4.25) specializes the FPD-optimal design (4.15) under Q , and consistent with the parameter transformation in (4.26). Therefore, (4.15) can be understood as the extension of the quantized design—available in [37]—to FPD for non-quantized parametric models.

4.3.2 Functionally tractable FPD: exponential family

In this part, we are interested in finite-dimensional representations of the FPD-optimal design of parameter pd, $A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$ (4.15). An important such case arises when A° is a member of the *conjugate exponential family* (CEF) [61]:

$$\begin{aligned}
A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K) &\equiv A^\circ(\boldsymbol{\theta}|\mathbf{V}_\lambda, \nu_\lambda) \\
&\propto [a(\boldsymbol{\theta})]^{\nu_\lambda} \exp[\langle \phi(\boldsymbol{\theta}), \mathbf{V}_\lambda \rangle].
\end{aligned} \tag{4.27}$$

Here, the shaping parameters, $\{\mathbf{V}_\lambda, \nu_\lambda\}$, are fixed, finite-dimensional functions of $\boldsymbol{\lambda}$, and therefore, of the knowledge objects in K (Definition 4.3.1). These are combined with kernels, $a(\boldsymbol{\theta})$ and $\phi(\boldsymbol{\theta})$, and via an appropriate scalar product, $\langle \cdot, \cdot \rangle$, in the manner specified in (4.27), and all quantities have compatible dimensions. The form (4.27) is very important for sequential processing of data, as expressed in Chapter 3. If adopted by the deliberator,

then *future* sequential observations, $x_i \stackrel{iid}{\sim} G_0(x_i|\boldsymbol{\theta})$, $i = 1, 2, \dots$, are projected invariantly into finite-dimensional statistics if $G_0(\cdot)$ has the associated exponential family form [61]. The following lemma provides sufficient conditions under which the FPD-optimal design, $A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$, achieves the (CEF) form (4.27).

Lemma 4.3.2 (CEF-FPD-optimal design of $A(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$). *Sufficient conditions under which the FPD-optimal parameter prior (4.15) achieves the conjugate exponential family (CEF) form (4.27) are as follows:*

1. *The deliberator's parametric pd, $G_0(x|\boldsymbol{\theta})$ (Definition 4.3.1), is an exponential family member, as follows:*

$$G_0(x|\boldsymbol{\theta}) \equiv a(\boldsymbol{\theta})b(x) \exp[\langle \phi(\boldsymbol{\theta}), d(x) \rangle]. \quad (4.28)$$

Also, $E_{G_l}[\log(b(x))] < \infty$ and $E_{G_l}[d(x)] < \infty$, $l = 1, \dots, s$, where the G_l are the participants' pds (item 2 of K , in Definition 4.3.1).

2. *$F_I(\boldsymbol{\theta}|K)$ (Definition 4.3.1) is chosen conjugate to $G_0(x|\boldsymbol{\theta})$, i.e. with the CEF form (4.27).*

Proof. By definition of KLD (2.2.1),

$$D[G_l||G_0(\cdot|\boldsymbol{\theta})] = E_{G_l}[\log G_l(x)] - E_{G_l}[\log G_0(x|\boldsymbol{\theta})]. \quad (4.29)$$

Substituting (4.28) into the final term (4.29), and noting that E_{G_l} is distributive across $\langle \cdot, \cdot \rangle$:

$$\begin{aligned} D[G_l||G_0(\cdot|\boldsymbol{\theta})] &= E_{G_l}[\log G_l] - E_{G_l}[\log(b(x))] \\ &\quad - \log[a(\boldsymbol{\theta})] - \langle \phi(\boldsymbol{\theta}), E_{G_l}[d(x)] \rangle. \end{aligned} \quad (4.30)$$

Substituting (4.30), $\forall l$, into the FPD-optimal parameter prior (4.15):

$$A^\circ(\boldsymbol{\theta}|\boldsymbol{\lambda}, K) = \frac{1}{c(\boldsymbol{\lambda})} F_I(\boldsymbol{\theta}|K)[a(\boldsymbol{\theta})]^{(\sum_{l=1}^s \lambda_l)} \exp \left\{ \langle \phi(\boldsymbol{\theta}), \sum_{l=1}^s \lambda_l E_{G_l}[d(x)] \rangle \right\}. \quad (4.31)$$

Imposing the condition 2 above, then (4.31) has the CEF form (4.27), with shaping

parameters,

$$\mathbf{V}_\lambda = \mathbf{V}^I + \sum_{l=1}^s \lambda_l E_{G_l}[d(x)], \quad (4.32)$$

$$\nu_\lambda = \nu^I + \sum_{l=1}^s \lambda_l, \quad (4.33)$$

where $\{\mathbf{V}^I, \nu^I\}$ are the shaping parameters of $F_I(\boldsymbol{\theta}|K)$.

□

Notice that the CEF-FPD form (4.31) suggests a class of $\mathcal{I}$'s parametric model where the FPD-optimal design (4.15) is functionally tractable. Also, Remark 4.3.1 proposes a guideline for processing K into the Kuhn-Tucker multipliers, $\boldsymbol{\lambda} \equiv \boldsymbol{\lambda}(\boldsymbol{\beta})$ (4.15), by FPD. Nevertheless, in general, $\boldsymbol{\lambda}$ are computationally intractable due to the lack of explicit mapping function. In the following we provide a special case where $\boldsymbol{\lambda}$ are computationally tractable.

4.3.2.1 An illustrative example: the normal case

In this part, we aim to implement the FPD-optimal predictor (4.21) into the normal example. Let $\mathcal{I}_l$, $l = 1, \dots, s$ be normal participants, each modelling $x \in \mathbb{R}$:

$$G_l(x|\boldsymbol{\eta}_l) \equiv N_x(\mu_l, r_l), \quad l = 1, \dots, s, \quad (4.34)$$

where all the participant-specific shaping parameters, $\boldsymbol{\eta}_l = (\mu_l, r_l)$, $l = 1, \dots, s$, are known. $\mathcal{I}_0$ adopts

$$G_0(x|\boldsymbol{\theta}) \equiv G_0(x|a) \equiv N_x(a, r), \quad (4.35)$$

where $\boldsymbol{\theta} \equiv a$ is an unknown mean and r is a *known* variance. Assuming $F_I(a|K)$ be a diffuse Normal, then the conditions (Lemma 4.3.2) for achieving CEF form of the deliberator's FPD-optimal prior design are satisfied. In this case, (4.31) specializes to the following:

$$\begin{aligned} A^o(a|K) &= N_a(\mu^o, r^o), \\ \mu^o &= \frac{\sum_{l=1}^s \lambda_l \mu_l}{\sum_{j=1}^s \lambda_j}, \\ r^o &= \frac{r}{\sum_{j=1}^s \lambda_j}. \end{aligned} \quad (4.36)$$

Notice that the optimal Kuhn-Tucker multipliers, $\boldsymbol{\lambda}^o$ is computed as follows. Substituting (4.36) into the constraints (4.19) (i.e. into the nonlinear system (4.20)), and summing over l , leads to

$$\boldsymbol{\lambda}^T \{\mathbf{b} - \mathbf{A}\boldsymbol{\lambda}\} = 0, \quad (4.37)$$

where

$$\mathbf{b} \equiv r s \mathbf{1}_s$$

(here, $\mathbf{1}_s$ is a column vector of 1's, of length s), and

$$\begin{aligned} \mathbf{A} \equiv & \left(2 \sum_{l=1}^s \mu_l \right) \mathbf{1}_s \boldsymbol{\mu}^T - s \boldsymbol{\mu} \boldsymbol{\mu}^T \\ & + \mathbf{1}_s \mathbf{1}_s^T \sum_{l=1}^s \left(r_l + \mu_l^2 + r \log \frac{r}{r_l} - r - 2r\beta_l \right). \end{aligned} \quad (4.38)$$

Using complementary slackness (4.18), the optimal Kuhn-Tucker multipliers, $\boldsymbol{\lambda}^o$, are deduced tractably as:

$$\boldsymbol{\lambda}^o = \max\{\mathbf{0}_s, \mathbf{A}^{-1}\mathbf{b}\}, \quad (4.39)$$

where max is applied elementwise. Finally, adopting (4.36) as the optimized prior and using $G_0(x|a) \equiv N_x(a, r)$ to (4.21), $\mathcal{I}_o$'s FPD-optimal predictor is

$$G_0^o(x|K) = N_x(\mu^o, r^o + r) \quad (4.40)$$

4.4 Comparative Performance Evaluation under Simulation

We now examine the form and behaviour of the deliberator's FPD-optimal design of parameter prior pd (4.15) in simulation.

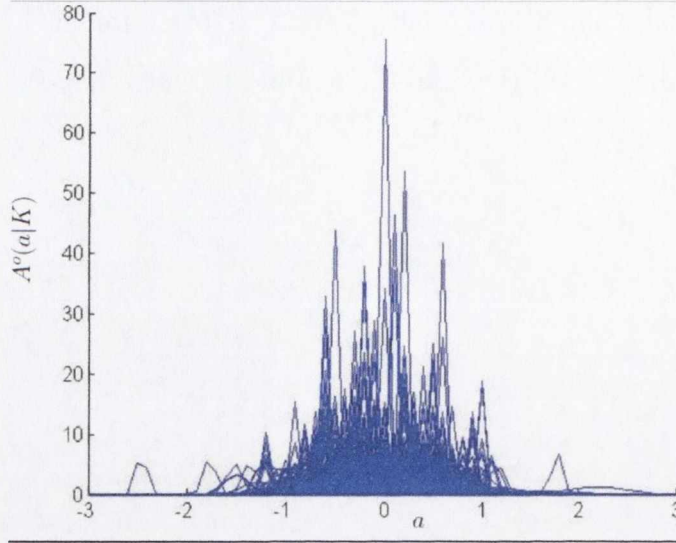


Figure 4.5. $A^o(a|K)$ for $s = 20$ with $MC = 1000$ runs. Normal case with $r = 1$.

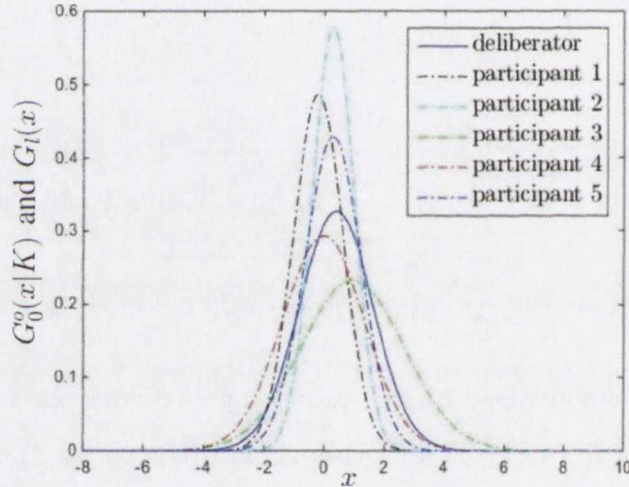


Figure 4.6. $\mathcal{I}_0$'s predictor, $G_0^o(x|K)$ (4.40), compared with the participants' pds, $G_l(x)$, for $s = 5$, with shaping parameters averaged across $MC = 100$ runs in all cases. Normal case with $r = 1$.

4.4.1 The hierarchical property of the FPD-optimal design (4.15)

For computational convenience, we consider the normal deliberator/participants case of Section 4.3.2.1. We simulate the normal participants' pds, independently drawing $r_l \sim iG(2, 1)$ (where iG denotes the *inverse Gamma* pd), and $\mu_l|r_l \sim N(0, r_l)$, $l = 1, \dots, s = 20$, in each of $MC = 1000$ runs. We design the simulation with $\mathcal{I}_0$'s variance being $r = 1$. In Figure 4.5, we plot $\mathcal{I}_0$'s FPD-optimal design, $A^o(a|K)$ (4.36), for each of these runs.

$\mathcal{I}_0$'s FPD-optimal predictor, $G_0^o(x|\boldsymbol{\lambda}, K)$ (4.40), is then computed in the case where

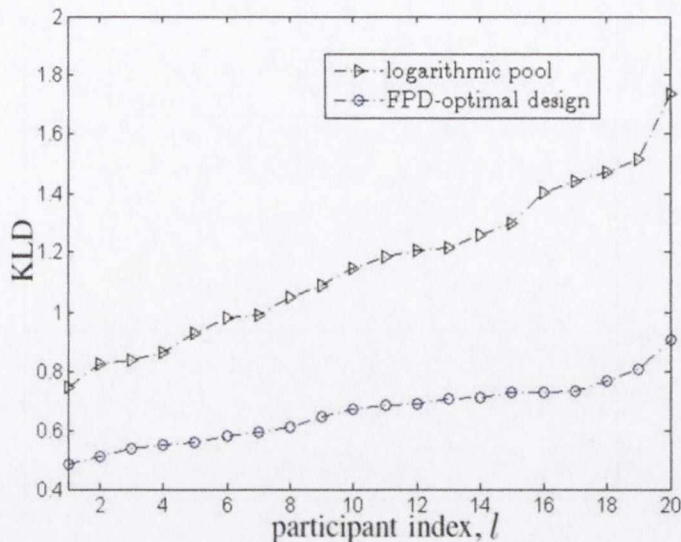


Figure 4.7. Comparison of $\mathcal{I}_0$'s predictor, $G_0^o(x|K)$ (4.40), and the logarithmic pool, $G_0^g(x)$ (4.4), in terms of the participants' expected loss (i.e. $D[G_l||G_0^o]$ and $D[G_l||G_0^g]$). The $s = 20$ participants are in the order of increasing β_l . Normal case with $r = 1$.

$s = 5$, and for each of $MC = 100$ runs. In Figure 4.6, we illustrate the predictor with shaping parameters averaged across the 100 runs. For comparison, we also plot the participants' contributed pds, $G_l(x)$, $l = 1, \dots, 5$, for the averaged shaping parameters of the respective participants.

4.4.2 Optimal design of the Kuhn-Tucker multipliers, λ

Next, we simulate the FPD-optimal design of the Kuhn-Tucker multipliers (weights), λ^o (4.39), associated with the participants, in the tractable normal case (Section 4.3.2.1). We manage the simulations in the same way as in Section 4.4.1, and use KLD as a measure of performance. In Figure 4.7, we plot both $D[G_l||G_0^o]$ and $D[G_l||G_0^g]$, where $G_0^g(x)$ is the logarithmic pooling predictor design (4.4) with uniform weights. We plot these quantities against the participants' index, $l = 1, \dots, s$, ordering the latter from lower to higher values of β_l (4.10) (higher to lower confidence of the deliberator in the respective participants). We find that those participants with lower β_l s have lower KLD, as the moral deliberator expects. Furthermore, $\mathcal{I}_0$'s FPD-optimal predictor design (4.40) outperforms the uniformly-weighted logarithmic pool design (4.4).

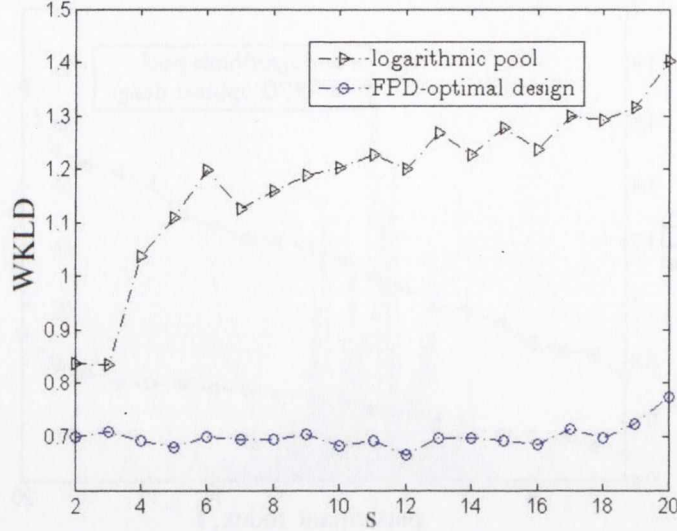


Figure 4.8. Comparison of $\mathcal{I}_0$'s $G_0^o(x|\boldsymbol{\lambda}, K)$ (4.40) and $G_0^g(x)$ (4.4) in terms of weighted KLD (4.41), as a function of the number of participants, s . Normal case with $r = 1$ and $w_l = \frac{1}{s}$.

In Figures 4.8 and 4.9, we illustrate the weighted KLD (WKLD), defined as follows:

$$WKLD = \sum_{l=1}^s w_l D[G_l || \cdot]. \quad (4.41)$$

In Figure 4.8, uniform weights, $w_l = \frac{1}{s}$, are used in (4.41) to compare $\mathcal{I}_0$'s FPD-optimal predictor design to the logarithmic pool, as a function of the number of participants, s (4.4). This figure shows that the logarithmic pool has an increasing trend (i.e. diminishing performance), while $\mathcal{I}_0$'s FPD-optimal design does not. Furthermore, $\mathcal{I}_0$'s FPD-optimal design outperforms the logarithmic pool for all values of s . In Figure 4.9, we use $w_l = \frac{\frac{1}{\beta_l}}{\sum_{j=1}^s \frac{1}{\beta_j}}$ (4.7) in (4.41). In this case, $WKLD$ increases with s for the logarithmic pooling predictor design, but decreases for the FPD-optimal design. The FPD-optimal design outperforms the logarithmic pool for $s \geq 5$.

4.5 Discussion

The simulation results of Section 4.4 demonstrate that the FPD-optimal predictor, $G_0^o(x|\boldsymbol{\lambda}, K)$ (4.21), outperforms logarithmic pooling for a range of operating conditions. Furthermore, the FPD-optimal parameter pd (4.15) can be arranged to belong to the conjugate exponential family, yielding functionally tractable algorithms for sequential processing of dynamic data. Lemma 4.3.2 restricts the class to which $G_0(x|\boldsymbol{\theta})$ must belong, if this is to

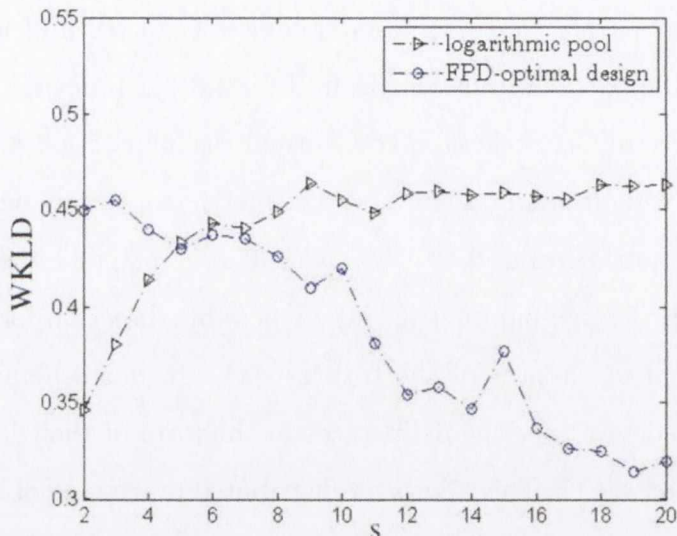


Figure 4.9. Comparison of $\mathcal{I}_0$'s $G_0^o(x|\boldsymbol{\lambda}, K)$ (4.40) and $G_0^g(x)$ (4.4) in terms of weighted KLD, as a function of the number of participants, s . Normal case with $r = 1$ and $w_l = \frac{1}{\sum_{j=1}^s \frac{1}{\beta_j}}$.

be guaranteed. In the next Chapter, we will explore richer classes which also yield this property, in particular mixtures of the participants' pds, i.e. $G_0(x|\boldsymbol{\alpha}, \boldsymbol{\theta}) \equiv \sum_{l=1}^s \alpha_l g_l(x)$, where $\boldsymbol{\alpha} \in \Delta_{s-1}$ are probabilistic which will be designed via FPD.

The FPD-optimal design (4.15) is functionally tractable for a rich class of $\mathcal{I}$'s parametric model i.e. (4.28), however they may not be computationally tractable. This is because processing of K , in particular $\boldsymbol{\beta}$ (as the part of knowledge), into $\boldsymbol{\lambda}$ (4.20) is highly nonlinear (Remark 4.3.1). Therefore, numerical manipulations are required. Some possible solutions will be expressed in Chapter 8.

The parametric hierarchical continuous model adopted by the deliberator (which extends the discrete case reported in [37, 38]) enjoys the usual advantage of Bayesian approaches in that uncertainty in the deliberator's knowledge, $G_0(x|\boldsymbol{\theta})$, is quantified, after merging, via the FPD-optimal parameter prior, $A^o(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$ (4.15), while also yielding the optimal, consistent predictor, $G_0^o(x|\boldsymbol{\lambda}, K)$ (Remark 5.3.2). A practical way to report the uncertainty in $G_0(x|\boldsymbol{\theta})$ is via random draws from the FPD-optimal prior: $\boldsymbol{\theta}^{(i)} \sim A^o(\boldsymbol{\theta}|\boldsymbol{\lambda}, K)$, $i = 1, 2, \dots$. The *randomized* deliberator parametric model, $G_0(x|\boldsymbol{\theta}^{(i)})$, $i = 1, 2, \dots$, can then be reported. This also circumvents the problem of intractable normalizing constant of (4.15) (i.e. intractable forms of A^o). Also, as discussed in the comments following Remark 5.3.2, this can provide a convenient alternative to computing an estimate, $\hat{\boldsymbol{\theta}}$, and it replaces the integral in (4.21) with a sum. The FPD-optimal parametric hierarchy of the deliberator

(Lemma 4.3.1) remains sensitive to the model choice, $G_0(x|\theta)$, and forms an important part of $\mathcal{I}_0$'s knowledge, K (see item 1 of Definition 4.3.1). Indeed, the nature of x will often justify the choice of G_0 , such as in the Normal case of Section 4.3.2.1. Nevertheless, important progress will be made when the parametric form can be fully relaxed to a continuous nonparametric form, $\theta \equiv G_0$, so that $x \sim G_0(x)$, where $G_0(x)$ is now a functionally unknown (i.e. nonparametric) pd. A possible direction for exploration of this task will be explained in chapter of the future work. In our scheme, each participant provides their knowledge of x to the deliberator in the form of their local predictor, $g_l(x)$ (see item 3 of Definition 4.3.1). This allows for distributed processing of local data into these respective predictors, leading to major savings in communication with the deliberator. For instance, in linear-Normal modelling of x by each participant, only the shaping parameters (mean and covariance) of the Normal (or, more generally, the Student- t) predictor need be reported to the deliberator. In future work, we will derive the specific solution for autoregressive participants.

While our deliberator-based (centralized) scheme delivers this advantage of distributed processing by local participants, the deliberation (i.e. decision-making) itself is confined to the deliberator, by definition. This may lead to scalability problems [29] in cases where the number of participants, s , is large. There is also concern about vulnerability of the overall system to deliberator node failure [109]. In order to overcome these problems, the participants—particularly in large- s cases—can be grouped, each group being moderated by a local deliberator. The current framework can then be adopted by each group, with each group's deliberator reporting their local FPD-optimal predictor (4.21) to the (global) deliberator, for final deliberation. Recursive FPD schemes such as this allow local interactions between participants to be considered, particularly when local deliberation is fed back to the participants in a closed-loop scheme. Another option is to introduce a group label, $\theta \rightarrow \{\theta, L = l\}$, into the deliberator's hierarchical model, allowing the deliberator to optimize their model for heterogeneous groups of participants. Ultimately, a global model of interaction between the participants and deliberator needs to be considered, as a formal means of achieving a fully distributed deliberation scheme, with local interactions modelled via a Markov random field [140].

The solution presented in this Chapter is a static, open-loop one, involving a single

invariant unknown, x . The framework can be extended to the dynamic case, involving modelling of a discrete-time random process, $x_i, i \in \mathbb{Z}$. In particular, we will show in future work how the deliberator's FPD-optimal inference may be fed back to the participants as a prior for their (conjugate) processing of local data, $x_i^l, l = 1, \dots, s$, achieving an FPD-optimal, closed-loop, Bayesian distributed filtering scheme.

Chapter 5 —

Multiple participant processing: external predictors

In this Chapter, we aim to process the probabilistic knowledge involving multiple participant predictors in the centralized deliberator-based scheme. The knowledge here does not involve the constraint 3 of K in Definition 4.3.1 which makes them quite different. This knowledge is formulated in Section 5.1. Processing K in different ways will require the specification of different ideal pds and consequently leads to different FPD solutions. Therefore, the ideal specification and the FPD solutions in this Chapter are totally different from those presented in the previous Chapter. In particular, here we process the same knowledge (in Definition 5.1.1) in three different ways yielding three different FPD solutions. We are going to distinguish these different FPD solutions as follows:

1. **FPD-optimal merging with augmentation (FPD-A):** in the first way of merging participants' predictors, unknown parameters, θ , are augmented by the label of participant. This yields $\mathcal{I}_0$'s FPD-optimal merging with augmentation which will be presented in Section 5.2.
2. **FPD-sequential merging (FPD-S):** the second way is merging participants' predictors in a sequence. This way yields $\mathcal{I}_0$'s FPD-sequential merging which will be presented in Section 5.3.
3. **FPD-optimal merging with restriction (FPD-R):** the third way is processing of participants' predictors where the knowledge-constrained set is restricted to the linear mixture of participants' predictors. This yields $\mathcal{I}_0$'s FPD-optimal merging with restriction (FPD-R) which will be presented in Section 5.4.

Notably, we specify the ideal pd differently for different ways of processing K in each Section. The first and third ways are one-batch processing (i.e. processing multiple predictors using only one KLD), while the second way is the sequential processing of participants' predictors. Recall that the main limitation of the proposed FPD-moral deliberator in Chapter 4 is the lack of generally tractable design of weights. $\mathcal{I}_0$'s FPD-A as our primary FPD solution, provides a principled design of non-uniform weights which is evaluated tractably and axiomatically for a rich class of $\mathcal{I}_0$'s parametric pd of x . The designed weights derived from the FPD-optimal prior will be used for computing the optimal predictor via Bayesian calculus.

5.1 Knowledge definition and modelling

In this section we define K in the way which is totally different from the knowledge in Definition 4.3.1. In particular, we relax the constraint 3 of Definition 4.3.1 such that the knowledge-constrained set, $\mathbb{A}_K$ is not confined to the bounded expectation of KLDs. Instead, The deliberator believes that one and only one of the participants will be accepted for designing of the prior of θ . We wish to design them optimally. For this reason, let L be an uncertain proposition which is treated as a random variable. The realization $L = l$ implies that we will choose the l^{th} — participant from s possible participants in order to learn about θ . $\mathcal{I}_0$'s augmented set of unknowns is:

$$(x, \theta, l) \in \mathbb{X} \times \Theta \times \{1, \dots, s\}. \quad (5.1)$$

Therefore, $\mathcal{I}_0$'s knowledge is defined as follows.

Definition 5.1.1. ($\mathcal{I}_0$ ' *Knowledge, K , without bounded expectation of KLD*). $\mathcal{I}_0$, and $\mathcal{I}_l, l = 1, \dots, s$, model a common unknown quantity, being the random variable, x . $\mathcal{I}_0$'s knowledge, K , comprises the following elements;

1. a known parametric form, $G_0(x|\theta)$, of their model for x , conditioned on the unknown, finite-dimensional parameter, θ .
2. known pds, $G_l(x)$, $l = 1, \dots, s$, are reported independently by the s participants to $\mathcal{I}_0$.

3. no special constraints are placed on s variational quantities i.e. $A(\boldsymbol{\theta}|l)$, $l = 1, \dots, s$, nor on pmf $P(l) \in \Delta_{s-1}$.

□

$\mathcal{I}_0$'s joint model of augmented set (5.1) is defined as follows.

Definition 5.1.2 ($\mathcal{I}_0$'s joint model of unknowns). Let M be $\mathcal{I}_0$'s joint model of $(x, \boldsymbol{\theta}, l)$. Then, using the chain rule:

$$M(x, \boldsymbol{\theta}, l|A, P, K) = M(x|\boldsymbol{\theta}, l, A, P, K)M(\boldsymbol{\theta}|l, A, P, K)M(l|A, P, K) \quad (5.2)$$

□

After knowing K , $\mathcal{I}_0$ specify their joint ideal pd of the augmented set (5.1) as follows:

$$\begin{aligned} M_I(x, \boldsymbol{\theta}, l|F_I, P_I, K) &\equiv M_I(x|\boldsymbol{\theta}, l, F_I, P_I, K)M_I(\boldsymbol{\theta}|l, F_I, P_I, K)M_I(l|F_I, P_I, K) \\ &\equiv G_0(x|\boldsymbol{\theta})F_I(\boldsymbol{\theta}|l, K)P_I(l|K), \end{aligned} \quad (5.3)$$

where F_I and $P_I \equiv P_I(l|K) \equiv w_{l,I}$ denote the ideal (pre-prior) pds of $\boldsymbol{\theta}$ and l respectively. Also the following assumptions are adopted:

- $\mathcal{I}_0$ replaces their parametric model of x , with l^{th} - participant's predictor, knowing l . It means that $\mathcal{I}_0$ model x independent of $\boldsymbol{\theta}$ conditioning on $G_l(x)$:

$$M(x|\boldsymbol{\theta}, l, A, P, K) \equiv G_l(x). \quad (5.4)$$

- By the usual conditional independence property of hierarchical model A :

$$M(\boldsymbol{\theta}|l, A, P, K) \equiv A(\boldsymbol{\theta}|l, K). \quad (5.5)$$

- By the definition of model P :

$$M(l|A, P, K) \equiv P(l|K) \equiv w_l.$$

Therefore, $\mathcal{I}_0$'s hierarchical model (5.2) becomes:

$$M(x, \boldsymbol{\theta}, l|A, P, K) \equiv G_l(x)A(\boldsymbol{\theta}|l, K)P(l|K) \quad (5.6)$$

We note that the joint model (5.6) involves $s+1$ variational quantities, i.e. s of them are $A_l \equiv A(\boldsymbol{\theta}|l, K)$, $l = 1, \dots, s$, and one of them is $P(l|K)$. For this reason, M is not unique and we use the joint ideal pd (5.3), to rank $M \in \mathbf{M}_K$, using KLD via FPD methodology.

5.2 FPD-optimal merging with augmentation (FPD-A)

In this section, we aim to derive the FPD-A of parameter prior, i.e. A^o , through processing K in Definition 5.1.1. For this reason, we exploit modelling assumptions (5.3) and (5.6) and define FPD principle for the model of unknown set (5.1) as follows.

Definition 5.2.1. (FPD principle for joint parametric pd) *The FPD-optimal design of unknown M — denoted by M^o , when processing knowledge, K (Definition 5.1.1) is given by the following:*

$$M^o \equiv \arg \min_{M \in \mathbf{M}_K} D[M||M_I]. \quad (5.7)$$

Recall that M denote the model of augmented set (5.1). Notably the only variational quantities are A_l and P in (5.7). Notice that (5.7) is a step-wise minimization,

$$A_l^o P^o \equiv \arg \min_{P \in \mathbf{P}_K} \left(\arg \min_{A_l \in \mathbb{A}_K} D[A_l P || F_{l,I} P_I] \right), \quad (5.8)$$

□

The following lemma proves that a tractable FPD-optimal solution of augmented model (5.6) is implied by the knowledge specified in Definition 5.1.1.

Lemma 5.2.1 (FPD for $\mathcal{I}_0$'s hierarchical model of unknown, processing K , in Definition 5.1.1). *The solution of (5.7), when processing K in Definition 5.1.1 underlying*

the regularity assumptions:

$$D[G_l||G_0(.|\boldsymbol{\theta})] < \infty, \quad l = 1, \dots, s, \quad (5.9)$$

is as follows:

$$M^o(x, \boldsymbol{\theta}, l|A^o, P^o, K) = G_l(x)A^o(\boldsymbol{\theta}|l, K)P^o(l|K) \quad (5.10)$$

where,

$$A^o(\boldsymbol{\theta}|l, K) \propto F_l(\boldsymbol{\theta}|l, K) \exp \{-D[G_l||G_0(.|\boldsymbol{\theta})]\}, \quad (5.11)$$

$$c_{A_l^o} \equiv \int_{\Theta} F_l(\boldsymbol{\theta}|l, K) \exp \{-D[G_l||G_0(.|\boldsymbol{\theta})]\} d\boldsymbol{\theta}. \quad (5.12)$$

$$P^o(l|K) = \frac{P_l(l|K)c_{A_l^o}}{\sum_{j=1}^s P_l(j|K)c_{A_j^o}} \equiv w_l^o \quad (5.13)$$

Proof. Substituting (5.3) and (5.6) into (5.7),

$$\begin{aligned} D[M||M_I] &= \sum_{l=1}^s \int_{\mathbb{X} \times \Theta} \left[\ln \frac{G_l(x)A(\boldsymbol{\theta}|l, K)P(l|K)}{G_0(x|\boldsymbol{\theta})F_l(\boldsymbol{\theta}|l, K)P_l(l|K)} \right] G_l(x)A(\boldsymbol{\theta}|l, K)P(l|K) dx d\boldsymbol{\theta} \\ &= \sum_{l=1}^s \int_{\Theta} \left[D[G_l||G_0(.|\boldsymbol{\theta})] + \ln \left(\frac{A(\boldsymbol{\theta}|l, K)P(l|K)}{F_l(\boldsymbol{\theta}|l, K)P_l(l|K)} \right) \right] A(\boldsymbol{\theta}|l, K)P(l|K) d\boldsymbol{\theta} \\ &= \sum_{l=1}^s \int_{\Theta} \left[D[G_l||G_0(.|\boldsymbol{\theta})] + \ln \left(\frac{A(\boldsymbol{\theta}|l, K)}{F_l(\boldsymbol{\theta}|l, K)} \right) + \ln \left(\frac{P(l|K)}{P_l(l|K)} \right) \right] A(\boldsymbol{\theta}|l, K)P(l|K) d\boldsymbol{\theta} \\ &= \sum_{l=1}^s \left\{ \int_{\Theta} \ln \left(\frac{A(\boldsymbol{\theta}|l, K)}{A^o(\boldsymbol{\theta}|l, K)} \right) A(\boldsymbol{\theta}|l, K) d\boldsymbol{\theta} - \ln c_{A_l^o} + \ln \left(\frac{P(l|K)}{P_l(l|K)} \right) \right\} P(l|K) \\ &= \sum_{l=1}^s \left\{ D[A_l||A_l^o] - \ln c_{A_l^o} + \ln \left(\frac{P(l|K)}{P_l(l|K)} \right) \right\} P(l|K) \\ &= \sum_{l=1}^s D[A_l||A_l^o]P(l|K) + \sum_{l=1}^s \left\{ -\ln c_{A_l^o} + \ln \left(\frac{P(l|K)}{P_l(l|K)} \right) \right\} P(l|K) \end{aligned} \quad (5.14)$$

By the properties of KLD, the first term on the (RHS) of (5.14) achieves its minimal value of zero if $A(\boldsymbol{\theta}|l, K) = A^o(\boldsymbol{\theta}|l, K)$, $l = 1, \dots, s$, where A^o is given in (5.11). Therefore, using step-wise minimization, the FPD optimal design of P is achieved by minimizing

$$\sum_{l=1}^s \left\{ -\ln c_{A_l^o} + \ln \left(\frac{P(l|K)}{P_l(l|K)} \right) \right\} P(l|K) = D[P||P^o] - \ln c_{P^o}. \quad (5.15)$$

Therefore (5.14) achieves its minimal value if $P(l|K) = P^o(l|K)$, where P^o is given in (5.13). Inserting (5.14) into (5.7), the FPD-optimal design of M is M^o (5.10). $\square$

Notice that we aim to deduce the FPD-optimal design of $A(\theta|K)$ using M^o (5.10). For this reason, we marginalize M^o (5.10) over l and x as follows.

Remark 5.2.1. (FPD-A of parameter prior.) Using the FPD-A of components (5.11) and weights (5.13), the FPD-A of $A(\theta|K)$ is, by marginalization of M^o (5.10) over l and x ,

$$\begin{aligned} A^o(\theta|K) &= \sum_{l=1}^s \int_{\mathbb{X}} M^o(x, \theta, l | A^o, P^o, K) dx \\ &= \sum_{l=1}^s w_l^o A^o(\theta|l, K), \end{aligned} \quad (5.16)$$

i.e. A^o is the mixing pds. $\square$

Remark 5.2.2. (Interpretation of FPD-A of parameter prior (5.16)) The FPD-A of parameter prior (5.16) might be interpreted as the optimal design of multi-channel model, where the deliberator adopts the different ideal (pre-prior) pds (i.e. $F_I(\theta|l, K)$, $l = 1, \dots, s$). Then the l^{th} ideal pd is updated using the knowledge provided by l^{th} -participant leading to the optimal local prior (5.11). Then $\mathcal{I}_0$'s optimal prior of θ conditioning on K , is tractably achieved using an optimal combination of these s local priors i.e. (5.16).

So far, we have updated multiple ideal pds (i.e. $F_I(\theta|l, K)$, $l = 1, \dots, s$), to the optimal prior pd (i.e. $A^o(\theta|K)$ (5.16)). Notice that A^o (5.16) is the parameter prior pd for sequential processing of data presented in Chapter 3 by $\mathcal{I}_0$. Alternatively, in the following we are interested in applying the optimal parameter prior (5.16) for achieving $\mathcal{I}_0$'s predictor in the optimal way.

5.2.1 FPD-A of the deliberator's predictor

Here, we use the FPD-A of $A(\theta|K)$, (i.e. A^o (5.16)) to derive FPD-A of $\mathcal{I}_0$'s predictor.

Remark 5.2.3. (FPD-A of $\mathcal{I}_0$'s predictor)

Notice that x is generated from $\mathcal{I}_0$'s parametric model, i.e.

$$x_i \stackrel{\text{ciid}}{\sim} G_0(x|\theta), \quad i = 1, 2, \dots \quad (5.17)$$

According to standard Bayesian calculus (i.e. Bayes' rule), $\mathcal{I}_0$'s prediction is an infinite mixture of observation model (5.17) and the optimal prior $pd A^\circ(\boldsymbol{\theta}|K)$ (5.16). Alternatively, it seems compelling to choose (as a predictor) a linear mixture of participants' $pd G_l(x)$ w.r.t optimal weights (5.13). These two predictive designs are now formally defined. After which we will specify when each should be used by $\mathcal{I}_0$.

- **FPD-A of $\mathcal{I}_0$'s predictor:** Replacing $\mathcal{I}_0$'s parametric observation model, i.e. $G_0(x|\boldsymbol{\theta})$ with $G_l(x)$ in M° (5.10), and its marginalization over $\boldsymbol{\theta}$ and l , it yields,

$$\begin{aligned} G_0^\circ(x|K) &= E_{A^\circ}[G_0(x|\boldsymbol{\theta})|K] \\ &= \sum_{l=1}^s w_l^\circ G_0^\circ(x|l, K) \end{aligned} \quad (5.18)$$

where,

$$G_0^\circ(x|l, K) = \int_{\Theta} G_0(x|\boldsymbol{\theta}) A^\circ(\boldsymbol{\theta}|l, K) d\boldsymbol{\theta} \quad (5.19)$$

- **linear pooling prediction with the FPD-optimal weights:** Returning to the actual form of M° (5.10) and its marginalization over $\boldsymbol{\theta}$ and l yields

$$G_0^{ao}(x|K) = \sum_{l=1}^s w_l^\circ G_l(x) \quad (5.20)$$

where, w_l° is given in (5.13).

□

Now we explain the role of each predictor as follows:

- Prediction in (5.20) is a linear pooling with tractable and consistent weights via FPD. Using this prediction, we can understand the role of linear pooling. When $\mathcal{I}_0$ wishes to predict the same x i.e. x_1 , prediction structure (5.20) is applied.
- Alternatively, $\mathcal{I}_0$ may wish to predict a future observation i.e. x_2 . In this case, the consistent prediction of x_2 is the structure (5.18), in FPD optimal sense.

The main difference between prediction (5.18) and (5.20) is the design of components. In prediction (5.18), the components (5.19) must be an infinite mixture of observation model (5.17) and the optimal prior $A^\circ(\boldsymbol{\theta}|K)$ (5.16). This structure allows us to consistently process the $\mathcal{I}_0$'s generative model of x , for the future data. On the other hand, the structure (5.20) is simply an optimal merging of participants' predictions. This structure does not propose a consistent way of processing $\mathcal{I}_0$'s future data. Notice that there is no uncertainty associated with the functional form of predictions (5.18) and (5.20). This is because $\mathcal{I}_0$'s parametric model, $G_0(x|\boldsymbol{\theta})$, conditioning on $\boldsymbol{\theta}$ is functionally known. Notably, the FPD-A of parameter prior (5.16) furnishes uncertainty in respect of $\mathcal{I}_0$'s parametric model, $x \sim G_0(x|\boldsymbol{\theta})$. Figure 5.1 shows the schematic diagram for processing K in Definition 5.1.1 into the FPD-A of parameter prior (5.16). Then, $\mathcal{I}_0$'s optimal predictor is deduced using Bayes' rule and applying marginalization.

Remark 5.2.4. (*Relationship to logarithmic pooling prediction*). Notably, $\mathcal{I}_0$ can propose a logarithmic pooling prediction by empirically designing w_l^g (4.4) with FPD-optimal weight w_l^o (5.13) as follows:

$$G_0^{go}(x|K) = \frac{1}{c_{G^{go}}} \prod_{l=1}^s [G_l(x)]^{w_l^o}, \quad (5.21)$$

where $c_{G^{go}} = \int \prod_{l=1}^s [G_l(x)]^{w_l^o} dx$.

□

5.3 FPD-sequential merging (FPD-S)

Recall that the knowledge, K , is defined in Definition 5.1.1. Using modelling assumptions (5.4) and (5.6), K constraints the possible set of models to $\mathbf{M}_K$ in (5.7). In fact, processing the joint model M (5.6) via (5.7) implies one-batch processing of multiple participants' predictors. In this section, we aim to employ the FPD-sequential processing (2.3.2) to process the knowledge, K . For this reason, the sequence of knowledge (2.15) is specialized to participant predictors, $G_l(x)$ s. In particular, $K_{1:j}$ involves G_1 to G_j . The schematic diagram of sequential processing is given in Figure 5.2. Processing knowledge in the sequential way leads to different solution from FPD-A of parameter prior (5.16). The

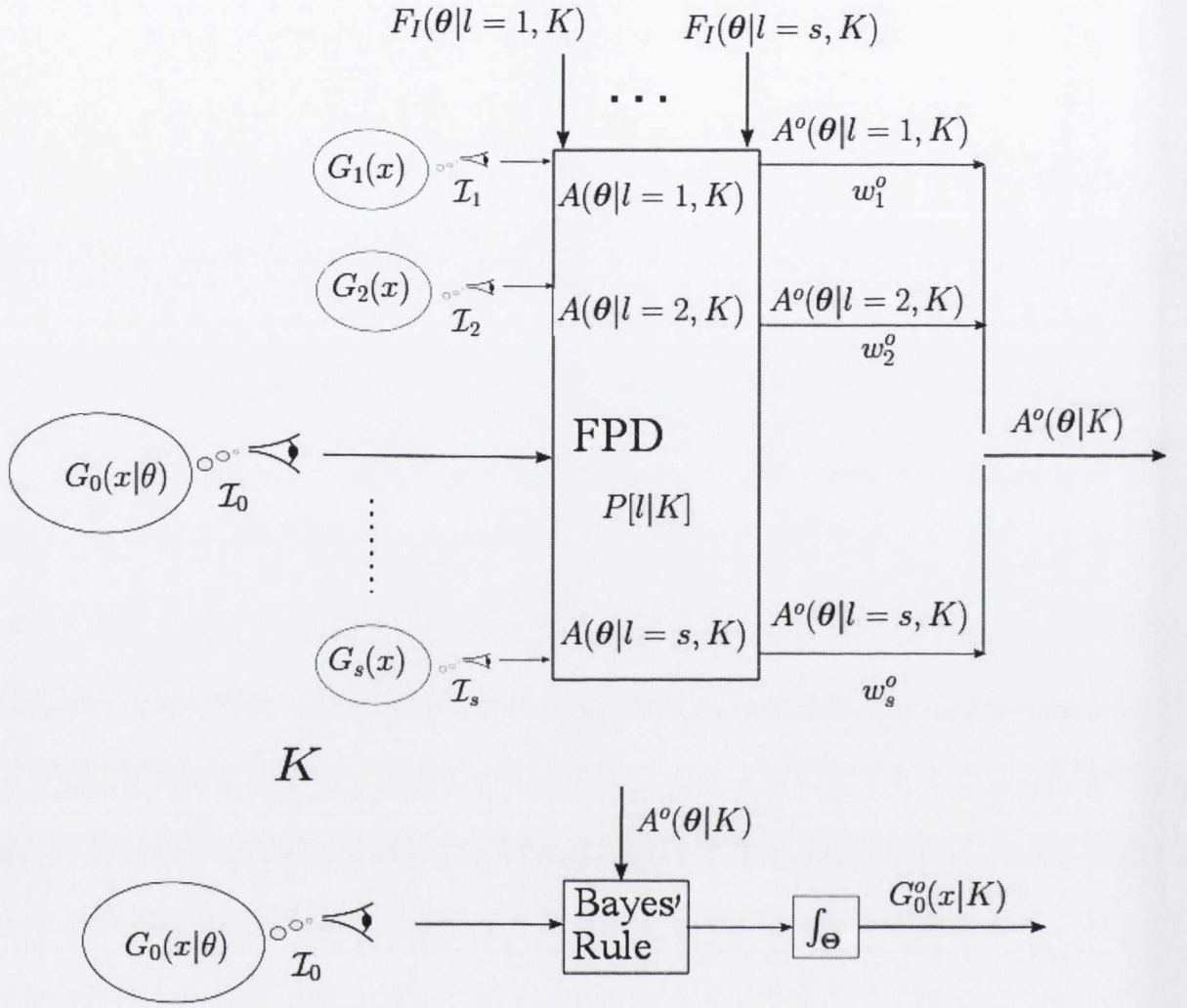


Figure 5.1. Schematic diagram for processing K (Definition 5.1.1) into the FPD-A (5.16), then deducing their optimal predictor. Here, $\int_{\Theta}$ denotes the marginalization operator.

guideline for sequential merging is formulated in the following lemma.

Lemma 5.3.1 (FPD-sequential merging of parameter prior). *Using (2.3.2), the sequential processing of $K_j, j \in \{1, \dots, s\}$, by $\mathcal{I}_0$, yields the FPD-sequential merging of $A(\theta|K)$ as follows:*

$$A^o(\theta|K) \propto F_I(\theta|K) \exp \left\{ - \sum_{l=1}^s D[G_l || G_0(\cdot|\theta)] \right\}, \quad (5.22)$$

where $F_I(\cdot)$ denotes the ideal pd conditioning on knowledge K , yet to be processed.

Proof. The ideal pd $F_I(\boldsymbol{\theta}|K)$ is specified as the pre-prior pd. Then we perform the following steps:

1. processing K_1 involving G_1 leads to,

$$A_1^o \equiv A^o(\boldsymbol{\theta}|K_1) \propto F_I(\boldsymbol{\theta}|K) \exp \{-D[G_1||G_0(\cdot|\boldsymbol{\theta})]\}, \quad (5.23)$$

$$c_{A_1^o} \equiv \int_{\Theta} F_I(\boldsymbol{\theta}|K) \exp \{-D[G_1||G_0(\cdot|\boldsymbol{\theta})]\} d\boldsymbol{\theta}. \quad (5.24)$$

2. in this step, the optimal design A_1^o (5.23) takes the role of pre-prior for processing $K_{1:2}$ which leads to,

$$A_2^o \equiv A^o(\boldsymbol{\theta}|K_{1:2}) \propto A^o(\boldsymbol{\theta}|K_1) \exp \{-D[G_2||G_0(\cdot|\boldsymbol{\theta})]\}, \quad (5.25)$$

$$\equiv F_I(\boldsymbol{\theta}|K) \exp \left\{ -\sum_{l=1}^2 D[G_l||G_0(\cdot|\boldsymbol{\theta})] \right\}, \quad (5.26)$$

$$c_{A_2^o} \equiv \int_{\Theta} F_I(\boldsymbol{\theta}|K) \exp \left\{ -\sum_{l=1}^2 D[G_l||G_0(\cdot|\boldsymbol{\theta})] \right\} d\boldsymbol{\theta}. \quad (5.27)$$

3. adopting the same strategy, the sequential processing of K in Definition 5.1.1 yields:

$$\begin{aligned} A_s^o \equiv A^o(\boldsymbol{\theta}|K_{1:s}) &\equiv A^o(\boldsymbol{\theta}|K) \\ &\propto A^o(\boldsymbol{\theta}|K_{1:s-1}) \exp \{-D[G_s||G_0(\cdot|\boldsymbol{\theta})]\}, \\ &\equiv F_I(\boldsymbol{\theta}|K) \exp \left\{ -\sum_{l=1}^s D[G_l||G_0(\cdot|\boldsymbol{\theta})] \right\}, \end{aligned} \quad (5.28)$$

$$c_{A_s^o} \equiv \int_{\Theta} F_I(\boldsymbol{\theta}|K) \exp \left\{ -\sum_{l=1}^s D[G_l||G_0(\cdot|\boldsymbol{\theta})] \right\} d\boldsymbol{\theta}. \quad (5.29)$$

□

Remark 5.3.1 (Properties of the FPD-sequential merging (5.22)). *The result (5.22) is invariant w.r.t. the order of processing participant' predictors and its functional form is totally different from the FPD-A of parameter prior (5.16). Notably, for the case where $s = 1$, they yield the same result which is also consistent with the result presented in [26].*

Using the FPD-S of parameter prior (5.22) and applying standard Bayesian calculus, the FPD-S of $\mathcal{I}_0$'s predictor is derived as follows.

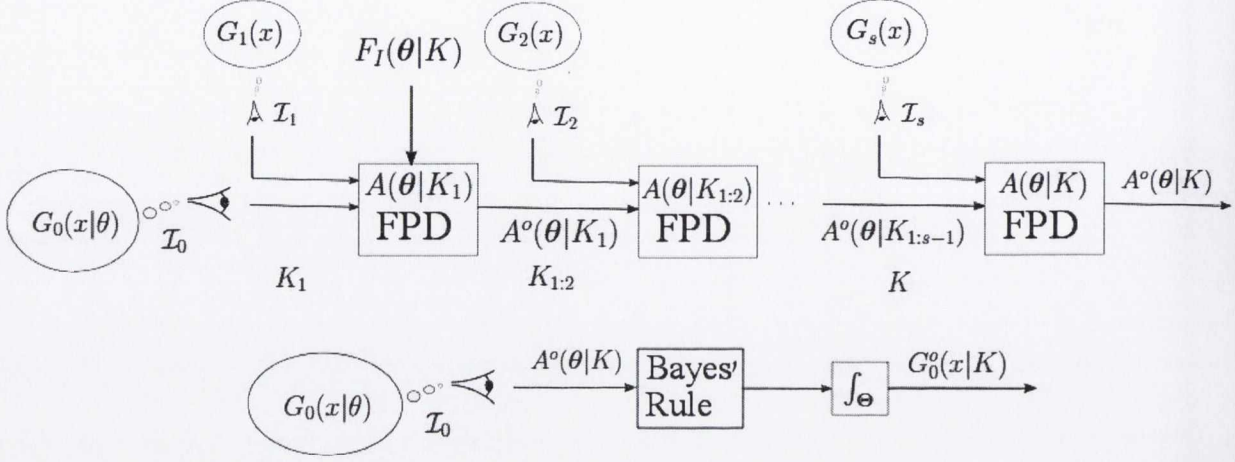


Figure 5.2. Schematic diagram for sequential processing of K (Definition 5.1.1) into the FPD-sequential merging (5.22), then deducing their sequential predictor. Here, $\int_{\Theta}$ denotes the marginalization operator.

Remark 5.3.2. (*FPD-S of $\mathcal{I}_0$'s predictor*) Using A^o (5.22), FPD-S of $\mathcal{I}_0$'s predictor, G_0^o , is computed as:

$$G_0^o(x|K) = E_{A^o(\theta|K)}[G_0(x|\theta, K)] \quad (5.30)$$

Recall that the FPD-sequential designs (5.22) and (5.30) do not provide non-uniform weights for the participants. Therefore, as an extension of sequential design we need to explore the other way of processing K in Definition 5.1.1 which allows us to optimally design non-uniform weights.

Figure 5.2 shows the schematic diagram for sequential processing of K in Definition 5.1.1 into the FPD-sequential merging (5.22). Then, FPD-S of $\mathcal{I}_0$'s predictor is deduced using Bayes' rule and applying marginalization.

5.4 FPD-optimal merging with restriction (FPD-R)

In this Section, we express the third way of processing the knowledge, K in Definition 5.1.1 where the knowledge-constrained set is the set of linear mixture of participant' predictors. Similar to the first way of knowledge processing in Section 5.2, this way is a one-batch

processing, however it yields different FPD solution which we refer it as FPD-optimal merging with restriction (FPD-R). This design can be considered as the extension of FPD-S (5.22). Since it proposes a principled and non-uniform design of weights. For this reason, we process the same knowledge, K , (Definition 5.1.1) and adopt a further assumption on the joint model (5.6) as follows,

$$A(\boldsymbol{\theta}|l, K) \equiv A(\boldsymbol{\theta}|K), \quad l = 1, \dots, s. \quad (5.31)$$

Applying (5.31) into (5.6) and marginalizing it over l yields:

$$\begin{aligned} M(x, \boldsymbol{\theta}|A, P, K) &\equiv \left(\sum_{l=1}^s P(l|K) G_l(x) \right) A(\boldsymbol{\theta}|K) \\ &\equiv \left(\sum_{l=1}^s w_l G_l(x) \right) A(\boldsymbol{\theta}|K). \end{aligned} \quad (5.32)$$

Also, we specify the ideal pd as,

$$M_I(x, \boldsymbol{\theta}|F_I, K) \equiv G_0(x|\boldsymbol{\theta}) F_I(\boldsymbol{\theta}|K), \quad (5.33)$$

Notice that we here do not specify different ideal pds for different participants— as we did in Section (5.3). Instead, one ideal pd is specified to process K . FPD-R of parameter prior is derived in the following corollary.

Corollary 5.4.1 (of Lemma 4.3.1, FPD-R of parameter prior). *Processing K (Definition 5.1.1) and underlying assumptions (5.31)-(5.33), the FPD-R of $A(\boldsymbol{\theta}|K)$ is:*

$$A^o(\boldsymbol{\theta}|K) \propto F_I(\boldsymbol{\theta}|K) \exp \left\{ - \sum_{l=1}^s \lambda_l^o D [G_l || G_0(\cdot|\boldsymbol{\theta})] \right\}, \quad (5.34)$$

where, $\boldsymbol{\lambda}^o = (\lambda_1^o, \dots, \lambda_s^o)$, denote the FPD-optimal design of weights,

$$\boldsymbol{\lambda}^o \equiv \underset{\boldsymbol{w}}{\operatorname{argmax}} c_{A^o}(\boldsymbol{w}), \quad (5.35)$$

where

$$\sum_{l=1}^s w_l = \sum_{l=1}^s \lambda_l^o = 1. \quad (5.36)$$

Proof. Substituting (5.32) and (5.33) into (5.7),

$$\begin{aligned}
D[M||M_I] &= \int_{\mathbb{X} \times \Theta} \left[\ln \frac{(\sum_{l=1}^s w_l G_l(x)) A(\boldsymbol{\theta}|K)}{G_0(x|\boldsymbol{\theta}) F_I(\boldsymbol{\theta}|K)} \right] \left(\sum_{l=1}^s w_l G_l(x) \right) A(\boldsymbol{\theta}|K) dx d\boldsymbol{\theta} \\
&= \int_{\Theta} \left[D[\sum_{l=1}^s w_l G_l || G_0(\cdot|\boldsymbol{\theta})] + \ln \left(\frac{A(\boldsymbol{\theta}|K)}{F_I(\boldsymbol{\theta}|K)} \right) \right] A(\boldsymbol{\theta}|K) d\boldsymbol{\theta} \\
&= D[A||A^*] - \ln c_{A^*}(\mathbf{w}),
\end{aligned} \tag{5.37}$$

where,

$$A^*(\boldsymbol{\theta}|K) \propto F_I(\boldsymbol{\theta}|K) \exp \left\{ -D[\sum_{l=1}^s w_l G_l || G_0(\cdot|\boldsymbol{\theta})] \right\}, \tag{5.38}$$

$$c_{A^*} = \int_{\Theta} F_I(\boldsymbol{\theta}|K) \exp \left\{ -D[\sum_{l=1}^s w_l G_l || G_0(\cdot|\boldsymbol{\theta})] \right\} d\boldsymbol{\theta}, \tag{5.39}$$

By the properties of KLD, the first term on the (RHS) of (5.37) achieves its minimal value of zero if $A(\boldsymbol{\theta}|K) = A^*(\boldsymbol{\theta}|K)$. Notably,

$$D[\sum_{l=1}^s w_l G_l || G_0(\cdot|\boldsymbol{\theta})] = \sum_{l=1}^s w_l E_{G_l} [\ln \sum_{l=1}^s w_l G_l(x)] - \sum_{l=1}^s w_l E_{G_l(x)} [\ln(G_0(x|\boldsymbol{\theta}))]. \tag{5.40}$$

Since, the first term in RHS (5.40) does not depend on $\boldsymbol{\theta}$, it can be ignored. Therefore, substituting (5.40) into A^* (5.38) yields,

$$\begin{aligned}
A^*(\boldsymbol{\theta}|\mathbf{w}, K) &\propto F_I(\boldsymbol{\theta}|K) \exp \left\{ -\sum_{l=1}^s w_l E_{G_l(x)} [G_0(x|\boldsymbol{\theta})] \right\}, \\
&\propto F_I(\boldsymbol{\theta}|K) \exp \left\{ -\sum_{l=1}^s w_l D[G_l || G_0(\cdot|\boldsymbol{\theta})] \right\},
\end{aligned} \tag{5.41}$$

$$c_{A^*} = \int_{\Theta} F_I(\boldsymbol{\theta}|K) \exp \left\{ -\sum_{l=1}^s w_l D[G_l || G_0(\cdot|\boldsymbol{\theta})] \right\} d\boldsymbol{\theta}, \tag{5.42}$$

Notably, A^* (5.41) involves the unknown weights, $\mathbf{w}$. Therefore, using step-wise maximization (5.35),

$$\boldsymbol{\lambda}^o \equiv \underset{\mathbf{w}}{\operatorname{argmax}} c_{A^*}(\mathbf{w}). \tag{5.43}$$

Substituting (5.43) back into (5.41) and (5.42), FPD-R of parameter prior (5.34) is achieved. $\square$

Notice that FPD-R of parameter prior (5.34) is derived underlying the assumption (5.31). This assumption restricts the knowledge-constrained set, $\mathbf{M}_K$, to the set of linear mixture of participants' predictors. For this reason, we refer to (5.34) as FPD-optimal merging with restriction.

Notice that FPD-R of parameter prior (5.34) can be evaluated more easily by using the underlying assumption (5.36). Recall that the evaluation of optimal weights in (4.15) is generally difficult. This issue can be improved using (5.43). In the following, we propose a class for which the FPD solutions is functionally tractable.

Remark 5.4.1 (Functionally tractable FPD solutions). *Using Definition 2.2.2, the FPD solutions (5.16), (5.22) and (5.34) are functionally tractable for the class of parametric model of x , $G_0(x|\boldsymbol{\theta})$, where $D[G_l||G_0(.|\boldsymbol{\theta})]$ is available in the closed form, e.g. $G_0(x|\boldsymbol{\theta})$ be a member of dynamic exponential family (DEF) [61]. The full explanation of employing FPD designs in AR model as an important member of DEF is presented in Chapter 6 .*

The FPD-R (5.34) is attained under assumption (5.32) stating that the joint space search, $\mathbf{M}_K$, involves the weighted linear mixture of participant' predictors. In particular, the variational space involves two variational quantises i.e. $A(\boldsymbol{\theta}|K)$ and $P(l|K) \in \Delta_{s-1}$. In the following Remark, FPD-R of $\mathcal{I}_0$'s predictor is presented.

Remark 5.4.2. (FPD-R of $\mathcal{I}_0$'s predictor) *Using $A^\circ(.)$ (5.34) and the optimal weights (5.35), FPD-R of $\mathcal{I}_0$'s predictor is deduced as:*

$$G_0^\circ(x|K) = E_{A^\circ(\boldsymbol{\theta}|K)}[G_0(x|\boldsymbol{\theta}, K)] \quad (5.44)$$

Figure 5.3 shows the schematic diagram for processing K in Definition 5.1.1 into FPD-R of parameter prior (5.34). Then, FPD-R of $\mathcal{I}_0$'s predictor is deduced using Bayes' rule and applying marginalization.

Notably FPD-R of predictor (5.44) and FPD-moral deliberator (4.21) have the same structure. But they are derived underlying different knowledges.

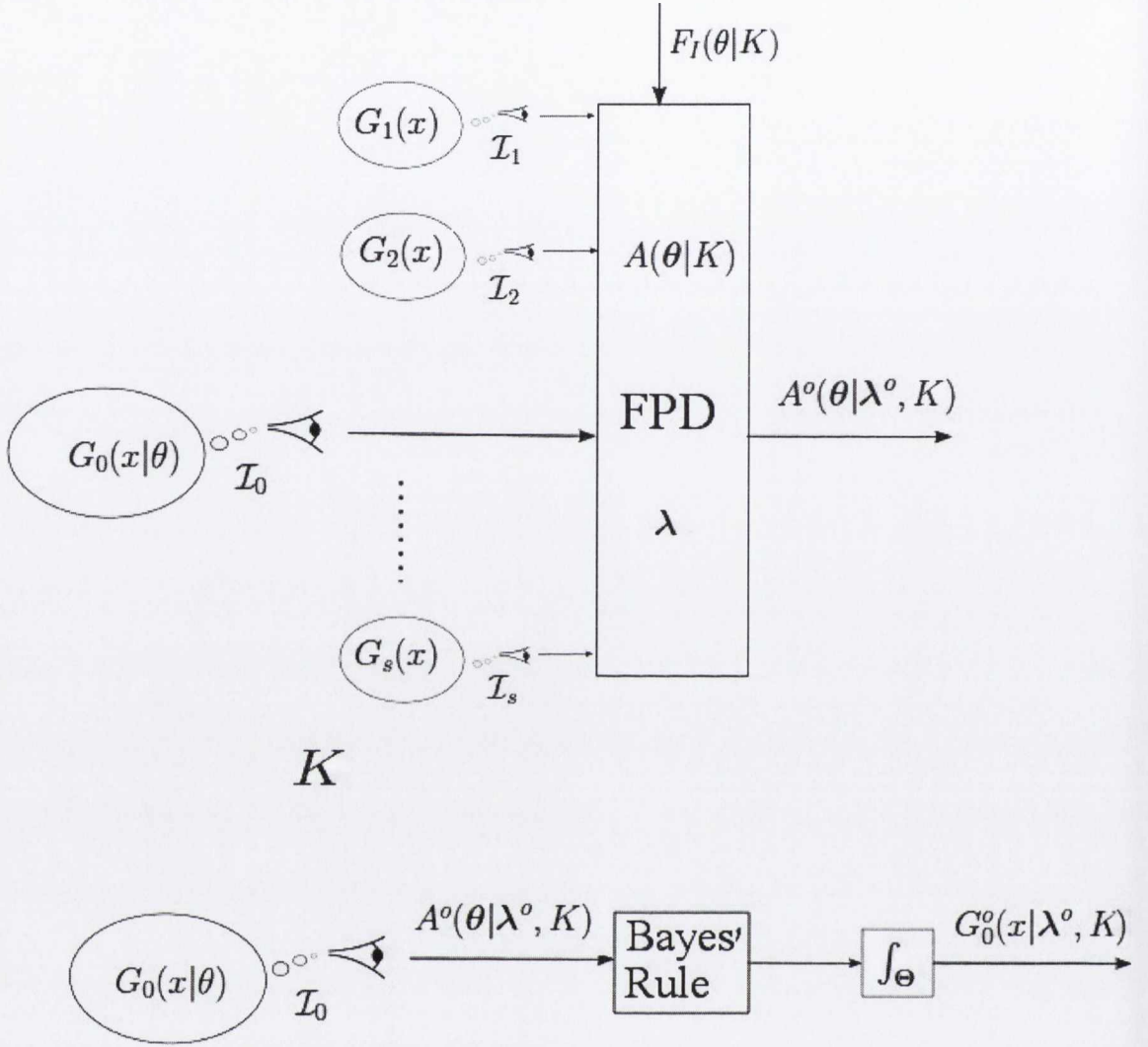


Figure 5.3. Schematic diagram for processing K (Definition 5.1.1) into the FPD-R (5.34), then deducing their optimal predictor. Here, $\int_{\Theta}$ denotes the marginalization operator.

5.5 Relationships among FPD-A, FPD-S and FPD-R

Recall that the way which K in (Definition 5.1.1) is processed leads to different knowledge-constrained set, ideal specifications and FPD solutions. In the following, we aim to compare and contrast the main perspectives of these three ways of knowledge processing and highlight the similarities and differences among FPD solutions (5.16), (5.22) and (5.34).

- In all FPD solutions, the same knowledge, K (Definition 5.1.1), is processed, but

underlying different assumptions and in different ways. In particular, FPD-A (5.16) and FPD-R (5.34) is derived under one-batch processing (i.e. one KLD for once). The difference between FPD-A (5.16) and FPD-R (5.34) is that the latter is attained under the extra assumption (5.31). On the other hand, the FPD-S (5.22) is achieved under sequential processing of s predictors (i.e. sequential processing of s KLDs).

- Processing knowledge, K , in different ways constraints the variational space to different number of variational quantities. For instance, the number of variational quantities of the knowledge-constrained set associated with FPD-A (5.16), FPD-R (5.34) and FPD-S (5.22) are respectively $s + 1$, 2 and 1. Notably FPD-A (5.16) is derived in the most flexible set and FPD-S (5.22) is achieved in the least flexible set.
- Different ideal pds are specified for different ways of processing K . For instance, FPD-A (5.16) requires us specify s ideal pds (5.3), i.e. $F_I(\boldsymbol{\theta}|l, K), l = 1, \dots, s$. While, for FPD-S (5.22), we specify one ideal pd in the first step. Then we use the optimal design of the previous step as the ideal pd of the next step in the Lemma 5.5. Furthermore, in FPD-R (5.34), only one ideal pd is specified i.e. $F_I(\boldsymbol{\theta}|K)$ (5.33).
- In FPD-S (5.22), any order of processing $G_l(x)$, yields the same result which asserts that this design is invariant w.r.t. the order of processing of predictors.
- FPD-A (5.16) and FPD-R (5.34) propose the principled and non-uniform design of weights, while FPD-S (5.22) assigns the uniform weights to all participants.
- None of these approaches proposes a non-uniform weight for the case where $s = 1$.

5.6 An illustrative example: normal case

Let participants contribute normal pd, each modelling $x \in \mathbb{R}$:

$$G_l(x|\boldsymbol{\eta}_l) \equiv N_x(\mu_l, r_l), \quad l = 1, \dots, s, \quad (5.45)$$

where $\boldsymbol{\eta}_l \equiv (\mu_l, r_l)$. $\mathcal{I}_0$ contribute parametric normal model:

$$G_0(x|\boldsymbol{\theta}) \equiv G_0(x|\mu, r) \equiv N_x(\mu, r) \quad (5.46)$$

where the unknown parameters are $\theta \equiv (\mu, r)$ with unknown mean μ and unknown variance r . In the following, we specialize the FPD solutions presented in Sections 5.2-5.4, into the normal case. Also, the corresponding predictor for this special case is deduced.

5.6.1 FPD-A: normal case

Adopting $F_l(\mu, r|l, K) \equiv NiG(\mathbf{V}_{l,I}, \nu_{l,I})$, $l = 1, \dots, s$, and substituting (5.45) and (5.46) into (5.11), the FPD-A of parameter pd for l^{th} -component becomes,

$$A^o(\mu, r|l, K) \propto NiG(\mathbf{V}_l^o, \nu_l^o) \quad (5.47)$$

where,

$$\mathbf{V}_l^o = \mathbf{V}_{l,I} + \begin{bmatrix} r_l + \mu_l^2 & \vdots & \mu_l \\ \dots\dots\dots & & \\ \mu_l & \vdots & 1 \end{bmatrix}, \quad (5.48)$$

$$\nu_l^o = \nu_{l,I} + 1. \quad (5.49)$$

Therefore, FPD-A of parameter prior (5.16) in the Remark 5.2.1, becomes:

$$A^o(\theta|K) = \sum_{l=1}^s w_l^o NiG(\mathbf{V}_l^o, \nu_l^o), \quad (5.50)$$

where,

$$w_l^o = \frac{w_{l,I} c_{A_l^o}}{\sum_{j=1}^s w_{j,I} c_{A_j^o}}, l = 1, \dots, s, \quad (5.51)$$

and,

$$c_{A_l^o} = \Gamma\left(\frac{\nu_l^o}{2}\right)(\Lambda_l^o)^{-\frac{\nu_l^o}{2}} |\mathbf{V}_l^o|^{-\frac{1}{2}} (2\pi)^{\frac{1}{2}}, l = 1, \dots, s. \quad (5.52)$$

Remark 5.6.1. ($\mathcal{I}_0$'s FPD-A of predictor (specialization of Remark (5.2.3))

- Notice that using (5.18) and (5.50), the FPD-A of predictor (5.18) becomes a mixture

where

$$c_A(\boldsymbol{\lambda}) = -\log \Gamma\left(\frac{\nu_{\boldsymbol{\lambda}}}{2}\right) + \frac{\nu_{\boldsymbol{\lambda}}}{2} \log \Lambda_{\boldsymbol{\lambda}} + \frac{1}{2} \log |V_{22,\boldsymbol{\lambda}}| - \frac{1}{2} \log(2\pi) \quad (5.59)$$

$$\begin{aligned} \Lambda_{\boldsymbol{\lambda}} &= V_{11,\boldsymbol{\lambda}} - \frac{V_{21,\boldsymbol{\lambda}}^2}{V_{22,\boldsymbol{\lambda}}} \\ &= \frac{(\sum_{l=1}^s \lambda_l) (\sum_{l=1}^s \lambda_l (r_l + \mu_l^2)) - (\sum_{l=1}^s \lambda_l \mu_l)^2}{\sum_{l=1}^s \lambda_l} \end{aligned} \quad (5.60)$$

Substituting $\sum_{l=1}^s \lambda_l = 1$ into (5.59) and (5.60), and using the “optimtool in MatLab”, the optimal value of $\boldsymbol{\lambda}$, i.e. $\boldsymbol{\lambda}^o$, is derived.

Remark 5.6.3. ($\mathcal{I}_0$ ’s FPD-R of predictor (specialization of Remark (5.4.2))

Adopting (5.55) into (5.4.2), the FPD-R of predictor for the normal case is the Student-t’s pd,

$$G_0^o(x|K) = St_x(\mu_{\boldsymbol{\lambda}^o}^o, \tau_{\boldsymbol{\lambda}^o}^o, \nu_{\boldsymbol{\lambda}^o}^o + 1), \quad (5.61)$$

where $\boldsymbol{\lambda}^o$, (5.51) and the Student-t’s shaping parameters, $\mu_{\boldsymbol{\lambda}^o}^o, \tau_{\boldsymbol{\lambda}^o}^o$, are derived from $\mathbf{V}_{\boldsymbol{\lambda}^o}^o$ (5.56) and $\nu_{\boldsymbol{\lambda}^o}^o$ (5.57).

Notice that the FPD-moral deliberator predictor of moral deliberator (4.21) for the normal case is quite similar to FPD-R (5.61), i.e.

$$G_0^o(x|K) = St_x(\mu_{\boldsymbol{\lambda}^o}^o, \tau_{\boldsymbol{\lambda}^o}^o, \nu_{\boldsymbol{\lambda}^o}^o + 1), \quad (5.62)$$

where $\boldsymbol{\lambda}^o$ does not have to be restricted to,

$$\sum_{l=1}^s \lambda_l^o = 1.$$

Remark 5.6.4. ($\mathcal{I}_0$ ’s FPD-S for the normal case) Substituting $\lambda_l^o = 1$, $l = 1, \dots, s$ into (5.55), the FPD-S of parameter prior (5.22) for the normal case is:

$$A^o(\mu, r|K) = NiG(\mathbf{V}^o, \nu^o). \quad (5.63)$$

Notice that adopting (5.63) to (5.30), the FPD-S of predictor is a Student- t 's pd, i.e.

$$G_0^o(x|K) = St_x(\mu^o, \tau^o, \nu^o + 1), \quad (5.64)$$

where the Student- t 's shaping parameters, μ^o, τ^o are derived from $(\mathbf{V}^o, \nu^o)$ in (5.63).

5.7 Comparative performance evaluation under simulation

Here we examine the simulation performance of different $\mathcal{I}$'s FPD solutions of predictor in the normal case presented in Section 5.6. As mentioned in Remark 5.6.2, the “optimtool” in MatLab is used for numerical evaluation of $\boldsymbol{\lambda}^o$. This simulation is performed for the normal case presented in the illustrative example Section 5.6. For this reason, we evaluate these $\mathcal{I}_0$'s predictors for $s = 2$ participants and $\mu_1 = 1$ and $\mu_2 = 5$, $r_1 = 1$ and $r_2 = 4$. Results here are derived in $MC = 1000$ runs.

The diffuse ideal pds are adopted and simulations are examined for operating factors i.e. the $\mathcal{I}$'s μ , r , n and ν_I in different figures. Here, the measure of performance is the mean squared error (MSE) defined as

$$MSE = \frac{1}{MC} \sum_{i=1}^{MC} (x^{(i)} - \hat{x}^{(i)})^2, \quad (5.65)$$

where, $x^{(i)}$ is a simulated value and $\hat{x}^{(i)}$ is the predicted value in i^{th} -MC run. Notice that the MSE (5.65) is computed for different approaches by replacing $\hat{x}^{(i)}$ with predicted values of corresponding approach in each run.

In Figure 5.4 the simulation is designed for $r = 1$, $n = 500$, $s = 2$. The participants report shaping parameters to $\mathcal{I}_0$ who process them and their own data, $\mathbf{x}_n$. Recall that here $\mathcal{I}_0$'s shaping parameters (μ, r) are unknown. The plot is sketched in terms of different $\mathcal{I}_0$'s simulated μ . Figure 5.4 shows that $\mathcal{I}$'s FPD-A of predictor (5.53) outperforms others. The difference of MSE is more distinctive where μ increases. An increasing trend is seen when μ grows.

In Figure 5.5 the simulation is designed for $\mu = 4$, $n = 500$, $s = 2$ in terms of

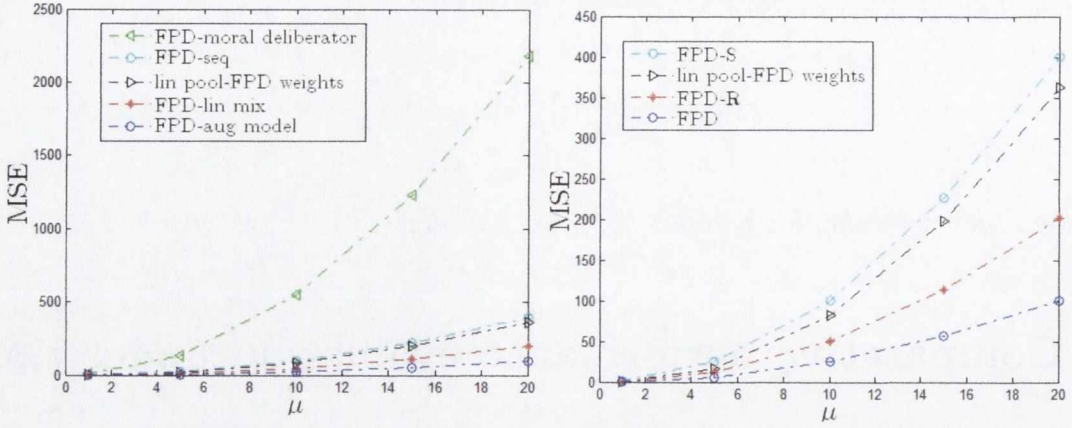


Figure 5.4. Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (5.53), lin pool-FPD weights (5.54), FPD-R (5.61), FPD-moral deliberator (5.62), FPD-S (5.64) against $\mathcal{I}_0$'s μ for $r = 1$, $n = 500$, $s = 2$.

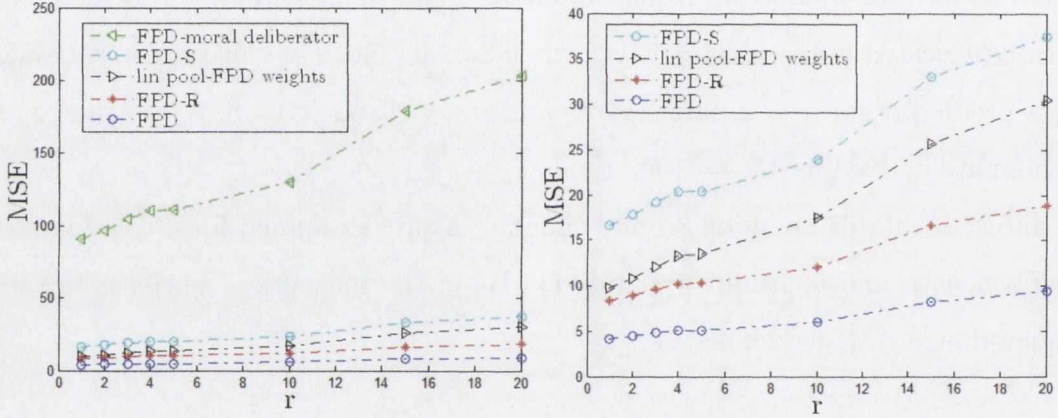


Figure 5.5. Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (5.53), lin pool-FPD weights (5.54), FPD-R (5.61), FPD-moral deliberator (5.62), FPD-S (5.64) against $\mathcal{I}_0$'s r for $\mu = 4$, $n = 500$, $s = 2$.

$r = \{1, 2, 3, 4, 5, 10, 15, 20\}$. This figure is designed for different simulated r . It shows that $\mathcal{I}$'s FPD-A of predictor (5.53) outperforms others with an increasing trend. The MSE of all predictors increases when r grows.

In Figure 5.6 the simulation is designed for $\mu = 4$, $r = 1$, $s = 2$ in terms of the number of observation $n = \{50, 100, 200, 300, 400, 500\}$. It shows that $\mathcal{I}$'s FPD-A of predictor (5.53) outperforms others with a flat trend.

In Figure 5.7 the simulation is designed for $\mu = 4$, $r = 1$, $n = 500$ and $s = 2$ in terms of $\nu = \{10, 20, 30, 40, 50\}$. It shows that $\mathcal{I}$'s FPD-A of predictor (5.53) outperforms others

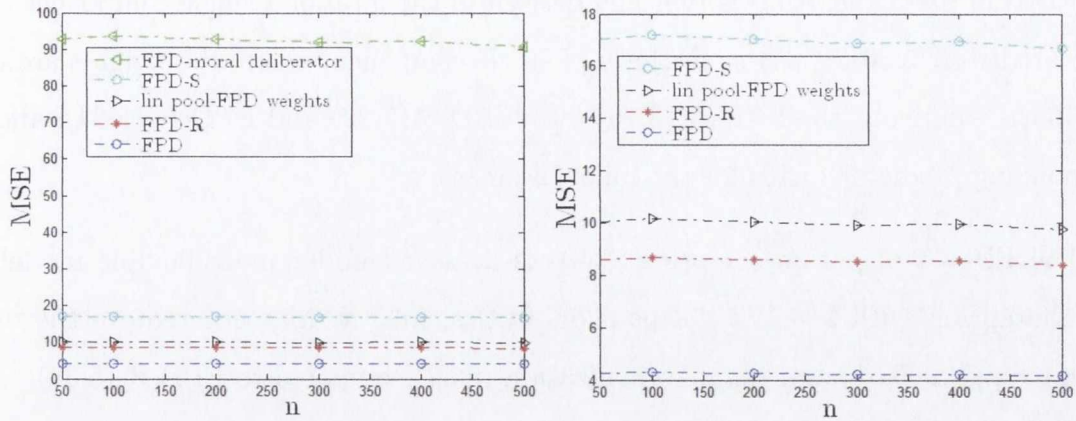


Figure 5.6. Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (5.53), lin pool-FPD weights (5.54), FPD-R (5.61), FPD-moral deliberator (5.62), FPD-S (5.64) against $\mathcal{I}_0$'s n for $\mu = 4$, $r = 1$, $s = 2$.

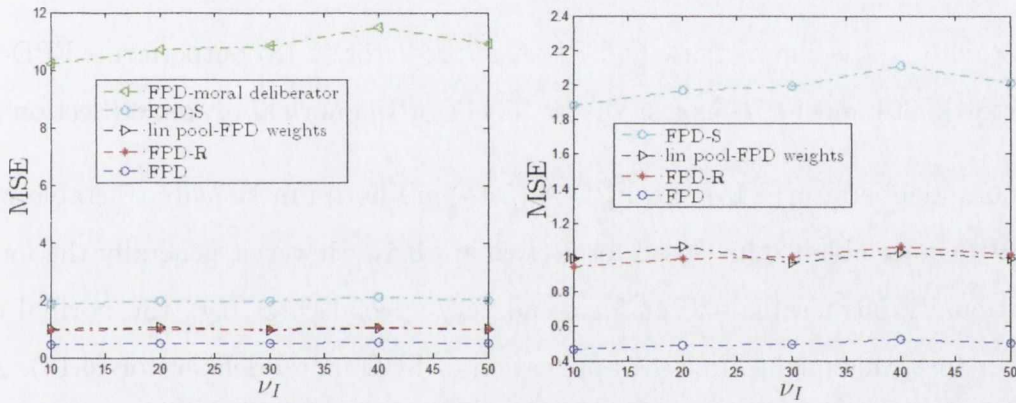


Figure 5.7. Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (5.53), lin pool-FPD weights (5.54), FPD-R (5.61), FPD-moral deliberator (5.62), FPD-S (5.64) against $\mathcal{I}_0$'s ν_I for $\mu = 4$, $r = 1$, $n = 500$.

with a flat trend.

5.8 Discussion

In this Chapter, we have proposed two FPD-optimal solutions (i.e. FPD-A and FPD-R) and one FPD sequential solution (i.e. FPD-S) of parameter prior, through merging the distributed predictive pds supplied by multiple participants. FPD-A (5.16) has provided the optimal and principled weights which are used to obtain the optimal predictor via Bayes' rule. Using FPD-A (5.16), the optimized predictor is in the form of linear mixture (5.18) for which both components and weights are tractably evaluated for many model-based contexts.

The FPD-R of predictor (5.44) is typically derived in the form of a single component. The FPD-S predictor (5.30) is also in the form of single component where the weights are unit, e.g. (5.64). We favour the FPD-A (5.16) over FPD-S (5.22) and FPD-R (5.34), and its corresponding predictor (5.18) for the following reasons:

- The FPD-A of parameter prior (5.16) is achieved under more flexible modelling assumption with $s + 1$ variational quantities, i.e. it does not require the extra assumption (5.31) and the marginalization (5.32) compared to FPD-R (5.34).
- The FPD-A of parameter prior (5.16) can be considered as the hierarchical extension of linear pooling approach yielding a linear weighted mixture of predictors with a tractable and optimal design of weights.
- In Section 5.7, it is shown that the FPD-A of predictor (5.18) outperforms FPD-S of predictor (5.30) and FPD-R of predictor (5.44) for the normal example (Section 5.6).

The FPD-moral deliberator (4.15) and FPD-A (5.16) are both functionally tractable when $D[G_l|G_0(.|\boldsymbol{\theta})]$ is available in the closed form (Remark 5.5). However, generally the former is computationally intractable except for some very special case (e.g. the normal case 4.3.2.1). Therefore, this might limit the application of FPD-moral deliberator (4.15). Also we have evaluated the normal case presented in Section 5.6 for at most $s = 2$ using the “optimtool in MatLab”. Recall that the computational issue of FPD-moral deliberator (4.15) in normal case (5.6) is related to solving s equations in (5.59), where the optimal Kuhn-Tucker multipliers, $\boldsymbol{\lambda}^o$ cannot be computed in an explicit function of K . Also, the simulation result (Section 5.7) shows that the performance of FPD-moral deliberator (4.15) is quite weak compared to FPD-A (5.16). Therefore, we will not proceed with FPD-moral deliberator (4.15) in the next Chapter. Instead, we will implement the FPD solutions presented in this Chapter for AR model.

FPD-A (5.16) is achieved under this assumption that just one of the participant is true. However, for more general contexts where two or more participants are true, this approach is not applied. The future direction of this extension will be explained in Chapter 8. Also, FPD-A (5.16) and FPD-moral deliberator (4.15) are designed for the knowledge-constrained set with $s + 1$ variational quantities. For instance, $\boldsymbol{M}_K$ corresponding to (5.16) involves $A(\boldsymbol{\theta}|l, K)$, $l = 1, \dots, s$, and pmf $P(l|K) \in \Delta_{s-1}$. Also, $\boldsymbol{M}_K$ corresponding to

FPD-moral deliberator (4.15) involves $A(\boldsymbol{\theta}|K)$, and λ_l , $l = 1, \dots, s$. However, they are derived under different knowledge-constrained set with different functional form. Also, FPD-moral deliberator (4.15) and FPD-R (5.34) have the same functional form, while they are derived within different knowledge-constrained set with different variational quantities. In this Chapter, the deliberator fully respects to the true participant and do not hold any proportion for their own parametric model i.e. $G_0(x|\boldsymbol{\theta})$. A possible direction will be provided in Chapter 8.

The FPD-optimal parametric hierarchy of the deliberator (Lemma 5.2.1 and Corollary 5.4.1) remains sensitive to the model choice, $G_0(x|\boldsymbol{\theta})$, and forms an important part of $\mathcal{I}_0$'s knowledge, K (see item 1 of Definition 5.1.1). The relaxation of parametric form to a continuous nonparametric form, $\boldsymbol{\theta} \equiv G$ (i.e. $x \sim G_0(x)$, where $G_0(x)$ is now a functionally unknown pd) is an important task which will be fully discussed in Chapter of the future work.

While our centralized deliberator-based scheme delivers important advantages in Remark 4.1.1, the deliberation (i.e. decision-making) itself is confined to the deliberator (centralized deliberation). In practice i.e. smart grid systems [109] and networks with local controllers [111], the centralized scheme might face with scalability problems [29] for large s , and vulnerability of the overall system to deliberator node failure [109]. In order to overcome these problems, the participants—particularly in large-scale systems—can be grouped, each group being moderated by a local deliberator. The current framework can then be adopted by each cluster, with each group's deliberator reporting their local FPD-A of predictor (5.18) to the global deliberator, for final decision.

Chapter 6 —

Processing multiple autoregressive nodes

In this Chapter, we aim to design the FPD-optimal parameter prior and deduce optimal predictor through processing both stochastic knowledge and data distributed among multiple nodes. For this reason, we exploit the technology presented in Chapter 3 for processing data and adopt the centralized deliberator-based scheme presented in Chapter 5 to process the probabilistic knowledge. In particular, we assume that each node (i.e. both deliberator and participants) models x with AR model. The participants adopt the conjugate prior and process their own local data into the OSA predictor (3.33). Then all participant' predictors are reported to the AR deliberator who processes these predictors with own parametric model (i.e. AR observation model) to design their optimal parameter prior using the methodology presented in Chapter 5. Notably, we will study the AR deliberator for both known and unknown model order, p . Then we will employ FPD-A of parameter prior (5.16) and corresponding optimal predictor (5.18) to derive the tractable and optimal design of parameter prior and predictor with principled and tractable design of weights. The FPD-optimal weights are exploited for linear pooling prediction. Furthermore, the FPD-R of parameter prior (5.34) and its corresponding predictor are derived for multiple AR nodes. Notably, FPD-A of predictor (5.18) is considered as our primary solution for the reasons discussed in the Section (5.8). We perform extensive simulations in order to compare FPD solutions with linear pooling predictor (with FPD-optimal design of weights) (5.20) for a range of operating factors.

6.1 FPD solutions for AR deliberator: known model order

Here, we assume that the deliberator and each participant are commonly modelling $x \in \mathbb{R}$ using AR model with known model orders, p and $p_l, l = 1, \dots, s$ respectively. The knowledge, K , for a system with multiple AR nodes is defined as follows.

Definition 6.1.1. ($\mathcal{I}_0$ ' Knowledge, K , for multiple AR nodes with known model order).

1. $\mathcal{I}_0$ adopt the parametric AR model with known model order (p) (3.5), i.e.

$$G_o(x_n | \mathbf{x}_{n-1}, \mathbf{a}^{(p)}, r, p) \equiv G_o(x_n | \psi_n^{(p)T}, \mathbf{a}^{(p)}, r) \equiv N_x(\psi_n^{(p)T} \mathbf{a}^{(p)}, r), \quad (6.1)$$

2. using (3.33), the contributed predictor by l^{th} - participant is Student's-t pd,

$$G_l(x_n | \mathbf{x}_{n-1,l}, p_l) = St_x(\mu_{l,n-1}^{(p_l)}, \tau_{l,n-1}^{(p_l)}, \nu_{l,n-1}^{(p_l)} + 1), \quad l = 1, \dots, s, \quad (6.2)$$

where $\mu_{l,n-1}^{(p_l)}, \tau_{l,n-1}^{(p_l)}, \nu_{l,n-1}^{(p_l)}$, are known shaping parameters of Student's-t pd. Recall that G_l converts data $\mathbf{x}_{n-1,l}$ into shaping parameters of Student's-t pd. For simplicity, we denote $Var(x_n | \mathbf{x}_{n-1}, p_l) \equiv \tau_{l,n-1}^{(p_l)}$.

3. the data sets, $\mathbf{x}_{n-1,l}, l = 1, \dots, s$, are assumed to be the part of knowledge. Notably, the deliberator and participants process their own data set possibly with different sample size. Here, to avoid notational confusion, we assume that the sample size of all nodes are the same.
4. $\mathcal{I}_0$'s model order, p , which is known.
5. participants' model order, $p_l, l = 1, \dots, s$, which are known.

Adopting the ideal pd,

$$F_I(\mathbf{a}^{(p)}, r | l, p, K) \equiv NiG(V_{l,I}^{(p)}, \nu_{l,I}^{(p)}), \quad (6.3)$$

we will employ FPD-A (5.16) and derive the optimal design of $A(\mathbf{a}^{(p)}, r|l, p, K)$ underlying K (Definition 6.1.1) in the following section.

6.1.1 FPD-optimal solution of deliberator's AR parameter prior: Application FPD-A (5.16) and FPD-R (5.34)

Processing K in Definition 6.1.1 and using ideal specification (6.3), the FPD-A of $A(\mathbf{a}^{(p)}, r|l, p, K)$ is:

$$\begin{aligned} A^o(\mathbf{a}^{(p)}, r|p, l, K) &\propto F_I(\mathbf{a}^{(p)}, r|p, l, K) \exp \left\{ -D[G_I || G_0(.|\mathbf{a}^{(p)}, r, p)] \right\} \\ &= NiG_{a,r}(V_{l,I}^{(p)}, \nu_{l,I}^{(p)}) r^{\frac{-1}{2}} \exp \left\{ \frac{-1}{2r} [\tau_{l,n-1}^{(p)} + (\mu_{l,n-1}^{(p)})^2 - 2\mu_{l,n-1}^{(p)} \psi_n^{(p)T} \mathbf{a}^{(p)} + (\psi_n^{(p)T} \mathbf{a}^{(p)})^2] \right\} \end{aligned} \quad (6.4)$$

Therefore,

$$A^o(\mathbf{a}^{(p)}, r|p, l, K) = NiG_{a,r}(\mathbf{V}_{l,n-1}^{(p)o}, \nu_{l,n-1}^{(p)o}) \quad (6.5)$$

where, $\mathbf{V}_{l,n-1}^{(p)o}$ enjoys the same partitioning as (3.20) with,

$$\mathbf{V}_{l,(n-1)11}^{(p)o} = \mathbf{V}_{l,I11}^{(p)} + \tau_{l,n-1}^{(p)} + (\mu_{l,n-1}^{(p)})^2 \left(1 + \{ \psi_n^{(p)} (\psi_n^{(p)} \psi_n^{(p)T})^{-1} \psi_n^{(p)} \} \right) \quad (6.6)$$

$$\mathbf{V}_{l,(n-1)a1}^{(p)o} = \mathbf{V}_{l,Ia1}^{(p)} + \mu_{l,n-1}^{(p)} \psi_n^{(p)}, \quad (6.7)$$

$$\mathbf{V}_{l,(n-1)aa}^{(p)o} = \mathbf{V}_{l,Iaa}^{(p)} + \psi_n^{(p)} \psi_n^{(p)T}, \quad (6.8)$$

and

$$\nu_{l,n-1}^{(p)o} = \nu_{l,I}^{(p)} + 1. \quad (6.9)$$

Let $P_I(l|p, K)$ be uniform, then using FPD-A (5.16) and (6.5), the FPD-A of $A(\mathbf{a}^{(p)}, r|p, K)$ is:

$$\begin{aligned} A^o(\mathbf{a}^{(p)}, r|p, K) &= \sum_{l=1}^s w_{l,n-1}^{(p)o} A^o(\mathbf{a}^{(p)}, r|p, l, K) \\ &= \sum_{l=1}^s w_{l,n-1}^{(p)o} NiG_{a,r}(\mathbf{V}_{l,n-1}^{(p)o}, \nu_{l,n-1}^{(p)o}), \end{aligned} \quad (6.10)$$

where $w_{l,n-1}^{(p)o}$ is tractably evaluated as:

$$w_{l,n-1}^{(p)o} = \frac{\kappa_{a,r}(\mathbf{V}_{l,n-1}^{(p)o}, \nu_{l,n-1}^{(p)o})}{\sum_{j=1}^s \kappa_{a,r}(\mathbf{V}_{j,n-1}^{(p)o}, \nu_{j,n-1}^{(p)o})} \quad (6.11)$$

where,

$$\kappa_{a,r}(\mathbf{V}_{l,n-1}^{(p)o}, \nu_{l,n-1}^{(p)o}) = \Gamma\left(\frac{\nu_{l,n-1}^{(p)o}}{2}\right) (\Lambda_{l,n-1}^{(p)o})^{-\frac{\nu_{l,n-1}^{(p)o}}{2}} |\mathbf{V}_{l,(n-1)aa}^{(p)o}|^{-\frac{1}{2}} (2\pi)^{\frac{1}{2}}, \quad (6.12)$$

$$\Lambda_{l,n-1}^{(p)o} = \mathbf{V}_{l,(n-1)11}^{(p)o} - (\mathbf{V}_{l,(n-1)1a}^{(p)o})^T (\mathbf{V}_{l,(n-1)aa}^{(p)o})^{-1} \mathbf{V}_{l,(n-1)1a}^{(p)o} \quad (6.13)$$

Remark 6.1.1. (FPD-R of deliberator's AR parameter prior: Application (5.34))

Recall that when we restrict the knowledge-constrained set to linear mixture of predictors, the FPD-R of $A(\mathbf{a}^{(p)}, r|p, K)$ through processing K (Definition 6.1.1) and underlying ideal specification (6.3) is:

$$A^o(\mathbf{a}^{(p)}, r|p, K) = NiG_{a,r}(\mathbf{V}_{\lambda^o,n-1}^{(p)o}, \nu_{\lambda^o,n-1}^{(p)o}) \quad (6.14)$$

where,

$$\mathbf{V}_{\lambda^o,n-1}^{(p)o} = \sum_{l=1}^s \lambda_{l,n-1}^{(p)o} \mathbf{V}_{l,(n-1)}^{(p)o}, \quad (6.15)$$

$$\nu_{\lambda^o,n-1}^{(p)o} = \sum_{l=1}^s \lambda_{l,n-1}^{(p)o} \nu_{l,(n-1)}^{(p)o}. \quad (6.16)$$

Here, $\nu_{l,(n-1)}^{(p)o}$ is given in (6.9) and the elements of $\mathbf{V}_{l,(n-1)}^{(p)o}$ are given in (6.6), (6.7) and (6.8). Notice that the similar guideline to Remark (5.6.2) can be exploited to evaluate the optimal weights, λ^o .

In the following Remark, we adopt FPD-A (6.10) and FPD-R (6.14) in order to deduce the optimal predictors underlying Definition 6.1.1.

Remark 6.1.2. ($\mathcal{I}_0$'s optimal predictors underlying Definition 6.1.1)

- **FPD-A of $\mathcal{I}_0$'s predictor (5.18) underlying Definition 6.1.1:** using the

FPD-A of $\mathcal{I}_0$'s predictor (5.18) and adopting the optimal prior (6.10) and the optimal weights (6.11), FPD-A of $\mathcal{I}_0$'s predictor of x_n , underlying Definition 6.1.1 is:

$$G_0^o(x_n | \mathbf{x}_{n-1}, p, K) = \sum_{l=1}^s w_{l,n-1}^{(p)o} \mathcal{S}t_x(\mu_{l,n-1}^{(p)o}, \tau_{l,n-1}^{(p)o}, \nu_{l,n-1}^{(p)o} + 1) \quad (6.17)$$

where, the counter, $\nu_{l,n-1}^{(p)o}$, is given in (6.9) and

$$\mu_{l,n-1}^{(p)o} = \psi_n^{(p)T} (\mathbf{V}_{l,(n-1)aa}^{(p)o})^{-1} \mathbf{V}_{l,(n-1)a1}^{(p)o}, \quad (6.18)$$

$$\tau_{l,n-1}^{(p)o} = \frac{\nu_{l,n-1}^{(p)o}}{\Lambda_{l,n-1}^{(p)o} \left(1 + \psi_n^{(p)T} (\mathbf{V}_{l,(n-1)aa}^{(p)o})^{-1} \psi_{l,n}^{(p)} \right)}. \quad (6.19)$$

Also, the FPD-A of $\mathcal{I}_0$'s predictor of x_{n+1} is

$$G_0^o(x_{n+1} | \mathbf{x}_n, p, K) = \sum_{l=1}^s w_{l,n}^{(p)o} \mathcal{S}t_x(\mu_{l,n}^{(p)o}, \tau_{l,n}^{(p)o}, \nu_{l,n}^{(p)} + 1) \quad (6.20)$$

- **linear pooling prediction with $\mathcal{I}_0$'s FPD-optimal weights underlying Definition 6.1.1:** using participants' predictors (6.2), the linear pooling predictor with the FPD-optimal weights (6.11) of x_n is

$$\begin{aligned} G_0^{ao}(x_n | \mathbf{x}_{n-1}, K) &= \sum_{l=1}^s w_{l,n-1}^{(p)o} G_l(x_n | \mathbf{x}_{n-1}, p_l) \\ &= \sum_{l=1}^s w_{l,n-1}^{(p)o} \mathcal{S}t_x(\mu_{l,n-1}^{(p_l)}, \tau_{l,n-1}^{(p_l)}, \nu_{l,n-1}^{(p_l)}), \end{aligned} \quad (6.21)$$

Notably, the linear pooling design (6.21) cannot be applied to deduce the optimal predictor of x_{n+1} . Instead, we can exploit the FPD-optimal design (6.20) to process $\mathcal{I}_0$'s future data.

- **FPD-R of $\mathcal{I}_0$'s predictor (5.44) underlying Definition 6.1.1:** using the FPD-R of predictor (5.44), and the optimal prior (6.14), FPD-R of $\mathcal{I}_0$'s predictor of x_n is:

$$G_0^o(x_n | \mathbf{x}_{n-1}, p, K) = \mathcal{S}t_x(\mu_{\lambda^o,n-1}^{(p)o}, \tau_{\lambda^o,n-1}^{(p)o}, \nu_{\lambda^o,n-1}^{(p)o} + 1), \quad (6.22)$$

where λ^o can be evaluated using Remark (5.6.2) with $\sum_{l=1}^s \lambda_l^o = 1$. $\square$

6.2 FPD solutions for deliberator's AR parameter prior: unknown participants' model order

In this Section, we aim to design the FPD-optimal prior and predictive pds underlying knowledge, K , (Definition 6.1.1) where constraint 5 is relaxed. For this reason, let the order of participants p_l , $l = 1, \dots, s$, be unknown. In this case each participant contributes the mixture of Student's-t pd, i.e.

$$\begin{aligned} G_l(x_n | \mathbf{x}_{n-1}) &= \sum_{p_l=1}^{p_{l,u}} \alpha_{l,n-1}^{(p_l)} G_l(x_n | \mathbf{x}_{n-1}, p_l) \\ &= \sum_{p_l=1}^{p_{l,u}} \alpha_{l,n-1}^{(p_l)} \mathcal{S}t_x(\mu_{l,n}^{(p_l)}, \tau_{l,n}^{(p_l)}, \nu_{l,n}^{(p_l)}), \quad l = 1, \dots, s \end{aligned} \quad (6.23)$$

Adopting $\mathcal{I}_0$'s parametric AR model (6.1) and using the ideal pd (6.3), the FPD-A of $A(\mathbf{a}^{(p)}, r | p, K)$ is:

$$A^o(\mathbf{a}^{(p)}, r | p, K) = \sum_{l=1}^s w_{l,n-1}^{(p)o} A^o(\mathbf{a}^{(p)}, r | p, l, K), \quad (6.24)$$

where, the l -conditional FPD-A of $A(\mathbf{a}^{(p)}, r | l, p, K)$ is:

$$A^o(\mathbf{a}^{(p)}, r | l, p, K) = NiG_{a,r}(\mathbf{V}_{l,n-1}^{(p)o}, \nu_{l,n-1}^{(p)o}). \quad (6.25)$$

Here, the elements of $\mathbf{V}_{l,(n-1)}^{(p)o}$ are:

$$\mathbf{V}_{l,(n-1)11}^{(p)o} = \mathbf{V}_{l,I11}^{(p)} + \tau_{l,n-1} + (\mu_{l,n-1})^2 \left(1 + \{ \psi_n^{(p)} (\psi_n^{(p)} \psi_n^{(p)T})^{-1} \psi_n^{(p)} \} \right) \quad (6.26)$$

$$\mathbf{V}_{l,(n-1)a1}^{(p)o} = \mathbf{V}_{l,Ia1}^{(p)} + \mu_{l,n-1} \psi_n^{(p)}, \quad (6.27)$$

$$\mathbf{V}_{l,(n-1)aa}^{(p)o} = \mathbf{V}_{l,Iaa}^{(p)} + \psi_n^{(p)} \psi_n^{(p)T}, \quad (6.28)$$

where,

$$\mu_{l,n-1} = \sum_{p_l=1}^{p_{l,u}} \alpha_{n-1}^{(p_l)} \mu_{l,n-1}^{(p_l)}, \quad (6.29)$$

$$\tau_{l,n-1} = \sum_{p_l=1}^{p_{l,u}} \alpha_{l,n-1}^{(p_l)} \left(\tau_{l,n-1}^{(p_l)} + (\mu_{l,n-1}^{(p_l)})^2 \right) - \left(\sum_{p_l=1}^{p_{l,u}} \alpha_{l,n-1}^{(p_l)} \mu_{l,n-1}^{(p_l)} \right)^2 \quad (6.30)$$

Notice that (6.29) and (6.30) are available in [141].

Remark 6.2.1. (*The application of FPD-R of prior (5.34) underlying Definition 6.1.1 for unknown p_l*)

Applying (6.26)-(6.28) into (6.15), the FPD-R of parameter prior is derived in the form of NiG (6.14). Since the computation is quite similar to (6.14), we avoid repetition.

Remark 6.2.2. (*$\mathcal{I}_0$'s optimal predictors underlying Definition 6.1.1 without constraint 5*) adopting $A^o(\cdot)$ (6.24) as the optimal prior in (6.17), (6.20) and (6.21), the FPD-A predictors underlying Definition 6.1.1 without constraint 5, are achieved. Also, using the FPD-R of parameter prior derived from Remark 6.2.1, the FPD-R of predictor is computed in the straight way.

6.3 FPD-optimal solutions of the AR deliberator: unknown model order

Here, the aim is to derive the FPD-optimal prior and predictor underlying Definition 6.1.1, relaxing constraint 4. For this reason, let p be unknown. Underlying assumptions 1-3, Section 3.4, and using p -conditional design (6.10), the FPD-A of an arbitrary subset of parameters, $A(\mathbf{a}_\delta, r | L \geq p_0, K)$ is:

$$\begin{aligned} A^o(\mathbf{a}_\delta, r | L \geq p_0, K) &= \sum_{p=p_0}^{p_u} \beta_{n-1}^{(p)} A^o(\mathbf{a}_\delta^{(p)}, r | p, K) \\ &= \sum_{p=p_0}^{p_u} \beta_{n-1}^{(p)} \sum_{l=1}^s w_{l,n-1}^{(p)o} \text{NiG}_{a,r}(\mathbf{V}_{l,n-1}^{(p)o}, \nu_{l,n-1}^{(p)o}), \end{aligned} \quad (6.31)$$

where the weights, $\beta_{n-1}^{(p)}$, is computed via (3.44). Furthermore, the FPD-R of parameter prior in the set of linear mixture is:

$$A^o(\mathbf{a}_\delta, r | L \geq p_0, K) = \sum_{p=p_0}^{p_u} \beta_{n-1}^{(p)} NiG_{a,r}(\mathbf{V}_{\lambda_{n-1}^o}^{(p)o}, \nu_{\lambda_{n-1}^o}^{(p)o}), \quad (6.32)$$

where, $\mathbf{V}_{\lambda_{n-1}^o}^{(p)o}$ and $\nu_{\lambda_{n-1}^o}^{(p)o}$ are respectively given in (6.15) and (6.16).

Remark 6.3.1. ($\mathcal{I}_0$'s predictors underlying Definition 6.1.1 relaxing constraint 4)

- **FPD-A of predictor for unknown p :** using the p -conditional FPD-A of predictor (6.17) and the Remark 3.4.3, the computation of FPD-A of $\mathcal{I}_0$'s predictor for unknown p is:

$$\begin{aligned} G_0^o(x_n | \mathbf{x}_{n-1}, K) &= \sum_{p=1}^{p_u} G_0^o(x_n, p | \mathbf{x}_{n-1}, K) \\ &= \sum_{p=1}^{p_u} \alpha_{n-1}^{(p)o} \sum_{l=1}^s w_{l,n-1}^{(p)o} \mathcal{S}t_x(\mu_{l,n-1}^{(p)o}, \tau_{l,n-1}^{(p)o}, \nu_{l,n-1}^{(p)o} + 1) \end{aligned} \quad (6.33)$$

- **linear pooling predictor with FPD-optimal weights for unknown p :** applying the same procedure, the linear pooling predictor of x_n with FPD-optimal weights for unknown p is computed as follows:

$$G_0^{ao}(x_n | \mathbf{x}_{n-1}, K) = \sum_{p=1}^{p_u} \alpha_{n-1}^{(p)o} \sum_{l=1}^s w_{l,n-1}^{(p_l)o} \mathcal{S}t_x(\mu_{l,n-1}^{(p_l)o}, \tau_{l,n-1}^{(p_l)o}, \nu_{l,n-1}^{(p_l)o}), \quad (6.34)$$

- **FPD-R of $\mathcal{I}_0$'s predictor for unknown p :** Adopting p -conditional FPD-R of parameter prior (6.14), the FPD-R of predictor for unknown p is obtained as follows:

$$G_0^o(x_n | \mathbf{x}_{n-1}, K) = \sum_{p=1}^{p_u} \alpha_{n-1}^{(p)o} \mathcal{S}t_x(\mu_{\lambda_{n-1}^o}^{(p)o}, \tau_{\lambda_{n-1}^o}^{(p)o}, \nu_{\lambda_{n-1}^o}^{(p)o} + 1) \quad (6.35)$$

where p -conditional optimal predictor is given in (6.22).

6.4 Comparative performance evaluation under simulation

Here we examine the simulation performance for different optimal predictors i.e. FPD-A (6.17), lin pool-FPD weights (6.21) and FPD-R (6.22) for known p in Section 6.1. Also we aim to compare FPD-A (6.33), lin pool-FPD weights (6.34) and FPD-R (6.35) for unknown p proposed in Sections 6.3.

6.4.1 Simulation results for $\mathcal{I}_0$'s predictors with known model order p

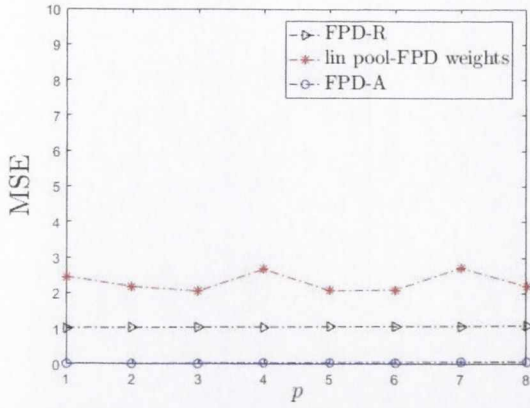
In this part, we wish to examine the performance of different $\mathcal{I}_0$'s predictors with known model order p presented in Section 6.1. For this reason, we consider $s = 10$ AR participants who process their own local data into OSA predictor with three shaping parameters. These shaping parameters are reported to $\mathcal{I}_0$. We assume that all participants and $\mathcal{I}_0$ are $AR(8)$, i.e. $p_l = 8$, $l = 1, \dots, 10$. Also, simulated data are generated from $AR(8)$. Notice that results here are derived in $MC = 1000$ runs. The ideal pds, $F_I(\boldsymbol{\theta}|l, p, K) \equiv NiG(0.01 \times I_{(p+1, p+1)}, 1)$ and $F_I[l] = 1/s$, $l = 1, \dots, s$ are adopted. We examine simulation for operating factors p , s in different figures. Here, the measure of performance is the mean squared error (MSE) defined as

$$MSE = \frac{1}{MC} \sum_{i=1}^{MC} (x^{(i)} - \hat{x}^{(i)})^2, \quad (6.36)$$

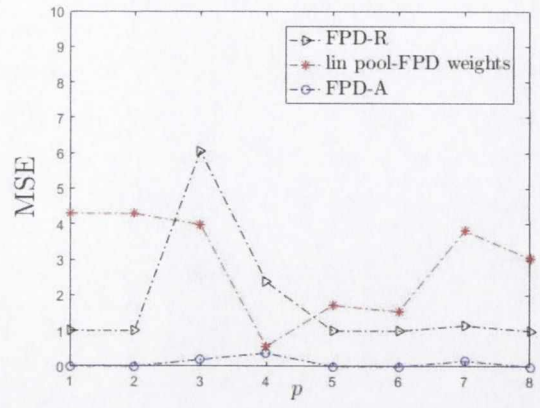
where, $x^{(i)}$ is a simulated value and $\hat{x}^{(i)}$ is the predicted value in the i^{th} -MC run. Notice that the MSE is computed for different approaches by replacing $\hat{x}^{(i)}$ with predicted values of corresponding approach in each run.

In Figures 6.1, parts (a)-(c), we plot the MSE for $\mathcal{I}_0$'s predictors (6.17), (6.21) and (6.22). These Figures are plotted against the model order p with $p_l = 8$, $l = 1, \dots, s$, for $n = 10$ in part (a), $n = 50$ in part (b) and $n = 500$ in part (c). These figures show that the FPD-A of predictor (6.17) outperforms others for $n = \{10, 50, 500\}$, which are plotted in terms of $p = \{1, \dots, 8\}$.

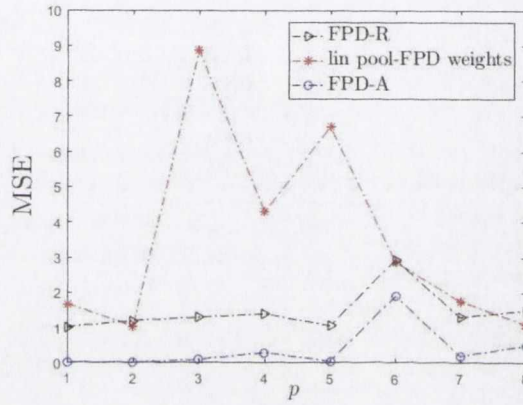
In Figure 6.2 the simulation is designed for $r = 1$, $p = 8$, $n = 100$, $p_l = 5$, $l = 1, \dots, s$ and the simulated data are generated from $AR(8)$. Recall that here participants generate



(a)



(b)



(c)

Figure 6.1. Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.17), lin pool-FPD weights (6.21), and FPD-R (6.22) for $p_l = 8$, $s = 10$. (a) $n = 10$, (b) $n = 50$, (c) $n = 500$.

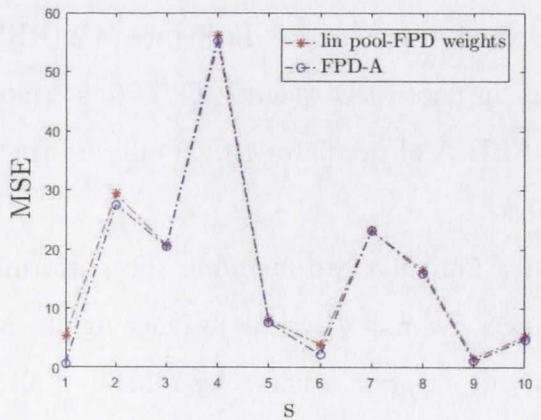
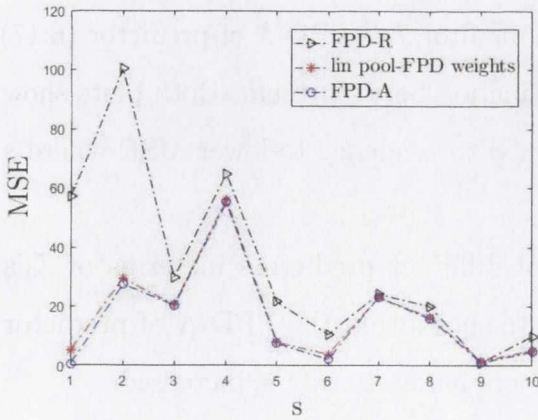


Figure 6.2. In LHS, comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.17), lin pool-FPD weights (6.21), and FPD-R (6.22) for $p_l = 5$, $l = 1, \dots, s$ and $n = 100$ in terms of s . In RHS, MSE for FPD-A (6.17) against lin pool-FPD weights (6.21), with the same operating factor.

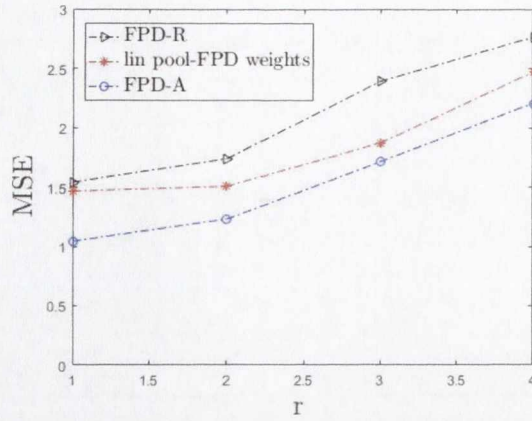


Figure 6.3. Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.17), lin pool-FPD weights (6.21), and FPD-R (6.22) for $p_l = 5$, $p_2 = 10$, $s = 2$ and $n = 100$ in terms of $\mathcal{I}_0$'s variance, r .

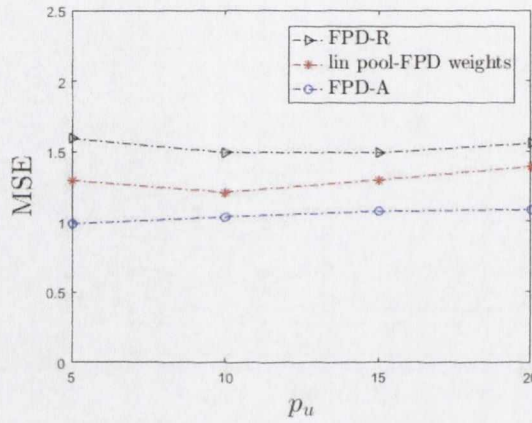


Figure 6.4. Comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.33), lin pool-FPD weights (6.34), and FPD-R (6.35) for $p_l = 5$, $s = 10$, $n = 100$ and $r = 1$ and unknown order p , in terms of p_u .

and process their own local data, then they report shaping parameters to $\mathcal{I}_0$. Here the plot is sketched in terms of s . In Figure 6.2 (RHS), we monitor $\mathcal{I}_0$'s FPD-A of predictor (6.17) versus lin pool-FPD weights (6.21) to see more distinction between them. Both plots show that FPD-A of predictor (6.17) outperforms others with tendency to lower MSE where s increases.

In Figure 6.3, we examine the performance of different predictors in terms of $\mathcal{I}_0$'s variance, i.e. $r = \{1, 2, 3, 4\}$. Once again this figure shows that the FPD-A of predictor (6.17) outperforms others. Also, MSE of all predictors increases as r is increased.

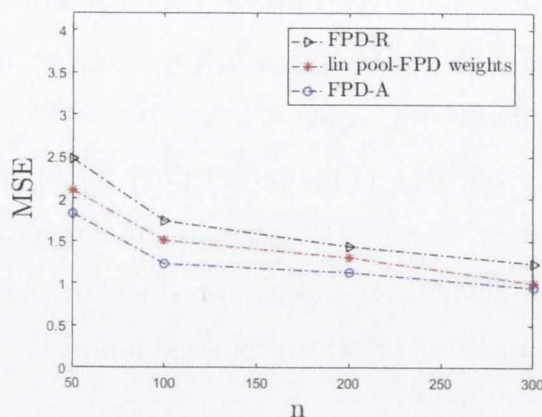


Figure 6.5. comparison of MSE for $\mathcal{I}_0$'s predictors: FPD-A (6.33), lin pool-FPD weights (6.34), and FPD-R (6.35) for $p_l = 5$, $s = 10$, and $r = 1$ with unknown order p , in terms of n .

6.4.2 Simulation results for $\mathcal{I}_0$'s predictors with unknown model order p

In this part, we examine the performance of $\mathcal{I}_0$'s optimal predictors (6.33)-(6.35) for unknown model order p . In Figure 6.4 the simulation is designed for $r = 1$, $n = 100$, $p_l = 5$, $l = 1, \dots, s$, $s = 10$. Once again, the participants locally generate and process their own data, then they are reported to $\mathcal{I}_0$. Here $\mathcal{I}_0$ does not know the model order, p , therefore, the deliberator applies (6.33)-(6.35) in terms of maximum order, p_u . Figure 6.4 shows that FPD-A of predictor (6.33) outperforms others. This figure shows that the MSE of predictors is not visually affected by changing in p_u .

In Figure 6.5 the simulation is designed for $r = 1$, $p_u = 20$, $p_l = 5$, $l = 1, \dots, s$, $s = 10$ in terms of $n = \{50, 100, 200, 300\}$ for the case where p is unknown. Using (6.33)-(6.35), the plot is sketched in terms of n for $p_u = 20$. Figure 6.5 shows that FPD-A (6.33) outperforms others.

6.5 discussion

In this Chapter, the FPD-A of parameter prior (5.16) and FPD-R of parameter prior (5.34) are implemented in AR model. In particular, they are used to compute the optimal predictors for known and unknown p for the AR model. In Section 6.4, it is shown that the FPD-A of predictor (5.16) outperforms FPD-R of predictor (5.34) for the AR model. Here,

$\mathcal{I}_0$'s (OSA) optimal predictors are derived through merging OSA participant' predictors (Section 3.3.2), by $\mathcal{I}_0$. One possible extension of this work is merging multiple MSA predictors which will be explained in Chapter 8.

Notice that FPD-moral deliberator of parameter prior (4.15) has the same functional form with FPD-R of parameter prior (5.34). The main difference between these designs is the evaluation of their normalizing constant, in particular, the evaluation of optimal weights λ^o . The FPD-R of predictor (5.34) is evaluated under,

$$\sum_{l=1}^s \lambda^o = 1, \tag{6.37}$$

while this assumption is not applied in FPD-moral deliberator (4.15). For the cases where the FPD-optimal prior involves normalizing constant with Gamma function, $\Gamma(\cdot)$, we have to solve highly non-linear equations for computation of λ^o . Using the assumption (6.37), the computation of λ^o will be quite easier. On the other hand, FPD-moral deliberator (4.15) does not allow us to exploit the assumption (6.37) for evaluation of λ^o , which requires many other manipulations. Therefore, this prior cannot be easily implemented for the system with multiple AR nodes.

Chapter 7 —

Bayesian filtering using logarithmic pooling design

In this Chapter, we aim to process the knowledge in the context of *Bayesian filtering* (BF). BF is a standard task of data processing where the aim is to infer the state variables (i.e. non-stationary signal) which parametrize a sequence of observation models. BF involves two steps for processing knowledge (i.e. data), which we refer as the *time step* and the *data step*. We will show that the time step of BF usually is intractable except for few special cases i.e. [52, 53, 142]. Therefore, this step is required to be designed or approximated instead of analytically derived. In particular, for the complete model—where the state evolution model is adopted, the distributional approximation is applied. On the other hand, for the incomplete model where state evolution model is not available, the distributional design needs to be implemented. In this Chapter, we focus on the BF problem for the incomplete model where the knowledge K , is in the form of (OSA) alternative state predictors and FPD should be addressed. However, in Section 7.4, we will show that none of the FPD solutions presented in previous Chapters, can be used to handle this task. Instead, we exploit the *stabilized forgetting* (SF) technique for merging a single alternative state predictor and more generally logarithmic pool (4.4) presented in Section 4.2.1 for designing the time step of BF through merging multiple state predictors. Here, the deliberator and participants contribute the pd of time-variant unknown quantities of interest—which we refer as *augmented states* variable, Ψ_n . Then the deliberator wishes to merge these state predictors to design the time step of BF. Using the designed time step of BF and $\mathcal{I}_0$'s time-variant parametric model of x_n , the data step is derived via Bayes' rule.

7.1 Bayesian filtering (BF) under first order Markov property

Here we aim to explore BF under first order Markov assumption. Let ψ_n be the state variable. The first order Markov property is defined as

$$F(\psi_n | \psi_{n-1}, \psi_{n-1}, \dots, \psi_1) = F(\psi_n | \psi_{n-1}) \quad (7.1)$$

The task of BF is to compute dynamically the posterior pd of augmented state variables, $F(\Psi_n | \mathbf{x}_n)$, $n = 1, 2, \dots$, where, $\mathbf{x}_n = [\mathbf{x}_{n-1}, x_n]$ denotes the vector of data. Recall that $\Psi_n = \{\psi_n, \theta_n\}$, where, θ_n could involve the time-variant parameters of the observations—denoted by $\theta_{n,x}$ and the time-variant parameters of the state variable—denoted by $\theta_{n,\psi}$, i.e $\theta_n = \{\theta_{n,x}, \theta_{n,\psi}\}$.

The joint pd of observed data and possible unknowns i.e., $(\psi_n, \theta_{n,x}, \theta_{n,\psi}, \mathbf{x}_n)$ can be factorized in the off-line scenario, under *conditional independency* as,

$$\begin{aligned} F(\psi_n, \theta_{n,\psi}, \theta_{n,x}, \mathbf{x}_n) &= F(\mathbf{x}_n | \psi_n, \theta_{n,x}, \theta_{n,\psi}) F(\psi_n | \theta_{n,x}, \theta_{n,\psi}) F(\theta_{n,\psi} | \theta_{n,x}) F(\theta_{n,x}) \\ &\equiv F(\mathbf{x}_n | \psi_n, \theta_{n,x}) F(\psi_n | \theta_{n,\psi}) F(\theta_{n,\psi}) F(\theta_{n,x}) \end{aligned} \quad (7.2)$$

where,

$$F(\mathbf{x}_n | \psi_n, \theta_{n,x}) = \prod_{i=1}^n F(x_i | \psi_i, \theta_{i,x}) \quad (7.3)$$

$$F(\psi_n | \theta_{n,\psi}) = F(\psi_1 | \theta_{1,\psi}) \prod_{i=2}^n F(\psi_i | \psi_{i-1}, \theta_{i,\psi}) \quad (7.4)$$

$$F(\theta_{n,\psi}) = F(\theta_{1,\psi}) \prod_{i=2}^n F(\theta_{i,\psi} | \theta_{i-1,\psi}) \quad (7.5)$$

$$F(\theta_{n,x}) = F(\theta_{1,x}) \prod_{i=2}^n F(\theta_{i,x} | \theta_{i-1,x}) \quad (7.6)$$

Note that $F(\theta_{1,x})$ and $F(\theta_{1,\psi})$ are prior pds, and (7.3)-(7.6) hold under the following assumptions:

- observation model (under conditionally independent distributed (cid) assumption):

$$F(x_i|\boldsymbol{\psi}_i, \theta_{i,x}, \theta_{i,\psi}, \mathbf{x}_{i-1}) \equiv F(x_i|\psi_i, \theta_{i,x}) \quad (7.7)$$

- state evolution model (under the first order Markov property),

$$F(\psi_i|\boldsymbol{\psi}_{i-1}, \boldsymbol{\theta}_{i,\psi}, \boldsymbol{\theta}_{i,x}, \mathbf{x}_{i-1}) \equiv F(\psi_i|\psi_{i-1}, \theta_{i,\psi}) \quad (7.8)$$

- state-parameters evolution model (under the first order Markov property),

$$F(\boldsymbol{\theta}_{i,\psi}|\boldsymbol{\theta}_{i-1,\psi}, \mathbf{x}_{i-1}) \equiv F(\theta_{i,\psi}|\theta_{i-1,\psi}) \quad (7.9)$$

- observation-parameters evolution model (under the first order Markov property)

$$F(\boldsymbol{\theta}_{i,x}|\boldsymbol{\theta}_{i-1,x}, \mathbf{x}_{i-1}) \equiv F(\theta_{i,x}|\theta_{i-1,x}) \quad (7.10)$$

The graphical model related to (7.7)-(7.10) is presented in Figure 7.1. Notably, Bayesian inference in off-line scenario (7.2) avoids marginalization over unknowns [143]. More interesting scenario is the on-line BF which happens in many practical research area such as control [144], telecommunication systems [61] and target tracking [145], etc. For this purpose, we need to re-factorize the joint pd (7.2) for the on-line scenario. The joint pd (7.2) can be factorized under assumptions (7.7) - (7.10) in the on-line scenario as follows,

$$\begin{aligned} F(\boldsymbol{\psi}_n, \boldsymbol{\theta}_{n,\psi}, \boldsymbol{\theta}_{n,x}, \mathbf{x}_n) = \\ F(x_n|\psi_n, \theta_{n,x})F(\psi_n|\psi_{n-1}, \theta_{n,\psi})F(\theta_{n,\psi}|\theta_{n-1,\psi})F(\theta_{n,x}|\theta_{n-1,x})F(\boldsymbol{\psi}_{n-1}, \boldsymbol{\theta}_{n-1,\psi}, \boldsymbol{\theta}_{n-1,x}, \mathbf{x}_{n-1}) \end{aligned} \quad (7.11)$$

Using (7.11), Bayesian conditional pd of unknowns at times $i = \{n-1, n\}$ conditioning on data is,

$$\begin{aligned} F(\psi_n, \psi_{n-1}, \theta_{n,\psi}, \theta_{n-1,\psi}, \theta_{n,x}, \theta_{n-1,x}|\mathbf{x}_n) \propto \\ F(x_n|\psi_n, \theta_{n,x})F(\psi_n|\psi_{n-1}, \theta_{n,\psi})F(\theta_{n,\psi}|\theta_{n-1,\psi})F(\theta_{n,x}|\theta_{n-1,x})F(\boldsymbol{\psi}_{n-1}, \boldsymbol{\theta}_{n-1,\psi}, \boldsymbol{\theta}_{n-1,x}|\mathbf{x}_{n-1}) \end{aligned}$$

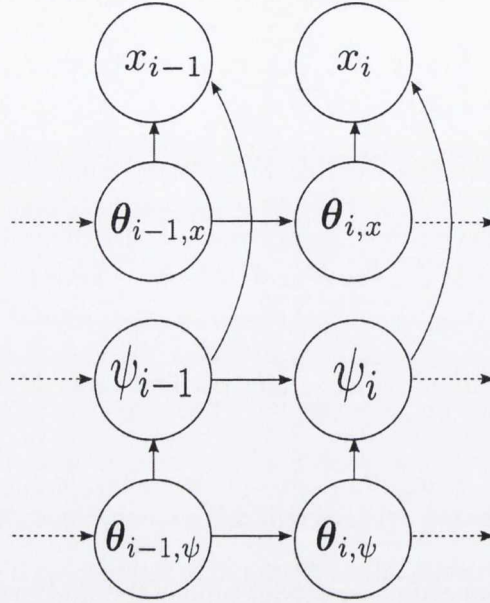


Figure 7.1. Graphical model of unknowns under assumptions (7.7) - (7.10)

(7.12)

where,

$$F(\psi_{n-1}, \theta_{n-1,\psi}, \theta_{n-1,x} | \mathbf{x}_{n-1}) = \int_{\theta_{n-2,x}} \int_{\theta_{n-2,\psi}} \int_{\psi_{n-2}} F(\psi_{n-1}, \theta_{n-1,\psi}, \theta_{n-1,x} | \mathbf{x}_{n-1}) d\psi_{n-2} d\theta_{n-2,\psi} d\theta_{n-2,x}, \quad (7.13)$$

Notice that $\theta_{n-2,x}$, $\theta_{n-2,\psi}$ and ψ_{n-2} are respectively the aggregated parameters of observation model, the aggregated parameters of state model and the aggregated state variables at time $n - 2$.

The BF involves two main steps as follows.

- **The time step:** using (7.12), the time step of BF is,

$$F(\psi_n, \theta_{n,\psi}, \theta_{n,x} | \mathbf{x}_{n-1}) =$$

$$\int_{\theta_{n-1,x}} \int_{\theta_{n-1,\psi}} \int_{\psi_{n-1}} F(\psi_n, \psi_{n-1}, \theta_{n,\psi}, \theta_{n-1,\psi}, \theta_{n,x}, \theta_{n-1,x} | \mathbf{x}_{n-1}) d\psi_{n-1} d\theta_{n-1,\psi} d\theta_{n-1,x} \quad (7.14)$$

- **The data step:** using (7.14) and Bayes' rule, the data step of BF is attained as follows,

$$F(\psi_n, \theta_{n,\psi}, \theta_{n,x} | \mathbf{x}_n) \propto F(x_n | \psi_n, \theta_{n,x}) F(\psi_n, \theta_{n,\psi}, \theta_{n,x} | \mathbf{x}_{n-1}) \quad (7.15)$$

Generally marginalization (7.14) in the time step is not (functionally) tractable. Therefore, a design or approximation of this step is required. Notice that if the integrand (7.14) (i.e. the second, third and fourth term in (7.12)) are explicitly available, then the model is complete and an approximation has to be applied. Otherwise, the model is incomplete and we need to design the time step of BF. Furthermore, the data step of BF (7.15) is not usually derived in the form of the standard pd. In particular, the normalizing constant is functionally intractable except for some special cases where again the distributional approximation needs to be exploited.

Remark 7.1.1. (*Conditional and marginal filtering pds*)

Depending on the context, one might be interested in inferring marginals or conditional filtering pds. Some of important pds are listed as follows,

- using (7.15), the marginal filtering pd of the state parameter, $\theta_{n,\psi}$,

$$F(\theta_{n,\psi} | \mathbf{x}_n) \propto \int_{\theta_{n,x}} \int_{\psi_n} F(\psi_n, \theta_{n,\psi}, \theta_{n,x} | \mathbf{x}_n) d\psi_n d\theta_{n,x} \quad (7.16)$$

- using (7.15), the marginal filtering pd of the state parameter, $\theta_{n,x}$,

$$F(\theta_{n,x} | \mathbf{x}_n) \propto \int_{\theta_{n,\psi}} \int_{\psi_n} F(\psi_n, \theta_{n,\psi}, \theta_{n,x} | \mathbf{x}_n) d\psi_n d\theta_{n,\psi} \quad (7.17)$$

$$(7.18)$$

- using (7.15), the marginal filtering pd of the state variable, ψ_n ,

$$F(\psi_n|\mathbf{x}_n) \propto \int_{\theta_{n,\psi}} \int_{\theta_{n,x}} F(\psi_n, \theta_{n,\psi}, \theta_{n,x}|\mathbf{x}_n) d\theta_{n,x} d\theta_{n,\psi} \quad (7.19)$$

$$(7.20)$$

- the conditional filtering pd of the state variable, ψ_n , conditioning on $\theta_{n,\psi}$,

$$F(\psi_n|\theta_{n,\psi}, \mathbf{x}_n) = \frac{F(\psi_n, \theta_{n,\psi}|\mathbf{x}_n)}{F(\theta_{n,\psi}|\mathbf{x}_n)} \quad (7.21)$$

In the following remark, we discuss some possible extensions of (7.7) and (7.8) by entering a hidden time-invariant pointer, p .

Remark 7.1.2. (Time-invariant pointer)

Each observation model (7.7) can be extended by a discrete scalar time-invariant pointer, p , (with p_u state), i.e. $F(x_n|\psi_n, \theta_{n,x}, p)$. In this case, p is a time-invariant parameter of the conditional observation model for $n = 1, 2, \dots$. Therefore, the observation model (7.7) is marginalised w.r.t. p , which is therefore a p_u -component mixture, regardless of the length of the data.

7.2 Exact BF

Recall that the exact BF is available for some special cases of the complete model. In this section, first we define the concept of the exact BF and mathematically derive the two steps of BF for these special cases. The exact BF is defined as follows.

Definition 7.2.1. Exact BF: BF is exact if; (a) marginalization in the time step (7.14) is tractable and, (b) the data step (7.15), preserves its functional form throughout time $i = 1, 2, \dots$

7.2.1 Kalman Filter (KF)

KF is an important and exact BF which satisfies Definition 7.2.1. KF has been widely used in different applications e.g. communication system [146], robotics [145], car and radar tracking [147], the weather prediction [145] , [148]. In this section, KF assumptions (in multivariate case) are fully explored and the time and data steps of BF in KF type—which we refer to as time and data steps of KF—is derived.

KF assumptions:

- *known assumption of parameters:* in KF, $\theta_{n,x}$ and $\theta_{n,\psi}$ are assumed to be known, therefore, the time step (7.14) and the data step (7.15) of KF is respectively reduced to,

$$F(\psi_n | D_{n-1}) = \int_{\psi_{n-1}} F(\psi_n | \psi_{n-1}) F(\psi_{n-1} | D_{n-1}) d\psi_{n-1}, \quad n = 2, 3, \dots \quad (7.22)$$

$$F(\psi_n | D_n) \propto F(x_n | \psi_n) F(\psi_n | D_{n-1}), \quad n = 1, 2, \dots \quad (7.23)$$

where ψ_n denotes the multivariate states at time n and D_{n-1} is the matrix of observation at time $n - 1$.

- observation model (7.7) and state evolution model (7.8) are binding with normal pd respectively as follows,

$$F(x_n | \psi_n) \equiv N_{x_n}(C_n \psi_n, R_x) \quad (7.24)$$

$$F(\psi_n | \psi_{n-1}) \equiv N_{\psi_n}(A_n \psi_{n-1}, R_\psi) \quad (7.25)$$

where C_n , R_x , A_n and R_ψ are known matrices.

- Furthermore, (7.24) and (7.25) are assumed to be linear in state variables and the associated Wold representation is as follows:

$$\psi_n = A_n \psi_{n-1} + v_n,$$

$$x_n = C_n \psi_n + \epsilon_n,$$

where A_n and C_n are known matrices and v_n and ϵ_n respectively denotes state and observation errors.

Let the KF at $i = n - 1$ be,

$$F(\psi_{n-1}|\mathbf{x}_{n-1}) = N_{\psi_{n-1}}(\boldsymbol{\mu}_{n-1}, \boldsymbol{\Sigma}_{n-1}), \quad (7.26)$$

underlying the KF assumptions in above, the time step (7.22) and the data step (7.23) are respectively,

1. The time step of KF:

$$\begin{aligned} F(\psi_n|\mathbf{D}_{n-1}) &\equiv F(\psi_n|\mathbf{x}_{n-1}) \\ &= N_{\psi_n}(\boldsymbol{\mu}_{n|n-1}, \boldsymbol{\Sigma}_{n|n-1}) \end{aligned} \quad (7.27)$$

where,

$$\boldsymbol{\mu}_{n|n-1} = \mathbf{A}_n \boldsymbol{\mu}_{n-1} \quad (7.28)$$

$$\boldsymbol{\Sigma}_{n|n-1} = \mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T \quad (7.29)$$

2. The data step of KF:

$$\begin{aligned} F(\psi_n|\mathbf{D}_n) &\equiv F(\psi_n|\mathbf{x}_n) \\ &= N_{\psi_n}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \end{aligned} \quad (7.30)$$

where,

$$\begin{aligned} \boldsymbol{\mu}_n &= \mathbf{A}_n \boldsymbol{\mu}_{n-1} + \{(\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \mathbf{C}_n^T [\mathbf{C}_n (\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \mathbf{C}_n^T \\ &\quad + \mathbf{R}_x]^{-1}\} [\mathbf{x}_n - \mathbf{C}_n \mathbf{A}_n \boldsymbol{\mu}_{n-1}], \quad n = 2, \dots \end{aligned} \quad (7.31)$$

$$\begin{aligned} \boldsymbol{\Sigma}_n &= (\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \{ \mathbf{I} - \mathbf{C}_n^T [\mathbf{C}_n (\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \mathbf{C}_n^T \\ &\quad + \mathbf{R}_x]^{-1} \mathbf{C}_n (\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \}, \quad n = 2, \dots \end{aligned} \quad (7.32)$$

Notably, $\boldsymbol{\mu}_n$ (A.24) and $\boldsymbol{\Sigma}_n$ (A.25) are sufficient statistics of $F(\psi_n|\mathbf{x}_n)$ which fully accomplish the task of BF. The mathematical derivations of two steps of KF are presented in Appendix A.7.

Notably, KF assumptions are too restrictive in the world applications. We wish to relax these assumptions so that (a) and (b) in Definition (7.2.1) still hold. A possible extension is employing Remark (7.1.2) in KF context which will be studied in the following.

7.2.2 Mixture Kalman Filter (MKF)

We extend beyond the KF and still maintain tractability of BF solution, namely mixture Kalman filter (MKF). In the MKF, we introduce the unknown, time-invariant pointer, p , (with p_u states) to the observation model (7.24). Without loss of generality, we consider the univariate case with $A_n = C_n = 1$. Then the p -conditional observation model (7.24) becomes:

$$F(x_n|\psi_n, p) \equiv N_{x_n}(\psi_n, r_x^{(p)}), \quad n = 1, 2, \dots, p = 1, \dots, p_u, \quad (7.33)$$

where $r_x^{(p)}$ is the known scalar. Also, the state evolution model becomes:

$$F(\psi_n|\psi_{n-1}) \equiv N_{\psi_n}(\psi_{n-1}, r_\psi), \quad (7.34)$$

where r_ψ is known scalar. Notably, (7.34) can be extended by p as well, but for simplicity, we work with state evolution model (7.34). The graphical scheme of the above structure is presented in Figure 7.2.

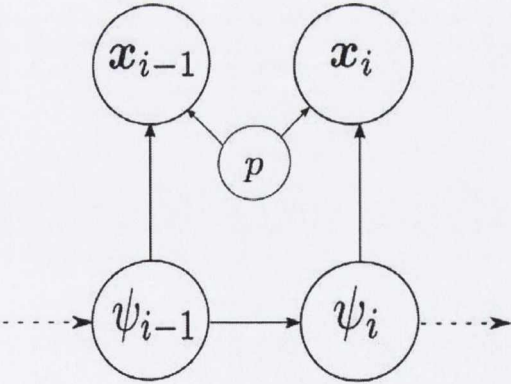


Figure 7.2. Graphical model for MKF

The discrete non-uniform prior of p is adopted as follows,

$$F(p) \equiv \alpha_0^{(p)}, p = 1, \dots, p_u \quad (7.35)$$

where $\alpha_0^{(p)}$ is known a priori. The joint prior pd of (ψ_0, p) is,

$$\begin{aligned} F(\psi_0, p) &= F(p)F(\psi_0|p) \\ &\equiv \alpha_0^{(p)} N_{\psi_0}(\mu_0^{(p)}, \sigma_0^{(p)2}). \end{aligned} \quad (7.36)$$

Using KF calculation in (Section 7.2.1) data step for p -conditional filtering pd is,

$$F(\psi_n|p, \mathbf{x}_n) \equiv N_{\psi_n}(\mu_n^{(p)}, \sigma_n^{(p)2}), \quad (7.37)$$

where,

$$\mu_n^{(p)} = \mu_{n-1}^{(p)} + \left\{ \frac{(r_\psi + \sigma_{n-1}^{(p)2})}{(r_\psi + \sigma_{n-1}^{(p)2}) + r_x^{(p)}} \right\} [x_n - \mu_{n-1}^{(p)}] \quad (7.38)$$

$$\sigma_n^{(p)2} = (r_\psi + \sigma_{n-1}^{(p)2}) \left\{ 1 - \frac{(r_\psi + \sigma_{n-1}^{(p)2})}{(r_\psi + \sigma_{n-1}^{(p)2}) + r_x^{(p)}} \right\}. \quad (7.39)$$

Using the chain rule,

$$\begin{aligned} F(\psi_n, p|\mathbf{x}_n) &= F(\psi_n|p, \mathbf{x}_n)F(p|\mathbf{x}_n) \\ &\propto \alpha_n^{(p)} N_{\psi_n}(\mu_n^{(p)}, \sigma_n^{(p)2}) \end{aligned} \quad (7.40)$$

where $\alpha_n^{(p)} \equiv \frac{\kappa_{\psi_n}^{(p)}}{\sum_{d=1}^{p_u} \kappa_{\psi_n}^{(d)}} = \frac{|\sigma_n^{(p)}|}{\sum_{d=1}^{p_u} |\sigma_n^{(d)}|}$. Therefore, the data step is achieved by marginalization (7.40) over p as follows,

$$\begin{aligned} F(\psi_n|\mathbf{x}_n) &= \sum_{p=1}^{p_u} F(\psi_n, p|\mathbf{x}_n) \\ &= \sum_{p=1}^{p_u} \alpha_n^{(p)} N_{\psi_n}(\mu_n^{(p)}, \sigma_n^{(p)2}) \end{aligned} \quad (7.41)$$

Notably, (7.41) is a mixture of KF (MKF).

Remark 7.2.1 (Kalman filter). *When $p_u = 1$, MKF collapses to the exact KF. The exact KF solution can be similarly obtained from (7.37)-(7.39).*

7.3 A review of Bayesian methods beyond KF

Recall that beyond KF assumptions, the time step of BF is functionally intractable due to the intractability of the marginalization (7.14). Also, specification of the state evolution model i.e. $F(\Psi_n|\Psi_{n-1})$ is the other challenging task. Notice that the incomplete models are the cases where state evolution model is not explicitly specified. Therefore, depending on using complete or incomplete model, either distributional approximation or distributional design is respectively exploited to handle the time step of BF.

Distributional approximation is fundamentally divided into stochastic methods and deterministic approaches presented in the Section 2.4. In BF, these approaches are basically used to preserve the functional form of the approximate pd during the updates. Some classic examples of deterministic approximation for BF are extended Kalman filter [149] and the unscented filter [150]. These filters are applied to linearize non-linear state evolution and observation models. Also Bayesian examples of deterministic techniques are VB approximation i.e. [61, 144, 151]. Deterministic approximation is typically computationally efficient, however the accumulation error grows by time [143]. In against, MCMC and Particle filtering (PF) (i.e. sequential Monte Carlo (SMC) techniques are widely-established as stochastic sampling techniques i.e. [61], [152], [153]. Recall that using these stochastic techniques, the intractable posterior pd (7.15) is approximated by the empirical pd (2.34) as follows:

$$\tilde{F}(\Psi_n|\mathbf{x}_n) = \sum_{i=1}^N w_n^{(i)} \delta(\Psi_n - \Psi_n^{(i)}), \quad (7.42)$$

where $w_n^{(i)}$ is data-driven weights corresponding to i^{th} particle and, $\Psi_n^{(i)}$ is i.i.d sample from *importance* function i.e. [61, 153]. Notice that (7.42) requires an incomplete model where we adopt the approximation. PF techniques have particularly proved popular for global approximation, since they can achieve arbitrarily high accuracy. However they involve the higher computational load. Therefore, deterministic approaches are usually more accurate than PF techniques with availability of the same computational load [61]. Also recent

successful approaches emphasis on exploiting a combination of deterministic and stochastic approaches in order to compromising computational load and accuracy [143], [154].

On the other hand, the distributional design is applied for designing the time step of BF for incomplete model specification. In particular, *stabilized forgetting* (SF)-based design is a well-established method for designing the time step of BF in the literature i.e. [52, 78, 155, 156]. In the rest of this Chapter, we focus on incomplete models of BF and we design the time step of BF through merging (alternative) state predictors.

7.4 Merging state predictors for designing the time step of BF

Recall that our primary methodology of distributional design for merging multiple predictors is FPD. However, the proposed FPD solutions of parameter prior, $A^o(\theta|K)$, presented in the previous Chapters do not allow us to derive the optimal design of $\mathcal{I}_0$'s state predictor i.e. $A^o(\psi_n|K)$, through merging multiple state predictors i.e. $F_l(\psi_n|\mathbf{x}_{n-1}) \equiv F_l(\psi_n|\mathbf{V}_{l,n-1}, \nu_{l,n-1})$, $l = 1, \dots, s$. Since in previous Chapters, we have processed multiple predictors of x , i.e. $G_l(x)$, where the deliberator's parametric model of x follows the case 1 in Remark 2.3.2. In contrast, here we need to process multiple predictors of the state variable where the observation model of x_n needs to be dynamically parametrized with the case 2 in Remark 2.3.2. These predictors are reported in different level of hierarchy. Therefore they require the FPD-optimal design of the hyper-parameter prior i.e. $S^o(\mathbf{V}_{0,n|n-1}, \nu_{0,n|n-1}|K)$ —which is functionally intractable too. For this reason, in place of FPD, we use the other KLD-based method, namely logarithmic pooling design—presented in Chapter 4—to handle the task of BF underlying an incomplete model. In particular, we exploit the SF technique as a special case of the logarithmic pooling design ($s = 1$) for merging alternative state predictor. Notice that the possible extension of FPD for processing knowledge in the context of BF will be discussed in Chapter 8.

7.4.1 SF-based design for merging alternative pd

Recall that distributional pools presented in Section 4.2.1 are popular distributional designs for merging participants' predictors. In particular, logarithmic pooling design can be

exploited for designing the time step of BF in incomplete models. Since it allows us to preserve the functional form of filtering pd through updating, while the linear pool (4.3) does not grantee this condition. Therefore, we will focus on logarithmic pooling design in the rest of this Chapter. SF-based design, as a special case of logarithmic pool, can be directly derived from (4.4). For this reason, let $s = 1$, $w_{l,n}^g = \lambda_n$ in (4.4), and redefining notation i.e. $F_1 \equiv G_1$, then the SF solution for the time step of BF is:

$$A_0^g(\psi_n | \mathbf{x}_{n-1}, \lambda_n) \propto [A_0^g(\psi_{n-1} | \mathbf{x}_{n-1})_{\psi_n}]^{\lambda_n} F_1^{(1-\lambda_n)}(\psi_n | \mathbf{x}_{n-1}), \quad (7.43)$$

Substituting A_0^g (7.43) into (7.15), the data step of BF is achieved via Bayes' rule as follows,

$$A_0^g(\psi_n | \mathbf{x}_n, \lambda_n) \propto F(x_n | \psi_n) A_0^g(\psi_n | \mathbf{x}_{n-1}, \lambda_n) \quad (7.44)$$

Remark 7.4.1. (*Properties of SF*)

- *The SF solution is a local and deterministic technique to design the time step of BF. In particular, forgetting operator, λ_n , commute with Bayes' rule which is quite natural in the case where an explicit time-evolution model (7.8)-(7.10) is not available [155],*
- *The SF solution is justified as an indispensable step in sequential local design [91].*
- *The SF solution is a special case of logarithmic pooling design where $s = 1$.*

In the following, the tractability of the SF solution (7.44) is defined.

Definition 7.4.1. (*Tractable SF*)

The SF solution is tractable if the filtering pd, $A_0^g(\psi_n | \mathbf{x}_n, \lambda_n)$, can be recursively updated via a fixed finite-dimensional (non-sufficient) statistics. This implies that the functional form of the SF solution is invariant over time.

7.5 Tractable SF for Exponential Family

In this section, we aim to relax the modelling assumptions presented in Section 7.2 as far as possible, such that the SF solution (7.44) remains tractable, i.e. it satisfies Definition 7.4.1.

This extends KF context to cases where the observation model belongs to exponential family. This extension allows us to model the capacity of channel, sensor, user, etc. Therefore, we study the SF solution in exponential family in the following.

7.5.1 SF solution in exponential family

Here, the normal assumption of the observation model in KF is relaxed in such a way that the SF solution is tractable. Notably, KF solution is no longer available. For this reason, let state variable, ψ_n , and the time-variant parameter θ_n be unknown and constitute the augmented state variables $\Psi_n = \{\psi_n, \theta_n\}$. Also, let the observation model (7.7) be a member in the exponential family as follows,

$$F(x_n|\Psi_n) \equiv a(\Psi_n)b(x_n) \exp[\langle \phi(\Psi_n), h(x_n) \rangle], \quad n = 1, 2, \dots \quad (7.45)$$

Here, $a(\cdot)$, $b(\cdot)$, $\phi(\cdot)$, and $h(\cdot)$ are known kernels, $\langle \cdot, \cdot \rangle$ is a known two-argument scalar operator. For tractability of the SF solution, we consider the conjugate exponential family (CEF) prior pd at time $n - 1$ as:

$$\begin{aligned} A_0^g(\Psi_{n-1}|\mathbf{x}_{n-1}) &= A_0^g(\Psi_{n-1}|\mathbf{S}_{0,n-1}) \\ &\propto a^{\nu_{0,n-1}}(\Psi_{n-1}) \exp[\langle \phi(\Psi_{n-1}), \tilde{\mathbf{V}}_{0,n-1} \rangle], \end{aligned} \quad (7.46)$$

where the fixed, finite-dimensional statistics, $\mathbf{S}_{0,n-1} = (\mathbf{V}_{0,n-1}, \nu_{0,n-1})$, are shaping parameters at time $n - 1$. An alternative CEF pd is adopted as follows:

$$F_1(\Psi_n|\mathbf{S}_{1,n-1}) \propto a^{\nu_{1,n-1}}(\Psi_n) \exp[\langle \phi(\Psi_n), \mathbf{V}_{1,n-1} \rangle],$$

where $\mathbf{S}_{1,n-1} = (\mathbf{V}_{1,n-1}, \nu_{1,n-1})$. Then the SF solution is computed as follows:

$$\begin{aligned} A_0^g(\Psi_n|\mathbf{x}_{n-1}, \lambda_n) &= A_0^g(\Psi_n|\mathbf{S}_{0,n-1}, \mathbf{S}_{1,n-1}, \lambda_n) \\ &\propto a^{\nu_{0,n|n-1}}(\Psi_n) \exp[\langle \phi(\Psi_n), \mathbf{V}_{0,n|n-1} \rangle], \end{aligned} \quad (7.47)$$

where,

$$\mathbf{V}_{0,n|n-1}(\lambda_n) = \lambda_n \mathbf{V}_{0,n-1} + (1 - \lambda_n) \mathbf{V}_{1,n-1}$$

$$\nu_{0,n|n-1}(\lambda_n) = \lambda_n \nu_{0,n-1} + (1 - \lambda_n) \nu_{1,n-1}$$

Therefore, using (7.45) and (7.47), and applying Bayes' rule, the data step of BF, yields the following filtering pd, which, by construction, has the same CEF form as (7.46),

$$\begin{aligned} F(\Psi_n | \mathbf{x}_n, \lambda_n) &= F(\Psi_n | \mathbf{S}_{0,n}, \lambda_n) \\ &\propto a^{\nu_{0,n}}(\Psi_n) \exp[\langle \phi(\Psi_n), \mathbf{V}_{0,n} \rangle] \end{aligned} \quad (7.48)$$

where $\mathbf{S}_{0,n} = (\mathbf{V}_{0,n}, \nu_{0,n})$, is computed via:

$$\mathbf{V}_{0,n}(\lambda_n) = \mathbf{V}_{0,n|n-1}(\lambda_n) + h(x_n) \quad (7.49)$$

$$\nu_{0,n}(\lambda_n) = \nu_{0,n|n-1}(\lambda_n) + 1 \quad (7.50)$$

From (7.48), the marginal state pds are:

$$F(\psi_n | \mathbf{x}_n, \lambda_n) = \int F(\Psi_n | \mathbf{S}_{0,n}, \lambda_n) d\theta_n \quad (7.51)$$

$$F(\theta_n | \mathbf{x}_n, \lambda_n) = \int F(\Psi_n | \mathbf{S}_{0,n}, \lambda_n) d\psi_n \quad (7.52)$$

7.5.2 SF solution for mixture components of exponential family

In this section, we aim to relax significantly the restrictive assumptions of the observation model in KF, such that the SF solution is still tractable, where the KF solution is not available. This is a useful extension of MKF presented in Section 7.2. The following lemma provides us with a sufficient condition (set of observation models) for which the SF solution can be computed tractably.

Lemma 7.5.1. *A sufficient condition for tractable SF is that the observation model (7.7) be in the form:*

$$F(x_n | \Psi_n, p) \equiv a_p(\Psi_n) b_p(x_n) \exp[\langle \phi_p(\Psi_n), h_p(x_n) \rangle], \quad n = 1, 2, \dots, \quad p = 1, \dots, p_u, \quad (7.53)$$

where $a_p(\cdot)$ and $b_p(\cdot)$ are p -conditional known, scalar kernels, and $\phi_p(\cdot)$, $h_p(\cdot)$ are p -conditional known kernels with compatible dimensions. p is an unknown time-invariant index (model pointer) with p_u states.

Proof. Using (7.47), the time step of SF can be effected as

$$\begin{aligned} A_0^g(\Psi_n|p, \mathbf{x}_{n-1}, \lambda_n) &\equiv A_0^g(\Psi_n|p, \mathbf{S}_{0,n-1}, \mathbf{S}_{1,n-1}, \lambda_n) \\ &\propto a_p^{(\nu_{0,n|n-1}^{(p)})}(\Psi_n) \exp[\langle \phi_p(\Psi_n), \mathbf{V}_{0,n|n-1}^{(p)} \rangle], \end{aligned} \quad (7.54)$$

where,

$$\begin{aligned} \mathbf{V}_{0,n|n-1}^{(p)}(\lambda_n) &= \lambda_n \mathbf{V}_{0,n-1}^{(p)} + (1 - \lambda_n) \mathbf{V}_{1,n-1}^{(p)} \\ \nu_{0,n|n-1}^{(p)}(\lambda_n) &= \lambda_n \nu_{0,n-1}^{(p)} + (1 - \lambda_n) \nu_{1,n-1}^{(p)}. \end{aligned}$$

Also, the data step of SF (7.48) becomes:

$$\begin{aligned} F(\Psi_n|p, \mathbf{x}_n, \lambda_n) &\equiv F(\Psi_n|p, \mathbf{S}_{0,n}, \lambda_n) \\ &\propto a_p^{(\nu_{0,n}^{(p)})}(\Psi_n) \exp[\langle \phi_p(\Psi_n), \mathbf{V}_{0,n}^{(p)} \rangle] \end{aligned} \quad (7.55)$$

where,

$$\mathbf{V}_{0,n}^{(p)}(\lambda_n) = \mathbf{V}_{0,n|n-1}^{(p)}(\lambda_n) + h_p(x_n) \quad (7.56)$$

$$\nu_{0,n}^{(p)}(\lambda_n) = \nu_{0,n|n-1}^{(p)}(\lambda_n) + 1. \quad (7.57)$$

Adopting a non uniform prior pd for p ,

$$F(p) \equiv \alpha_0^{(p)}, \quad p = 1, \dots, p_u,$$

Then, the posterior pd of p is derived as:

$$\begin{aligned} \tilde{\alpha}_n^{(p)} &= F(p|\mathbf{x}_n) \\ &= \frac{\alpha_0^{(p)} \kappa_{\Psi_n}^{(p)}(\mathbf{S}_{0,n}^{(p)})}{\sum_{d=1}^{p_u} \alpha_0^{(d)} \kappa_{\Psi_n}^{(d)}(\mathbf{S}_{0,n}^{(d)})}, \quad p = 1, \dots, p_u \end{aligned} \quad (7.58)$$

where $\kappa_{\Psi_n}^{(p)}(\cdot)$ is the (available) normalizing constant of the recovered p -conditioned CEF

form (7.48). Therefore, the joint filtering pd and its marginals are respectively a mixture of p_u components of (7.48), (7.51) and (7.52), with mixture weights given by (7.58):

$$F(\Psi_n | \mathbf{x}_n, \lambda_n) = \sum_{p=1}^{p_u} \tilde{\alpha}_n^{(p)} F(\Psi_n | p, \mathbf{x}_n, \lambda_n) \quad (7.59)$$

$$F(\psi_n | \mathbf{x}_n, \lambda_n) = \sum_{p=1}^{p_u} \tilde{\alpha}_n^{(p)} F(\psi_n | p, \mathbf{x}_n, \lambda_n) \quad (7.60)$$

$$F(\theta_n | \mathbf{x}_n, \lambda_n) = \sum_{p=1}^{p_u} \tilde{\alpha}_n^{(p)} F(\theta_n | p, \mathbf{x}_n, \lambda_n). \quad (7.61)$$

□

Notably the filtering pds (7.59)-(7.61)—which we refer to the *mixture* of SF (MSF), are in the form of mixture of the SF solution (7.48)-(7.52).

7.6 Illustrative examples and comparative performance under simulation

In this Section, we derive the SF solution for two examples. Then the simulation result is provided for each example. These examples are provided for different reasons as follows:

- In the first example, we derive the SF solution for the case where the observation model is (7.33). Though unnecessary and the exact KF solution is available in this case, it is worthwhile to examine that how much the SF solution is close to KF solution for tracking the true signal.
- In the second example, the SF solution is derived for an incomplete model where the observation is exponential (i.e. a special case of (7.53)). Notably for this case, the exact solution is not available. We examine the performance of the SF solution for tracking the true state in simulation.

Example 7.6.1. *Let the conditional observation model (7.53) be the univariate normal (7.33). Therefore, using (7.47) and (7.48), the SF solution at time n is conjugate normal as,*

$$F(\psi_n | p, \mathbf{x}_n, \lambda_n) \equiv N_{\psi_n}(\tilde{\mu}_n^{(p)}, \tilde{\sigma}_n^{(p)2}) \quad (7.62)$$

specializing, $h_p(\mathbf{x}_n)$ (7.56) with, $[x_n \ 1]^T[x_n \ 1]$, and constructing

$$\tilde{V}_n^{(p)} = \begin{bmatrix} \tilde{V}_{n,11}^{(p)} & \vdots & \tilde{V}_{n,21}^{(p)T} \\ \dots\dots\dots & & \\ \tilde{V}_{n,21}^{(p)} & \vdots & \tilde{V}_{n,22}^{(p)} \end{bmatrix}, \quad \text{SF-designed of EIM}, \quad (7.63)$$

the shaping parameters (7.62) are as follows,

$$\tilde{\mu}_n^{(p)} = \frac{\tilde{V}_{n,21}^{(p)}}{\tilde{V}_{n,22}^{(p)}}, \quad \tilde{\sigma}_n^{(p)2} = \frac{r_x^{(p)}}{\tilde{V}_{n,22}^{(p)}}.$$

Therefore,

$$F(\psi_n|\mathbf{x}_n, \lambda_n) = \sum_{p=1}^{p_u} \tilde{\alpha}_n^{(p)} \cdot N_{\psi_n}(\tilde{\mu}_n^{(p)}, \tilde{\sigma}_n^{(p)2}) \quad (7.64)$$

where,

$$\tilde{\alpha}_n^{(p)} = \frac{|\tilde{\sigma}_n^{(p)}|}{\sum_{d=1}^{p_u} |\tilde{\sigma}_n^{(d)}|}, \quad p = 1, \dots, p_u.$$

Further specialization of (7.53) can be considered when $p_u = 1$. Therefore, the resulting SF solution is for this case attained similarly, as follows:

$$F(\psi_n|\mathbf{x}_n, \lambda_n) \propto N_{\psi_n}(\tilde{\mu}_n, \tilde{\sigma}_n^2)$$

7.6.1 Simulation study: evaluation of the example 7.6.1

We examine the performance of the MSF solution against the exact MKF solution. The results are attained in $MC = 1000$ runs. In Figure 7.3, we design the simulation for comparing the performance of the exact KF solution versus the SF solution (with three values of forgetting factor $\lambda = \{0.7, 0.8, 0.9\}$, in the normal case, with $r_x = r_\psi = 1$, and the diffuse prior $\tilde{\mu}_1 = 0$, $\tilde{\sigma}_1^2 = 1$ is adopted. Figure 7.3 shows *reconstruction* of (true) simulated state, ψ_n . It is clear that the exact KF solution outperforms others for tracking ψ_n , while the SF solution with $\lambda = 0.9$ also performs well. This is because the variation of the state r_ψ is small (slowly time-varying). The upper and lower Bayesian interval, $[E(\psi_n|\mathbf{x}_n) - \sqrt{\text{var}(\psi_n|\mathbf{x}_n)}, E(\psi_n|\mathbf{x}_n) + \sqrt{\text{var}(\psi_n|\mathbf{x}_n)}]$, around the SF solution with $\lambda = 0.9$ is also displayed. In Figure 7.4, we design the simulation with $r_\psi = 1$, $p_u = 5$ and $C^{(p)} = \{1.2, 1.4, 1.6, 1.8, 2\}$. The data are generated from $N(\psi_n, 2)$ and the state evolution

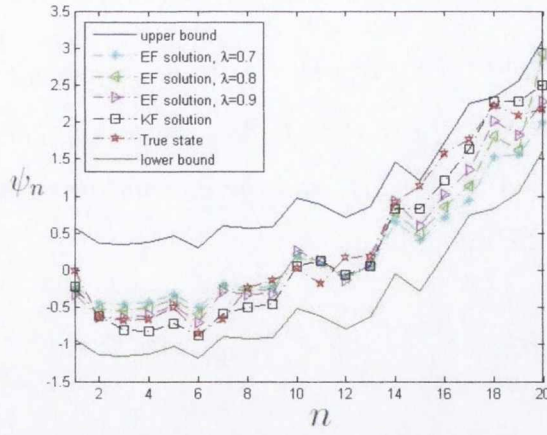


Figure 7.3. State reconstruction using SF versus KF in normal case $r_x = 1$, $r_\psi = 1$.

model is $N(2\psi_{n-1}, 1)$. We adopt the diffuse prior $\tilde{\mu}_1 = 0$, $\tilde{\sigma}_1^2 = 1$. In this figure, the performance of the MKF solution (7.41) versus MSF (7.64) solution is compared. MSF with the highest λ_n outperforms other. This is because the state is again slow-varying, and therefore, the maximum value of λ_n best matches the data in this case. Nevertheless, most of the inferred state trajectories remain inside the standard interval for all tested values of λ_n .

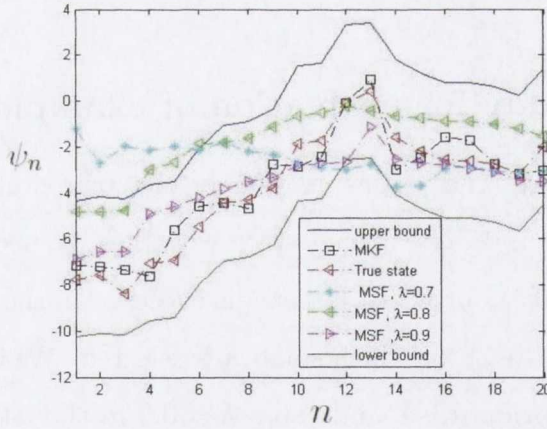


Figure 7.4. State reconstruction using MKF vs MSF (normal case).

Example 7.6.2. (*Exponential Mixture with Time-Variant Rate*)

In this example, the exponential observation model as a special case of the (7.53) is investigated. It models, for example, the inter-arrival times of particles released by a fixed mass of a radio-isotope, known to be one of p_u possible species ($p = 1, \dots, p_u$),

$$F(x_n|\psi_n, p) \stackrel{cid}{=} \mathcal{Exp}(\psi_n^{(p)}), \quad x_n > 0, \quad \psi_n^{(p)} > 0$$

$$a_p(\psi_n) = \phi_p(\psi_n) = \psi_n^{(p)}, \quad h_p(x_n) = x_n.$$

where $\psi_n^{(p)}$ is the time-varying rate parameter of the exponential observation model for the p^{th} species, and $a_p(\cdot)$ and $\phi_p(\cdot)$ are the exponential family kernels (7.53). The CEF pd (7.46) at time $n - 1$ is $Ga(\tilde{\nu}_{n-1}^{(p)}, \tilde{V}_{n-1}^{(p)})$, and the alternative pd is also chosen as $Ga(\cdot)$. Therefore using (7.48), the SF solution is:

$$F(\psi_n | p, \mathbf{x}_n, \lambda_n) \propto [\psi_n^{(p)}]^{\tilde{\nu}_n^{(p)}} \exp[\langle \psi_n^{(p)}, \tilde{V}_n^{(p)} \rangle] \quad (7.65)$$

$$\tilde{V}_n^{(p)} = \tilde{V}_{n|n-1}^{(p)} + x_n$$

$$\tilde{\nu}_n^{(p)} = \tilde{\nu}_{n|n-1}^{(p)} + 1$$

Furthermore, substituting (7.65) into (7.60), the MSF solution can be deduced as

$$F(\psi_n | \mathbf{x}_n, \lambda_n) \propto \sum_{p=1}^{P_u} \{ \tilde{\alpha}_n^{(p)} [\psi_n^{(p)}]^{\tilde{\nu}_n^{(p)}} \cdot \exp[\langle \psi_n^{(p)}, \tilde{V}_n^{(p)} \rangle] \}$$

where $\tilde{\alpha}_n^{(p)}$ is obtained by substituting the normalizing constant, $\kappa_{\psi_n}^{(p)}(\tilde{V}_n^{(p)}, \tilde{\nu}_n^{(p)}) = \frac{\Gamma(\tilde{\nu}_n^{(p)})}{(\tilde{V}_n^{(p)})^{\tilde{\nu}_n^{(p)}}}$, into (7.58).

7.6.2 Simulation study for evaluation of example 7.6.2

Figures 7.5 and 7.6 illustrate the reconstruction of the true state, ψ_n , using the SF solution in the case of slow and fast-varying state variables, respectively. In Figure 7.5, $F(\ln(\psi_n) | \ln(\psi_{n-1})) = N(0, 0.5)$, for which larger λ is chosen. On the other hand, in Figure 7.6, $F(\ln(\psi_n) | \ln(\psi_{n-1})) = N(0, 2)$ for which small λ is selected. We infer (reconstruct) the state using $\lambda = 0.9$ in the former case, and using $\lambda = 0.2$ in the latter case. Once again, the true state trajectory is dominantly contained within the inferred Bayesian interval estimate of the state.

7.7 Optimal tuning of the forgetting factor (FF), λ_n

We aim to assign an optimal data-driven FF sequentially i.e. $\lambda_n = \lambda(\mathbf{x}_n)$ involving a fixed, finite-dimensional, optimal projection of the available data, $\mathbf{x}_n$ at each time. In general, λ_n

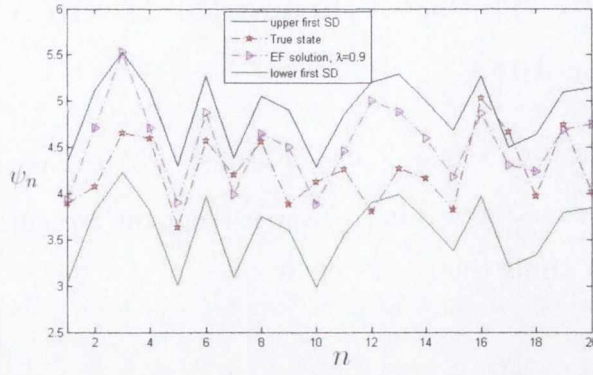


Figure 7.5. State reconstruction using SF for the exponential observation model in the slowly-varying state case.

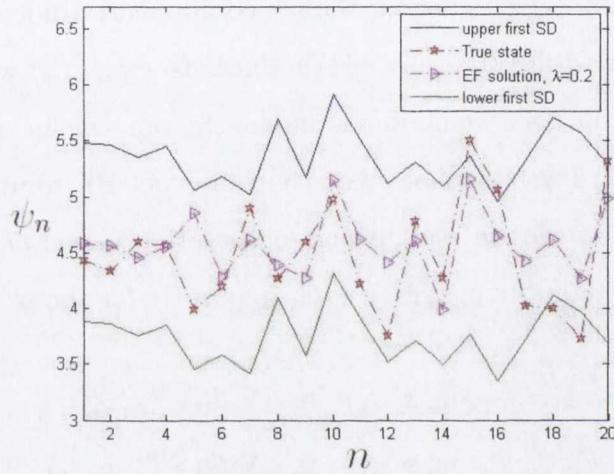


Figure 7.6. State reconstruction using SF for the exponential observation model in the fast-varying state case.

is unknown, but note that updating the SF-based (7.49) and (7.50) depends on the choice of $\lambda_n, \forall n$. Available approach to sequential data-driven FF assignment is [157], which infers sequential λ_n using an iterative variational Bayes (IVB) algorithm. Here we do not redreive caculations and directly use the result [157] i.e. an IVB-based algorithm which sequentially estimates λ_n . Here, λ_n is interpreted as an unknown parameter, assigned a uniform prior pd on $[0, 1]$, and the approximate VB marginal posterior pd is then sequentially computed. The data-driven assignment of λ_n is chosen as the expected value of this VB approximate posterior pd, which we explicitly denote by $\hat{\lambda}_n \equiv \hat{\lambda}_n(\mathbf{x}_n)$. Recall that λ_n enters (7.49) and (7.50), and so an iterative algorithm needs to be implemented at each step n .

7.7.1 Application: SF-based design for BF in a multiple signal-dependent noise problem

In this section, we investigate a signal-dependent noise problem as an application. Notice that here the FF is designed using IVB in [157]. We adopt the normal model, with unknown time-variant variances. Assume that,

$$r_{n,x}^{(p)} = c^{(p)} r_{n,\psi}, \quad p = 1, \dots, p_u, \quad (7.66)$$

where $r_{n,\psi}$ is the unknown variance of a normal random walk for the state variable, ψ_n , and $c^{(p)}$ is a time-invariant known gain for the p^{th} component. Under assumption (7.66), the dimension of state variables $(\psi_n, r_{n,\psi}, r_{n,x}^{(p)})$ reduces to $(\psi_n, r_{n,\psi})$ with the presence of time-invariant index p . Note that we no longer model the time-evolution of the augmented state variables $(\psi_n, r_{n,\psi})$ using Gaussian pd, so that an exact BF solution is not available in the current context. Consider the conditional observation model (7.53), and adopt the NiG pd as the filtering pd and its alternative at time $n - 1$ for the p^{th} component:

$$F(\psi_{n-1}, r_{n-1,\psi} | p, \mathbf{x}_{n-1}) \equiv NiG(\tilde{V}_{n-1}^{(p)}, \tilde{\nu}_{n-1}^{(p)}) \quad (7.67)$$

$$F_1(\psi_n, r_{n,\psi} | p, \mathbf{x}_{n-1}) \equiv NiG(\bar{V}_{n-1}^{(p)}, \bar{\nu}_{n-1}^{(p)}) \quad (7.68)$$

Therefore, SF design under IVB design of FF, at time n is attained as

$$F(\psi_n, r_{n,\psi} | p, \mathbf{x}_n, \hat{\lambda}_n) = NiG(\tilde{V}_n^{(p)}, \tilde{\nu}_n^{(p)}) \quad (7.69)$$

where

$$\tilde{V}_n^{(p)} = \hat{\lambda}_n \tilde{V}_{n-1}^{(p)} + (1 - \hat{\lambda}_n) \bar{V}_{n-1}^{(p)} + \frac{1}{c^{(p)}} [x_n, 1]^T [x_n, 1]$$

$$\tilde{\nu}_n^{(p)} = \hat{\lambda}_n \tilde{\nu}_{n-1}^{(p)} + (1 - \hat{\lambda}_n) \bar{\nu}_{n-1}^{(p)} + 1$$

Applying (7.69) into (7.60) and (7.61), the marginal pd of ψ_n is the mixture of p -conditional Student's-t pds and the marginal pd of $r_{n,\psi}$ is the mixture of p -conditional iG pds.

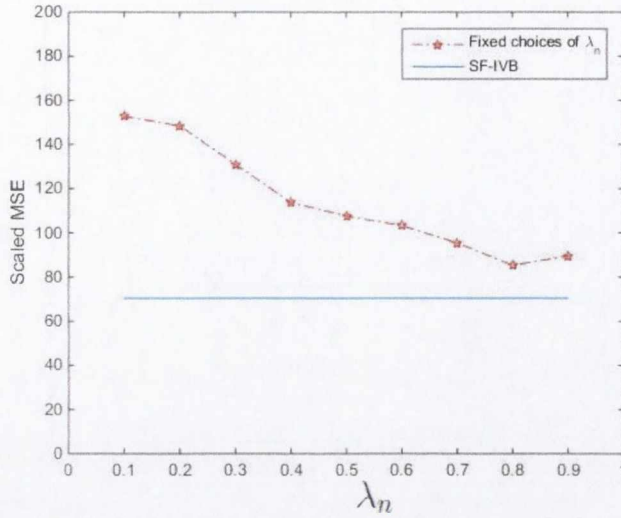


Figure 7.7. Comparison of scaled mean squared error for fixed choices of λ versus SF based on VB design: slowly-varying case.

7.7.2 Simulation study: multiple single-dependent noise problem

In this section, we compare the performance of the SF solution based on IVB design versus fixed choices of FF. Here, the results are evaluated based on 1000 MC runs. We design the simulation with $\psi_0 = 0.2$, $u_0 = 1$, $c^{(p)} = 1.5$, $p_u = 5$ for slow-varying (so that $r_{n,\psi}$, $n = 1, 2, \dots$, is kept small) hidden states. Note that an optimal fixed choice of FF is not available in general, but can be noted in the current Monte Carlo context. Figure 7.7 shows that $\lambda = 0.8$ (for the current simulation conditions) has the minimum scaled mean squared error (SMSE), defined as

$$SMSE(\lambda) = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \left(\frac{\psi_{ij} - E(\psi_{ij} | \mathbf{X}_{ij}, \lambda)}{\psi_{ij}} \right)^2, \quad (7.70)$$

where N is the number of time steps and M is the number of Monte Carlo runs. This optimal value of fixed FF is chosen for the purposes of comparison with the data-driven, variable assignment arising with VB. Also $F(\psi_n | \psi_{n-1}) = N(\psi_{n-1}, 0.7)$ and $F(u_n | u_{n-1}) = \ln N(\ln u_{n-1}, 0.9)$ is adopted in the simulation. In Figure 7.8, SF for optimal fixed choice of FF, $\lambda_n = 0.8$, is compared with the data-driven IVB assignment, with the latter outperforming the former. In fact, the VB-based assignment outperforms all fixed choices of FF. In Figures 7.9 and 7.10, we design the simulation with $F(\psi_n | \psi_{n-1}) = N(\psi_{n-1}, 3)$ and $F(u_n | u_{n-1}) = \ln N(\ln u_{n-1}, 4)$ for fast-varying ($r_{n,\psi}$ is kept large) hidden states. Figure 7.9

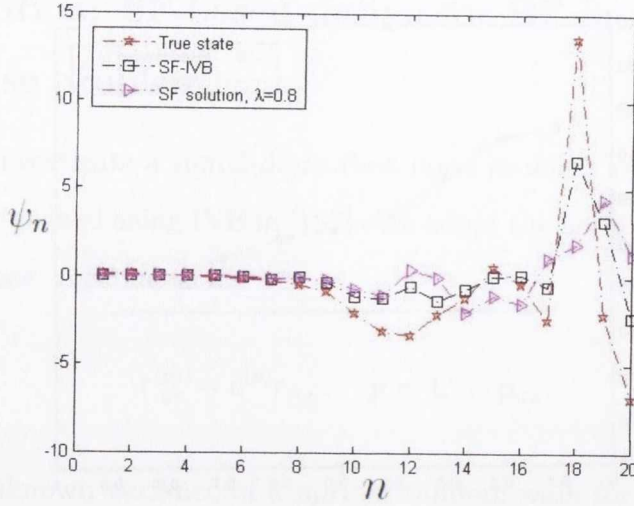


Figure 7.8. SF-based reconstruction of the state using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.8$: slow-varying case.

shows that $\lambda = 0.2$ minimizes the criterion (7.70) for different fixed choices of FF in this case.

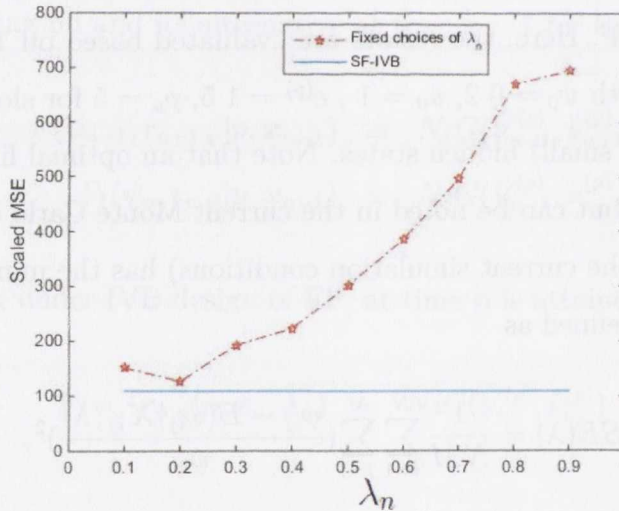


Figure 7.9. Comparison of scaled mean squared error for fixed choices of λ versus SF based on VB design: fast-varying case.

Comparison of SF for the optimal fixed choice of FF versus the data-driven VB-based sequential assignment of the FF in the fast-varying case is illustrated in Figure 7.10. For $n < 16$, we do not see a clear difference for the alternative schedules due to the large scale of differences for $n \geq 15$. Figure 7.10 shows SF design with the VB assignment of FF outperforms the solution for all fixed choices of FF, even in the case when the optimal

value, $\lambda = 0.2$, is chosen. In Figures 7.11 and 7.12, we assess performance of SF design

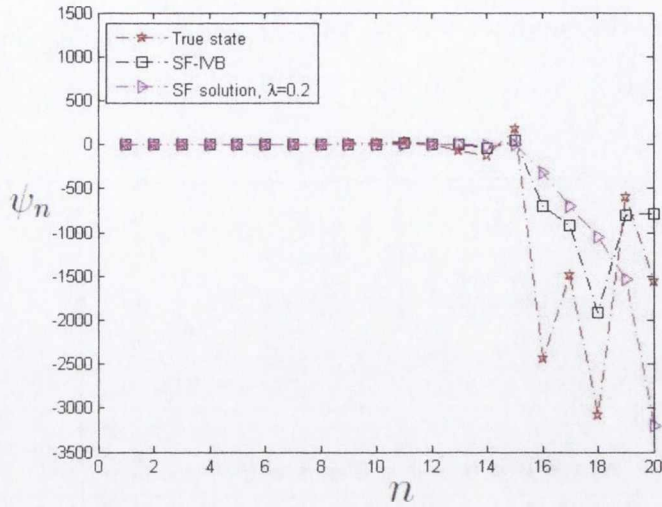


Figure 7.10. SF-based reconstruction of the state using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.2$: fast-varying case.

for reconstructing state variable u_n , under different schedules for assignment of the FF. Figure 7.11 illustrates a slow-varying case for u_n ($F(u_n|u_{n-1}) = \ln N(\ln u_{n-1}, 0.9)$) where the SF design with a data-driven VB assignment of FF outperforms that using an optimal fixed value of FF, $\lambda_n = 0.8, \forall n$. For increasing n , this distinction becomes ever more evident. Also the lower and upper first standard deviations (SD) are depicted, providing

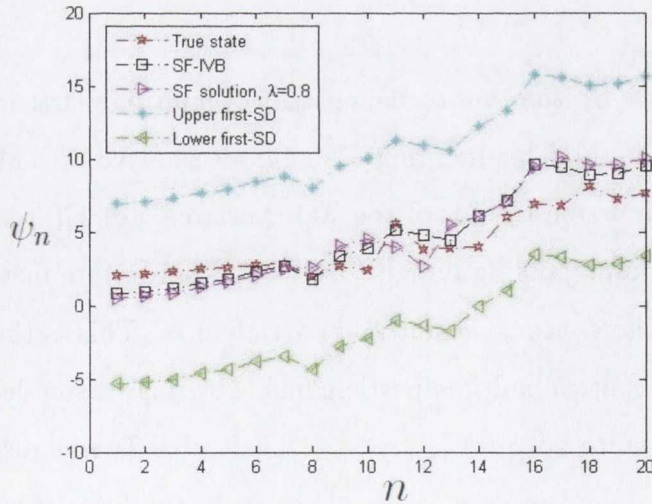


Figure 7.11. SF-based reconstruction of u_n using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.8$: slow-varying case.

Bayesian intervals that summarize the SF-design of sequential filtering pds for u_n , for

the case of the data-driven VB design of the FF. In figure 7.12, we compare the different schedules for a fast-varying case of u_n ($F(u_n|u_{n-1}) = \ln N(\ln u_{n-1}, 2.5)$). Once again, the SF design under VB-based assignment of the FF outperforms the case where the optimal fixed choice of FF ($\lambda_n = 0.2, \forall n$) is adopted.

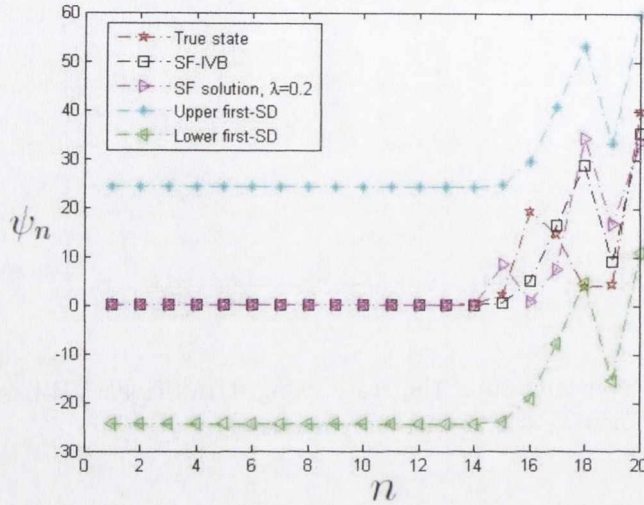


Figure 7.12. SF-based reconstruction of u_n using data-driven VB-based design of the FF versus an optimal fixed choice $\lambda_n = 0.2$: fast-varying case.

7.8 The logarithmic pooling design for incomplete models in MP

Here we aim to extend the SF solution to the context of multiple participants. Notice that SF is the design-based method for merging a single alternative pd. More generally, we might merge the multiple predictive pds of the state variable, i.e. $F_l(\psi_n|\mathbf{x}_{n-1}), l = 1, \dots, s$. We note that the logarithmic pooling design (4.4) is an appropriate method for designing the time step of BF in the context of multiple participants. This is the extension of SF (7.43) from single participant to multiple participants. For this reason, let each participant reports the time-variant state pd $F_l(\psi_n|\mathbf{x}_{n-1}), l = 1, \dots, s$, to $\mathcal{I}_0$ who also contribute their time-invariant filtering pd, and process these pds to reach the logarithmic pool as follows,

$$\begin{aligned} A_0^g &\equiv A_0^g(\psi_n|\mathbf{x}_{n-1}, \mathbf{w}_n^g) \\ &\propto [A_0^g(\psi_{n-1}|\mathbf{x}_{n-1})_{\psi_n}]^{w_{0n}^g} \prod_{l=1}^s [F_l(\psi_n|\mathbf{x}_{n-1})]^{w_{ln}^g}, \end{aligned} \quad (7.71)$$

where $w_{ln}^g \in \Delta_s$, i.e. $\sum_{l=0}^s w_{ln}^g = 1$. Using observation model (7.7) and applying (7.71) into BF, the data step of BF (7.15) is computed.

7.9 Discussion

In this chapter, we processed the probabilistic knowledge (i.e. pd of state variables) and data (i.e. x_n , $n = 1, 2, \dots$), in order to reach a tractable design of filtering pd in the context of BF. In particular, we employ SF as a special case of the logarithmic pool (4.4) for designing the time step of BF for a sufficiently rich class of observation models (7.53) presented in Lemma 7.5.1.

The four classes of solutions for BF i.e., $A, \dots, D$, are summarized in Figure 7.13. In particular, the boundary of each set of BF solution with related class of observation model—denoted inside $[\cdot]$ —is sketched in this Figure. Notably, the sufficiently large set of the observation model is the mixture of the exponential families, i.e. the class D .

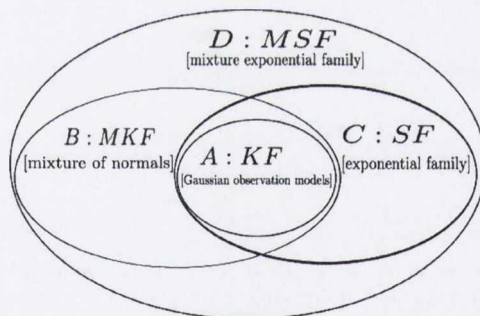


Figure 7.13. Schematic boundary of different filters (solutions) denoting with A-D. For each filter, the associated class of observation models is expressed in $[\text{observation model}]$.

Lemma 7.5.1 remains true for different design of the time-variant forgetting factor, λ_n . Although a data-driven design of the forgetting factor using IVB is presented in Section (7.7), there is the interest in the design of a schedule for λ_n using more elegant and principled design. As mentioned, we specified the SF solution as a special case of the distributional logarithmic pool. Therefore, various approaches for designing the weights in logarithmic pool can be adopted for designing λ_n . In particular, the elegant empirical design presented in Section 4.2.1.1 could be used. Several proposals are available in the Bayesian literature

such as methods presented in [48, 91].

The lemma 7.5.1 does not claim to provide a necessary condition for tractability of SF solution of BF, and, indeed, we will explore further extensions to time-variant parameter models for which tractability is preserved in future work.

SF has been adopted as a design of the time step in BF for state variables, Ψ_n , where the latter are augmented by the time-variant model parameters. SF has ensured that fixed, finite-dimensional statistics are maintained in the resulting filtering algorithm. In Section 7.7, the time-variant forgetting factor (FF) has been assigned optimally in a sequential data driven scheme, $\hat{\lambda}_n \equiv \hat{\lambda}_n(\mathbf{x}_n)$, using an iterative VB algorithm. A FPD-based proposal for a data-driven FF [91] will be explored in future research. Furthermore, we have restricted our focus to observation models with a mixture of p_u components, $\forall n$. This is the case when the pointer, p , is time-invariant. In future work, we will relax this simplification, adopting a time-variant pointer p_n , in which case the number of components increases as p_u^n , i.e. exponentially with n . Sequential design strategies are then required in order to bound the dimension of the model, and of the computed statistics.

The task of BF —addressed in this Chapter — remains to be done using FPD. In Chapter 8, we will discuss more technical details on how it might be developed.

Chapter 8 —

Conclusions and Future Work

In this Chapter, we summarize the core results of each Chapter i.e. Chapters 3-7. Some possible extensions of the future work corresponding to each chapter will be discussed in sequel. Finally we summarize key contributions of this work and remaining tasks are listed.

8.1 Conclusion of Chapter 3

In this Chapter, we process data (as a special case of knowledge processing) to deduce the unique and consistent Bayesian inference of the model order, the subset selection of the global set of unknown parameters, and OSA and MSA predictors for AR model. Furthermore, recursive algorithms for processing data into shaping parameters (i.e. EIMs) of posterior pd are proposed. Widely used criteria of model order estimation are discussed and specialized into the AR case. Moreover, automatic Ockham's regularization of BPO has been mathematically justified and a comprehensive explanation of emerging Ockham's term in estimation strategies of model order in AR case is provided. Furthermore, recursions introduced in Section 3.6 provide us efficient tools to compute the Bayesian data processing task. These recursions might be the good replacement of LDL decompositions which is popular for Bayesian computation in the current Bayesian literature i.e. [55, 91].

8.1.1 Possible extensions and future directions of Chapter 3

Recall that recursion in order (3.108), the recursion in time (3.111) and full recursion (3.112) at the level of EIMs (Section 3.6), allow us efficiently process data into shaping parameters of pds for computing different Bayesian inferential tasks in Section 3.3. However, these recursions do not provide a recursive updating of pds. Therefore, an important progress

will be exploring efficient recursions at the level of projection of EIMs in the future.

8.2 Conclusion for Chapter 4

In Chapter 4, the centralized moral deliberator-based scheme allows consistently processing distributed knowledge in multiple participants. This scheme is (i) parametric, and (ii) moral. Under (i), the participants' knowledge is merged into a prior on the deliberator's parameter. This prior is chosen optimally, using FPD. Under (ii), this optimal prior bounds the expected loss imposed on each participant. The scheme allows for a wide class of continuous parametric deliberators and participants to be handled. It quantifies uncertainty in the deliberator's parametric model, while also yielding the optimal predictor. This chapter has provided functionally tractable solutions in the conjugate case, and, specifically, for normal participants and deliberator. Extensive simulations point to the advantages of the scheme over logarithmic pooling, particularly in respect of uncertainty quantification, randomization, and predictive accuracy.

8.2.1 Possible extensions and future directions of Chapter 4

In this Section, some potential future tasks of Chapter 4 are pointed out. Despite the FPD-optimal parameter pd (4.15) and the optimal predictor (5.18) can be functionally tractable for a rich class of $\mathcal{I}_0$'s parametric model of x , $G_0(x|\boldsymbol{\theta})$, generally they are not computationally tractable. Recall that the main issue is deriving $\boldsymbol{\lambda}^\circ$ as an explicit function of $\boldsymbol{\beta}$ and K . Some of the well-known numerical techniques are *penalty function approximation* [137] and the *grid search* [158]. Penalty function approximation has been considered for similar tasks [38], and might be extended to our more general context. Also, the grid search is a popular algorithm for hyperparameter optimization where the hyperparameters are tuned in the space of hyperparameters. We would not aim to implement this algorithm. Because, KT multiplier $\lambda_l \geq 0$, $l = 1, \dots, s$, which implies that hyperparameter space includes unbounded value spaces for $\boldsymbol{\lambda}$. Therefore, discretization of space may be necessary before applying the grid search. Furthermore, the grid search alone does very poorly in practice [159].

Generally, it is difficult for the deliberator to specify their bounds, $\boldsymbol{\beta}$ (4.10), on the

expected loss imposed on the respective participants. To solve this issue, we might re-interpret λ as the primary parameter of the deliberator's model for x . In this case we can adopt an ideal pd (pre-prior) to λ . There is then the opportunity to derive the FPD-optimal design of λ through processing K . A similar idea is adopted for the logarithmic pool in [160].

Adopting an ideal pd of λ represents the quantification of $\mathcal{I}_0$'s opinion via hyperparameters. Then it will be updated after processing K , i.e. $A^o(\lambda|K)$. In the future work, this task will be fully explored.

8.3 Conclusion for Chapter 5

In Chapter 5, different knowledge, K , is defined and processed for designing the FPD-optimal solutions. In particular, K in Definition (5.1.1) is no longer constrained to the bounded expectation of KLDs (i.e. constraint 3 in Definition 4.3.1). This K is processed in three different ways with three different specification of ideal pd. Each way of processing K yields the different FPD solution. In particular, FPD-A (5.16) in Section 5.2 and FPD-R (5.34) in Section 5.4 are one-batch processing of K . While FPD-S (5.22) in Section 5.3 is the sequential processing of participants' predictors where FPD is applied s times. We note that FPD-A (5.16) yields a principled and tractable design of non-uniform weights. While the design (5.34) requires numerical techniques for evaluation of weights. Furthermore, FPD-S (5.22) assigns the unit weight to each participant. The main difference between (5.16) and (5.34) is that the further assumption (5.31) on the joint model (5.6) is adopted for deriving (5.34). A thorough comparison will be presented in Table 8.1. Notably these optimal priors can be used for sequential processing of data by $\mathcal{I}_0$.

We favour FPD-A (5.16) over FPD-S (5.22) and FPD-R (5.34) for the reasons mentioned in Section 5.8. The optimized prior (5.16) is exploited to deduce the optimal predictor. Specifically, the optimal predictor (5.18) is in the form of the linear mixture for which both components and weights are tractably evaluated for a rich class of $\mathcal{I}_0$'s parametric model, i.e. DEF. On the other hand, $\mathcal{I}_0$'s predictor (5.30) and (5.44) have the similar structure to (4.21) which is totally different from the optimal predictor (5.18).

8.3.1 Possible extensions and future directions of Chapter 5

The proposed scheme in Chapters 4 and 5, indicates that participants provide their knowledge of same x to the deliberator, $G_l(x)$. A possible extension of these works is to merge pds of partially common $\mathbf{x}$, reported by participants. A related work is provided in [38] for the discrete case. Also [161, 162] can be used where the joint pd of multiple transformations of x (the output of multiple distributed sensors) is constructed using hypothesis testing for the purpose of classification. In particular, the constructed pd is asymptotically optimal, i.e. the closest pd to the true pd in KLD sense. We will explore merging multiple pds in more general cases in the future work. The other possible future exploration is to implement the proposed FPD-optimal design for the recursive schemes where the $\mathcal{I}_0$'s deliberation is fed back to the participants in a closed-loop scheme.

The FPD-optimal design presented in Chapters 4 and 5 is a static, open-loop one, involving a single invariant unknown, x . The framework can be extended to the dynamic case, involving modelling of a discrete-time random process, x_i , $i \in \mathbb{Z}$. In particular, we will show in future work how the deliberator's FPD-optimal inference may be fed back to the participants as a prior for their (conjugate) processing of local data, x_i^l , $l = 1, \dots, s$, achieving an FPD-optimal, closed-loop, Bayesian distributed filtering scheme. The related work is found in [111] in the case where the scheme is decentralized and the weights are informally derived.

The FPD-optimal design (5.16) and (5.34) are achieved under this assumption that just one of the participants is true. However, this is not applied for more general contexts where two or more participants might be true. This generalization will be explored (e.g. by rearranging knowledge) in the future work. Also, in Chapter 5, the deliberator fully respects to one participant and replace their parametric pd with the participant's pd (5.4). A possible direction of the future work is to generalize the context to the case where the deliberator considers the participants' predictors partially and involves their parametric model in formulating their joint model (5.6) i.e. replacing the assumption (5.4) with,

$$M(x|l, \boldsymbol{\theta}, K) \equiv \alpha G_l(x) + (1 - \alpha) G_0(x|\boldsymbol{\theta}) \quad (8.1)$$

or,

$$M(x|l, \boldsymbol{\theta}, K) \propto G_l^\alpha(x) G_0^{1-\alpha}(x|\boldsymbol{\theta}). \quad (8.2)$$

Recall that FPD designs (5.16), (5.22) and (5.34) do not propose non-uniform weights for the case where $s = 1$. We note that this problem will be solved applying (8.1) in (5.32). The FPD-optimal parametric hierarchy of the deliberator in Chapters 4 and 5, constraints the model choice, $G_0(x|\boldsymbol{\theta})$, and forms an important part of $\mathcal{I}_0$'s knowledge, K . In particular, this is the application of item 1 Remark 2.3.2. An important parametric extension is adopting item 2 Remark 2.3.2. Using that, we can provide the uncertainty measure on the $\mathcal{I}_0$'s optimal predictor, where this is not currently available. More important progress will be made when the parametric form can be fully relaxed to a continuous nonparametric form, $\boldsymbol{\theta} \equiv G_0$, so that $x \sim G_0(x)$, where $G_0(x)$ is now a functionally unknown (i.e. nonparametric) pd. In particular, [19] develops the FPD framework for the design of priors in nonparametric hierarchies, when processing external knowledge of a very general kind. In future work, we will apply that framework in MPS. This will also yield uncertainty measures in respect of the deliberator's predictor, $G_0(x|K)$ (Remark 5.3.2).

Despite the advantages of our centralized deliberator-based scheme, the deliberation (i.e. decision-making) itself is confined to the deliberator. As mentioned in Section 4.1, this may lead to issues of scalability and vulnerability of the failure in the overall system in the large-scale systems. In the future, in order to overcome these problems, the participants—particularly in large-scale systems—can be grouped, each group being moderated by a local deliberator. The current framework can then be adopted by each group, with each group's deliberator reporting their local FPD-optimal predictor (4.21) to the (global) deliberator, for final deliberation. This scheme is referred as a *hybrid of centralized and decentralized scheme* in the related work i.e. [163].

8.4 Conclusion for Chapter 6

In Chapter 6, the FPD-optimal solutions of parameter prior and predictor derived in Chapter 5, are adopted for merging multiple AR nodes for the purpose of deriving $\mathcal{I}_0$'s optimal OSA predictor, through processing K (Definition 6.1.1). In particular, the deliberator

processes AR parametric observation model and OSA predictors (Section 3.3.2)— reported by the participants— in order to derive $\mathcal{I}_0$'s optimal parameter prior and also OSA optimal predictor. These FPD- optimal designs are derived for both known and unknown model order, p .

8.4.1 Possible extensions and future directions of Chapter 6

One direct extension of the work presented in Chapter 6 might be deriving the $\mathcal{I}_0$'s optimal MSA predictor through processing multiple MSA predictors provided by AR participants. There is an opportunity to use MSA predictor in Section 3.3.3 for processing local data of AR participants, Then reporting them to the deliberator for designing their optimal MSA predictor. This task will be reported in the future work. We might also implement FPD-moral deliberator (4.15) for processing multiple AR nodes after finding an effective technique for evaluation of the Kuhn-Tucker multiples, λ .

8.5 Conclusion for Chapter 7

In Chapter 7, we derive the SF solution for designing the time step of BF through processing alternative predictor of time-variant states. The designed SF is derived for a rich class of observation models, i.e.(7.53) presented in Lemma 7.5.1. Furthermore, the designed SF is used for processing data in the data step of BF via Bayes' rule. Then we have extended the SF solution to the logarithmic pooling design (4.4) for the purpose of processing multiple (alternative) predictors of time-variant states in MPS.

8.5.1 Possible extensions and future directions of Chapter 7

Recall that the proposed FPD solutions—presented in Chapters 4 and 5 cannot be used in this Chapter for the reasons mentioned in Section 7.4. The main reason is that the optimal design of the hyper-parameter prior of states i.e. $S^o(\mathbf{V}_{0,n|n-1}, \nu_{0,n|n-1}|K)$, is required. This will be possible using the second case of Remark 2.3.2, while the proposed FPD solutions are derived using case 1 Remark 2.3.2. Using the guideline provided in [19] and applying the second case of Remark 2.3.2, the FPD-optimal design of the hyper-parameter prior i.e. $S^o(\mathbf{V}_{0,n|n-1}, \nu_{0,n|n-1}|K)$ can be computed. Therefore, the FPD-optimal design of $A(\psi_n|K)$

will be derived in sequel as:

$$A^o(\psi_n | \mathbf{V}_{0,n|n-1}, \nu_{0,n|n-1}, K) = E_{S^o}[A(\psi_n | K)] \quad (8.3)$$

We wish to extend the current work to fully hierarchical FPD [19] via (8.3) and use it for merging multiple state predictors in the context of BF. We note that the stochastic approximation is also required to evaluate this hierarchical FPD-optimal design, $S^o(\cdot)$. Such computational tasks will be explored in the Future work.

8.6 Conclusion

Recall that in this thesis, the main aim was the optimal distributional design through processing knowledge distributed among multiple nodes. We have evoked FPD as our primary approach in a centralized deliberator-based scheme. This thesis extends the current context in the following way:

1. The FPD-optimal design is derived for a rich class of $\mathcal{I}$'s parametric model of x (i.e. DEF), as a common quantity of interest for all nodes, in the centralized scheme. Only solution for discrete pd was already available.
2. FPD-optimal designs of parameter prior allow quantification of the deliberator's uncertainty in their model choice, i.e. $G_0(x|\boldsymbol{\theta})$ —compared to distributional pooling design.
3. This work provides us the FPD extension (5.18) of linear pooling design. This extension allows designing the weights axiomatically and fully (functionally and computationally) tractably. Recall that the design of weights is the main challenge in distributional pooling designs.
4. The proposed FPD solutions derived for a centralized scheme might be adopted to the decentralized scheme too. Recall that in the decentralized scheme (i.e. the RHS Figure 4.1), each node will be capable to parametrize its own model of x and merge it with the received pds from the neighbouring nodes in order to derive optimal design.

Therefore, each node will be able to derive the FPD-optimal design through available knowledge by adopting our proposed solution.

5. We have understood that the FPD solutions in MPS are important applications of fundamental FPD (Definition 2.3.4) where the knowledge of the system is defined and processed in different ways. In particular, these solutions are derived through implementing the parametric hierarchical FPD expressed in item 1 Remark 2.3.2. A thorough summary of different FPD solutions proposed in the thesis is presented in Table 8.1. Recall that we favour FPD-A (5.16) over FPD-S processing (5.22) and FPD-R (5.34) for the reasons mentioned in Section 5.8. A critical comparison is between the core result of Chapter 5 i.e. FPD-A (5.16) and the main result of Chapter 4 i.e. FPD-moral deliberator (4.15). As shown in Table 8.1, both FPD-moral deliberator and FPD-A are functionally tractable and derived in the knowledge-constrained set, $\mathcal{M}_K$ with $s + 1$ variational quantities. However, FPD-moral deliberator (4.15) is not computationally tractable. In particular, numerical techniques is required to evaluate its normalizing constant and more specifically designing weights. On the other hand, FPD-A (5.16) is computationally tractable for a rich class of $\mathcal{I}_0$'s parametric model with principled and tractable design of weights. Therefore, we will consider (5.16) as our primary FPD solution for the future work.

properties \ FPD sol.	FPD-moral del.	FPD-A	FPD-R	FPD-S
Knowledge, K	K in Def. 4.3.1	K in Def. 5.1.1	K in Def. 5.1.1	K in Def. 5.1.1
Var. quant. in $\mathcal{M}_K$	$s + 1$	$s + 1$	2	1
Funct. tractable	Yes	Yes	Yes	Yes
Comput. tractable	No	Yes	No	Yes
Design of weights	num. eval.	opt. and tract.	num. eval.	unit weights

Table 8.1. A summary of proposed FPD solutions in the thesis with their properties

However, the following task remain for the future work:

- In this thesis, all $\mathcal{I}_0$'s FPD solutions are the application of parametric hierarchical FPD (i.e. case 1 Remark 2.3.2). This work has involved parametric observation model, $G_0(x|\boldsymbol{\theta})$, and optimization of parameter prior (i.e. $A^\circ(\cdot)$) that allows us the uncertainty about the parametric observation model. Notice that there is no

uncertainty on the optimized predictor. A important extension of the current work will be implementing the second setting of Remark 2.3.2 where parametric prior, A , is parametrized with finite-dimensional hyperparameter. Then the task is to derive the FPD-optimal design of hyperparameter prior. This setting will allow us quantify the uncertainty about the optimal design of $A(\boldsymbol{\theta}|\boldsymbol{\eta}, K)$. Further progress will be relaxing the constraint parametric hierarchies by allowing G_0 to be nonparametric process prior. Then we will exploit the nonparametric hierarchical FPD setting (2.18) in which $x \sim G_0(x)$, where $G_0(x)$ is unknown, nonparametric, infinite-dimensional parameter. Then we will design the nonparametric prior.

- Notice that none of the PFD solutions summarized in Table 8.1, provides the FPD-optimal design of the time step of BF. For this reason, we will use the parametric hierarchical FPD (i.e. the second setting of Remark 2.3.2) which allows us to derive the FPD-optimal design of hyperparameter prior, $S^o(V_{0,n}, \nu_{0,n}|K)$. Then using sampling methods from S^o , the FPD-optimal design of filtering pd i.e. $A^o(\psi_n|V_{0,n}, \nu_{0,n}, K)$ is achieved.
- In this thesis, all nodes provide the pd of the common x . A possible extension of this work is performing DM task for merging multiple pds of partially common x . For simplicity, let the deliberator and two participants respectively contribute $G_0(x_0|\boldsymbol{\theta})$, $G_1(x_1)$ and $G_2(x_2)$, where $x_j \neq x_l$, for $j \neq l$, $j, l = 0, 1, 2$ and $\cup_{l=0}^2 x_l = x$, then $\mathcal{I}_0$'s task will be designing the FPD-optimal predictor of x . There are some proposed methods that might be exploited. For instance, [38] proposed the merging of multiple conditional and marginal pmfs to construct the joint pmfs using Bernardo's prescription (2.36). Furthermore, [164] proposed the construction of copula-based joint pd using marginal pds of statistically dependent multi-modal random variables in multisensor fusion problem. Further related works of designing joint pds from dependent marginal pds are i.e. [161, 162] for multiple sensor processing.
- For the large-scale systems, there might be the scalability problem with the proposed centralized scheme. As a future task, the participants can be grouped, each group being moderated by a local deliberator. Then the current framework is used to merge local deliberator's predictors by a global deliberator, for final deliberation.

This work provided us with fully tractable FPD-optimal predictor of common x which was only explored for discrete pds in the state-of-the-art. As stated in the future work, this achievement can be exploited in various directions.

Appendix A—

Probability Distributions

A.1 Exponential Family

The class of standard pds which belong to exponential family (EF) has the form:

$$F(x|\boldsymbol{\theta}) = a(\boldsymbol{\theta})b(x) \exp[\langle \phi(\boldsymbol{\theta}), h(x) \rangle], \quad (\text{A.1})$$

where $a(\cdot)$ and $b(\cdot)$ are the known, scalar kernels, $\langle \cdot, \cdot \rangle$ is a known two-argument scalar operator which is distributive across addition in its second argument, and $\phi(\cdot)$, $h(\cdot)$ are the known kernels with compatible dimensions. In the following, we review some of standard pds belonging to EF (A.1) which are used through this thesis.

- **Multinomial**, $\mathcal{Mu}(1, \mathbf{p})$: Let $\boldsymbol{\theta} \equiv \mathbf{p} \equiv [p_1, p_2, \dots, p_m]^T \in \Delta_{m-1}$, where m denotes the number of states, and,

$$\Delta_{m-1} = \left\{ \mathbf{p} \mid \sum_{j=1}^m p_j = 1, 0 \leq p_j \leq 1, j = 1, \dots, m \right\}, \quad (\text{A.2})$$

is the open probability simplex of the dimension $m - 1$. The multinomial pd is a special case of EF (A.1), where $a(\boldsymbol{\theta}) \equiv a(\mathbf{p}) \equiv 1$, $\phi(\boldsymbol{\theta}) \equiv \phi(\mathbf{p}) \equiv \ln \mathbf{p}$.

- **Exponential**, $\mathcal{Exp}(\theta)$: the exponential pd is a special case of EF (A.1), where $x > 0$, $a(\boldsymbol{\theta}) \equiv \theta > 0$ and $b(x) \equiv 1$, $\phi(\boldsymbol{\theta}) \equiv \theta$, $h(x) \equiv x$, and $\langle \phi, h \rangle \equiv -\phi \cdot h$ (“ $\cdot$ ” denotes inner product).
- **Normal**, $\mathcal{N}(\mu, r)$: Let $\boldsymbol{\theta} \equiv (\mu, r)$, $\mu \in \mathbb{R}$, is the mean, and $r > 0$ is the variance. The normal pd is a special case of EF (A.1), where $x \in \mathbb{R}$, $a(\boldsymbol{\theta}) \equiv a(\mu, r) \equiv r^{-\frac{1}{2}}$ and $b(x) \equiv \frac{1}{\sqrt{2\pi}}$, $\phi(\boldsymbol{\theta}) \equiv \phi(\mu, r) \equiv -\frac{1}{2r}[-1 \quad \mu]^T \cdot [-1 \quad \mu]$, $h(x) \equiv [x \quad 1]^T \cdot [x \quad 1]$, (i.e. an

outer product of an extended data vector), and $\langle \phi, h \rangle \equiv \text{tr}(\phi \cdot h)$ (“ $\text{tr}(\cdot)$ ” denotes the trace of matrix).

A.2 Conjugate Exponential Family (CEF)

The general form of CEF is as follows:

$$\begin{aligned} F(\boldsymbol{\theta}|\mathbf{x}_n) &\equiv F(\boldsymbol{\theta}|\mathbf{V}_n, \nu_n) \\ &\propto a^{\nu_n}(\boldsymbol{\theta}) \exp[\langle \phi(\boldsymbol{\theta}), \mathbf{V}_n \rangle], \end{aligned} \quad (\text{A.3})$$

where, $(\mathbf{V}_n, \nu_n)$ together are sufficient statistics. Some of special cases of CEF are listed below.

- **Dirichlet**, $\text{Dir}(\mathbf{V}_n - \mathbf{1})$: Let $\boldsymbol{\theta} \equiv \mathbf{p} \equiv (p_1, p_2, \dots, p_m) \in \Delta_{m-1}$, where $\Delta_{m-1} = \{\mathbf{p} | \sum_{j=1}^m p_j = 1, 0 \leq p_j \leq 1, j = 1, \dots, m\}$, is the open probability simplex of the dimension $M - 1$. The Dirichlet pd is a special case of CEF (A.3), where $a(\boldsymbol{\theta}) \equiv a(\mathbf{p}) \equiv 1$, $\phi(\boldsymbol{\theta}) \equiv \phi(\mathbf{p}) \equiv \ln \mathbf{p}$, $\mathbf{V}_n = [V_{1,n}, \dots, V_{M,n}]^T$, and $\langle \phi, \mathbf{V}_n \rangle \equiv \phi^T \cdot \mathbf{V}_n$.
- **Gamma**, $\mathcal{G}a(\nu_n - 1, \mathbf{V}_n)$: the Gamma pd is a special case of CEF (A.3), where $a(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}) \equiv \theta$, and $\langle \phi, \mathbf{V}_n \rangle \equiv -\phi \cdot \mathbf{V}_n$. Notably, the Gamma pd also belongs to EF (A.1).
- **Normal inverse Gamma**, $\mathcal{NiG}(\mathbf{V}_n, \nu_n)$: the NiG pd is a special case of CEF (A.3), where $a(\boldsymbol{\theta}) \equiv a(\mu, r) \equiv r^{-\frac{1}{2}}$ and $\phi(\boldsymbol{\theta}) \equiv \phi(\mu, r) \equiv -\frac{1}{2r}[-1 \quad \mu]^T \cdot [-1 \quad \mu]$, and $\langle \phi, \mathbf{V}_n \rangle \equiv \text{tr}(\phi \cdot \mathbf{V}_n)$.

A.3 Non-Exponential Family

In the following, we list some of the non-EF pds which are used throughout this thesis.

- **Uniform**, $\mathcal{U}(a, b)$: let $\boldsymbol{\theta} \equiv \{a, b\}$, the continuous uniform pd on the support $x \in (a, b)$ is:

$$F(x|a, b) = \frac{1}{b-a}, \quad x \in (a, b), \quad a, b \in \mathbb{R} \quad (\text{A.4})$$

Notably, if $a = 0$ and $b = 1$, then the uniform pd (A.4) is called *standard* uniform pd.

- **The mixture of EF:** A widely used form of non-EF pd is the mixture of EF, i.e.

$$F(x|\boldsymbol{\theta}) = \sum_{j=1}^m \alpha_j F_j(x|\boldsymbol{\theta}), \quad (\text{A.5})$$

where, $\alpha_j \in \Delta_{m-1}$ is corresponding weight to j^{th} - component, and $F_j(x|\boldsymbol{\theta})$ is a member of EF in the form (A.1).

- **Student's- t pd, $St(\mu, \tau, \eta)$:** One of the standard non-EF pds is the Student's- t pd [165], i.e.,

$$F(x|\mu, \tau, \eta) = \frac{\Gamma(\frac{\eta+1}{2})}{\Gamma(\frac{\eta}{2})\sqrt{\pi\tau\eta}} \left\{ 1 + \frac{1}{\tau\eta}(x - \mu)^2 \right\}^{-\frac{\eta+1}{2}} \quad (\text{A.6})$$

A thorough list of standard pds is available in [13].

Appendix: Proofs

In the following, we present the proofs of lemma/expression that we left in the Chapters.

A.4 The proof of p -conditional OSA predictor (3.34)

Proof. Simplifying predictive pd as,

$$\begin{aligned}
 F(x_{n+1}|\mathbf{x}_n, p) &= \frac{F(\mathbf{x}_{n+1}|p)}{F(\mathbf{x}_n|p)} \\
 &= \frac{\kappa_{a,r}(\mathbf{V}_{n+1}^{(p)}, \nu_{n+1}^{(p)})}{\sqrt{2\pi} \cdot \kappa_{a,r}(\mathbf{V}_n^{(p)}, \nu_n^{(p)})} \\
 &= \frac{\Gamma\left(\frac{\nu_{n+1}^{(p)}}{2}\right)}{\Gamma\left(\frac{\nu_n^{(p)}}{2}\right) (\lambda_n^{(p)})^{\frac{1}{2}} \sqrt{2\pi}} \left(\frac{\lambda_{n+1}^{(p)}}{\lambda_n^{(p)}}\right)^{-\frac{1}{2}(\nu_n^{(p)}+1)} \left|\frac{\mathbf{V}_{n+1,aa}^{(p)}}{\mathbf{V}_{n,aa}^{(p)}}\right|^{-\frac{1}{2}} \quad (\text{A.7})
 \end{aligned}$$

Using (A.7), the main kernel of predictive pd is via,

$$\begin{aligned}
 \frac{\lambda_{n+1}^{(p)}}{\lambda_n^{(p)}} &= \frac{1}{\lambda_n^{(p)}} \{ \mathbf{V}_{n+1,11}^{(p)} - \mathbf{V}_{n+1,a1}^{(p)T} (\mathbf{V}_{n+1,aa}^{(p)})^{-1} \mathbf{V}_{n+1,a1}^{(p)} \} \\
 &= \frac{1}{\lambda_n^{(p)}} \{ \mathbf{V}_{n,11}^{(p)} + x_{n+1}^2 - [\mathbf{V}_{n,a1}^{(p)T} + x_{n+1} \psi_{n+1}^{(p)T}] [\mathbf{V}_{n,aa}^{(p)} + \psi_{n+1}^{(p)} \psi_{n+1}^{(p)T}]^{-1} [\mathbf{V}_{n,a1}^{(p)} + x_{n+1} \psi_{n+1}^{(p)}] \} \quad (\text{A.8})
 \end{aligned}$$

Notice that the inversion lemma states,

$$[\mathbf{V}_{n,aa}^{(p)} + \psi_{n+1}^{(p)} \psi_{n+1}^{(p)T}]^{-1} = (\mathbf{V}_{n,aa}^{(p)})^{-1} - \frac{(\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)} \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1}}{1 + \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)}} \quad (\text{A.9})$$

Moreover, using (A.9), we can simplify (A.8) as follows,

$$\begin{aligned}
\frac{\lambda_{n+1}^{(p)}}{\lambda_n^{(p)}} &= \frac{1}{\lambda_n^{(p)}} \left\{ \lambda_n^{(p)} + x_{n+1}^2 [1 - \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)} + \frac{(\psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)})^2}{1 + \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)}}] \right. \\
&\quad \left. - 2x_{n+1} [V_{n,a1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)} - \frac{(V_{n,a1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)})^2}{1 + \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)}}] + \frac{(V_{n,a1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)})^2}{1 + \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)}} \right\}
\end{aligned} \tag{A.10}$$

Therefore, using (A.10), OSA predictive pd (A.7), can be attained as follows,

$$\begin{aligned}
F(x_{n+1} | \mathbf{x}_n, p) &= \frac{\Gamma\left(\frac{\nu_n^{(p)} + 1}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{\nu_n^{(p)}}{2}\right) [\lambda_n^{(p)} (1 + \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)})]^{\frac{1}{2}}} \\
&\quad \cdot \left\{ 1 + \frac{1}{\lambda_n (1 + \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} \psi_{n+1}^{(p)})} \left(x_{n+1} - \psi_{n+1}^{(p)T} (\mathbf{V}_{n,aa}^{(p)})^{-1} V_{n,a1}^{(p)} \right)^2 \right\}^{-\frac{(\nu_n^{(p)} + 1)}{2}} \\
&= St\left(\hat{\mathbf{a}}^{(p)T} \psi_{n+1}^{(p)}, \frac{\nu_n^{(p)}}{\lambda_n^{(p)} (1 + \psi_{n+1}^{(p)T} \mathbf{V}_{n,aa}^{(-1)} \psi_{n+1}^{(p)})}, \nu_n^{(p)} + 1\right)
\end{aligned} \tag{A.11}$$

□

A.5 Proof of Lemma 3.6.1

Proof. Using (3.109) and (3.110) we have,

$$\begin{aligned}
N_{I-o} - N_{d-o} &= 3(n-1) + \sum_{p=2}^{p_u-1} (n-p)(p+1) - (n-p_u) \frac{p_u(p_u+1)}{2} \\
&= \frac{p_u^3}{6} - \frac{7p_u^2}{6} + \frac{4p_u}{3} - 1 = O(p_u^3)
\end{aligned} \tag{A.12}$$

Therefore, $N_{I-o} > N_{d-o}$, for $p_u \geq 8$.

□

A.6 Proof of Lemma 3.6.2

Proof. Assume that we exploit the on-line scenario to compute all EIMs available in off-line scenario $(\mathbf{V}_n^p; p = 1, \dots, p_u)$. Using the flow in Figure 3.3, the best on-line trajectory has to be passed, at least, through time recursion from $i = 2$ to arrive at $i = n$ which requires $\frac{p_u^3}{6} - \frac{5p_u}{3} + \frac{1}{2} = O(p_u^3)$ multiplications more than using decremental way in off-line scenario. Therefore, the lemma is proved. $\square$

A.7 The proof of the time step (7.27) and data step (7.30) of KF

Proof. The proof of the time step of KF (7.27): using (7.25) and (7.26) we have,

$$\begin{aligned} F(\psi_n | \mathbf{x}_{n-1}) &= \int_{\psi_{n-1}} F(\psi_n | \psi_{n-1}) F(\psi_{n-1} | \mathbf{x}_{n-1}) d\psi_{n-1} \\ &\propto \int_{\psi_{n-1}} \exp\{-L_n\} d\psi_{n-1} \end{aligned} \quad (\text{A.13})$$

where,

$$L_n = \frac{1}{2}(\psi_n - \mathbf{A}_n \psi_{n-1})^T \mathbf{R}_\psi^{-1} (\psi_n - \mathbf{A}_n \psi_{n-1}) + \frac{1}{2}(\psi_{n-1} - \boldsymbol{\mu}_{n-1})^T \boldsymbol{\Sigma}_{n-1}^{-1} (\psi_{n-1} - \boldsymbol{\mu}_{n-1}) \quad (\text{A.14})$$

By decomposing L_n as,

$$L_n = L_n(\psi_{n-1}, \psi_n) + L_n(\psi_n) \quad (\text{A.15})$$

where, $L_n(\psi_{n-1}, \psi_n)$ is considered as a quadratic form in ψ_{n-1} and obtained by calculating the first two derivatives and minimizing it w.r.t ψ_{n-1} as,

$$\begin{aligned} L_n(\psi_{n-1}, \psi_n) &= \frac{1}{2}(\psi_{n-1} - (\mathbf{A}_n^T \mathbf{R}_\psi^{-1} \mathbf{A}_n + \boldsymbol{\Sigma}_{n-1}^{-1})^{-1} [\mathbf{A}_n^T \mathbf{R}_\psi^{-1} \psi_n + \boldsymbol{\Sigma}_{n-1}^{-1} \boldsymbol{\mu}_{n-1}])^T \\ &\quad (\mathbf{A}_n^T \mathbf{R}_\psi^{-1} \mathbf{A}_n + \boldsymbol{\Sigma}_{n-1}^{-1}) (\psi_{n-1} - (\mathbf{A}_n^T \mathbf{R}_\psi^{-1} \mathbf{A}_n + \boldsymbol{\Sigma}_{n-1}^{-1})^{-1} [\mathbf{A}_n^T \mathbf{R}_\psi^{-1} \psi_n + \boldsymbol{\Sigma}_{n-1}^{-1} \boldsymbol{\mu}_{n-1}]) \end{aligned} \quad (\text{A.16})$$

Notably (A.16) is the common quadratic form of the negative exponent of a normal distribution satisfying in decomposition (A.15), whereas there are other quadratic forms

holding (A.15). Consequently $L_n(\psi_n)$ can be obtained as follows,

$$\begin{aligned} L_n(\psi_n) &= \frac{1}{2}\psi_n^T \mathbf{R}_x^{-1} \psi_n + \frac{1}{2}\boldsymbol{\mu}_{n-1}^T \boldsymbol{\Sigma}_{n-1}^{-1} \boldsymbol{\mu}_{n-1} \\ &- \frac{1}{2}[\mathbf{A}_n^T \mathbf{R}_\psi^{-1} \psi_n + \boldsymbol{\Sigma}_{n-1}^{-1} \boldsymbol{\mu}_{n-1}]^T (\mathbf{A}_n^T \mathbf{R}_\psi^{-1} \mathbf{A}_n + \boldsymbol{\Sigma}_{n-1}^{-1})^{-1} [\mathbf{A}_n^T \mathbf{R}_\psi^{-1} \psi_n + \boldsymbol{\Sigma}_{n-1}^{-1} \boldsymbol{\mu}_{n-1}] \end{aligned} \quad (\text{A.17})$$

where (A.17) does not depend on ψ_{n-1} and we are allowed to pull $L_n(\psi_n)$ out of integration and (A.13) can be rewritten as,

$$\begin{aligned} F(\psi_n | \mathbf{x}_{n-1}) &\propto \exp(-L_n(\psi_n)) \\ &\propto \mathcal{N}_{\psi_n}(\boldsymbol{\mu}_{n|n-1}, \boldsymbol{\Sigma}_{n|n-1}) \end{aligned} \quad (\text{A.18})$$

By minimizing $L_n(\psi_n)$ w.r.t ψ_n and using the inversion lemma, shaping parameters of the time step of KF are obtained as follows,

$$\boldsymbol{\mu}_{n|n-1} = \mathbf{A}_n \boldsymbol{\mu}_{n-1} \quad (\text{A.19})$$

$$\boldsymbol{\Sigma}_{n|n-1} = \mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T \quad (\text{A.20})$$

The proof of the data step of KF (7.30): once again applying similar calculation to latter part of the time step of KF, $F(\psi_n | \mathbf{x}_n)$ can be obtained as follows,

$$\begin{aligned} F(\psi_n | \mathbf{x}_n) &\propto F(\mathbf{x}_n | \psi_n) F(\psi_n | \mathbf{x}_{n-1}) \\ &\propto \mathcal{N}_{\mathbf{x}_n}(\mathbf{C}_n \psi_n, \mathbf{R}_x) \mathcal{N}_{\psi_n}(\boldsymbol{\mu}_{n|n-1}, \boldsymbol{\Sigma}_{n|n-1}) \\ &\propto \mathcal{N}_{\psi_n}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \end{aligned} \quad (\text{A.21})$$

where,

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n|n-1} + (\boldsymbol{\Sigma}_{n|n-1} \mathbf{C}_n^T (\mathbf{C}_n \boldsymbol{\Sigma}_{n|n-1} \mathbf{C}_n^T + \mathbf{R}_x)^{-1} [\mathbf{x}_n - \mathbf{C}_n \boldsymbol{\mu}_{n|n-1}]) \quad (\text{A.22})$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_{n|n-1} - \boldsymbol{\Sigma}_{n|n-1} \mathbf{C}_n^T (\mathbf{C}_n \boldsymbol{\Sigma}_{n|n-1} \mathbf{C}_n^T + \mathbf{R}_x)^{-1} \mathbf{C}_n \boldsymbol{\Sigma}_{n|n-1} \quad (\text{A.23})$$

Replacing $\boldsymbol{\mu}_{n|n-1}$ and $\boldsymbol{\Sigma}_{n|n-1}$ with right-hand side of (A.19) and (A.20) respectively, we have,

$$\boldsymbol{\mu}_n = \mathbf{A}_n \boldsymbol{\mu}_{n-1} + \{(\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \mathbf{C}_n^T [\mathbf{C}_n (\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \mathbf{C}_n^T]$$

$$+ \mathbf{R}_x]^{-1}\}[\mathbf{x}_n - \mathbf{C}_n \mathbf{A}_n \boldsymbol{\mu}_{n-1}], \quad n = 2, \dots \quad (\text{A.24})$$

$$\begin{aligned} \boldsymbol{\Sigma}_n &= (\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \{ \mathbf{I} - \mathbf{C}_n^T [\mathbf{C}_n (\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \mathbf{C}_n^T \\ &+ \mathbf{R}_x]^{-1} \mathbf{C}_n (\mathbf{R}_\psi + \mathbf{A}_n \boldsymbol{\Sigma}_{n-1} \mathbf{A}_n^T) \}, \quad n = 2, \dots \end{aligned} \quad (\text{A.25})$$

Therefore, the proof is completed. □

Bibliography

- [1] T. H. Davenport and L. Prusak, *Information ecology: Mastering the information and knowledge environment*. Oxford University Press on Demand, 1997.
- [2] D. Stenmark, "The relationship between information and knowledge," in *Proceedings of IRIS*, vol. 24, pp. 11–14, 2001.
- [3] G. Bellinger, D. Castro, and A. Mills, "Data, information, knowledge, and wisdom," 2004. Accessed on <http://www.outsights.com/systems/dikw/dikw.htm>.
- [4] T. H. Davenport and L. Prusak, *Working knowledge: How organizations manage what they know*. Harvard Business Press, 1998.
- [5] J. M. Bernardo, "Bayesian statistics," in *Encyclopedia of Life Support Systems (EOLSS)*, Oxford, UK, 2003.
- [6] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, pp. 209–230, 1973.
- [7] J. O. Berger, *Statistical decision theory and Bayesian analysis*. Springer, 1985.
- [8] J. Bernardo and A. Smith, *Bayesian theory*. Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons, 1997.
- [9] L. Cui, W. Zhang, H. Zhai, X. Zhang, and X. Xie, "Modeling and application of data correlations among heterogeneous data sources," in *Proceedings of 2nd International Conference on Signal Processing Systems (ICSPS)*, vol. 2, pp. V2–413, IEEE, 2010.
- [10] J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, *Signal processing VI: Theories and applications*. Elsevier, 2012.
- [11] S. Chakraborty, "Some applications of Dirac's delta function in statistics for more than one random variable," *Applications and Applied Mathematics*, vol. 3, no. 1, pp. 42–54, 2008.
- [12] R. Frieden, *Probability, statistical optics, and data testing: a problem solving approach*, vol. 10. Springer, 2012.
- [13] A. Quinn, "Lecture notes in statistical signal processing (5c3 course)," 2014. Accessed on <http://www.mee.tcd.ie/~aquinn/5c3/>.
- [14] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

- [15] J. J. O. Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian methods applied to signal processing*. Springer, 2012.
- [16] M. H. DeGroot, *Optimal Statistical Decisions*, vol. 82. John Wiley & Sons, 2005.
- [17] M. Kárný, "Towards fully probabilistic control design," *Automatica*, vol. 32, no. 12, pp. 1719–1722, 1996.
- [18] M. Kárný and T. Kroupa, "Axiomatisation of fully probabilistic design," *Information Sciences*, vol. 186, no. 1, pp. 105–113, 2012.
- [19] A. Quinn, M. Kárný, and T. Guy, "Fully probabilistic design of hierarchical Bayesian models," *Information Sciences*, vol. 369, pp. 532–547, 2016.
- [20] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.
- [21] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.
- [22] S. F. Gull, "Bayesian inductive inference and maximum entropy," in *Maximum-entropy and Bayesian methods in science and engineering*, pp. 53–74, Springer, 1988.
- [23] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
- [24] J. M. Bernardo, "Expected information as expected utility," *The Annals of Statistics*, vol. 7, no. 3, pp. 686–690, 1979.
- [25] J. Kracík and M. Kárný, "Merging of data knowledge in Bayesian estimation," in *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics*, vol. 2, pp. 229–232, 2005.
- [26] A. Quinn, M. Kárný, and T. V. Guy, "Optimal design of priors constrained by external predictors," *International Journal of Approximate Reasoning*, 2017.
- [27] R. R. Yager and F. Petry, "An intelligent quality-based approach to fusing multi-source probabilistic information," *Information Fusion*, vol. 31, pp. 127–136, 2016.
- [28] V. P. Savchuk and H. F. Martz, "Bayes reliability estimation using multiple sources of prior information: binomial sampling," *IEEE Transactions on Reliability*, vol. 43, no. 1, pp. 138–144, 1994.
- [29] A. M. Aziz, "A new multiple decisions fusion rule for targets detection in multiple sensors distributed detection systems with data fusion," *Information Fusion*, vol. 18, pp. 175–186, 2014.

- [30] S. García-Gómez, J. González-Ordás, F. J. García-Algarra, R. Toribio-Sardón, A. Sedano-Frade, and F. Buisan-Garcia, "Kowlan: A multi agent system for Bayesian diagnosis in telecommunication networks," in *Joint International Conferences on Web Intelligence and Intelligent Agent Technologies, WIC/ACM*, pp. 195–198, IET, 2009.
- [31] A. Ben Brahim and M. Limam, "Robust ensemble feature selection for high dimensional data sets," in *International Conference on High Performance Computing and Simulation (HPCS), 2013*, pp. 151–157, IEEE, 2013.
- [32] R. Olfati-Saber, A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.
- [33] L. Cholvy and A. Hunter, "Merging requirements from a set of ranked agents," *Knowledge-Based Systems*, vol. 16, no. 2, pp. 113–126, 2003.
- [34] A. E. Abbas, "A Kullback-Leibler view of linear and log-linear pools," *Decision Analysis*, vol. 6, no. 1, pp. 25–37, 2009.
- [35] M. Rufo, J. Martín, and C. Pérez, "Inference on exponential families with mixture of prior distributions," *Computational Statistics & Data Analysis*, vol. 53, no. 9, pp. 3271–3280, 2009.
- [36] M. Rufo, J. Martín, C. Pérez, *et al.*, "Log-linear pool to combine prior distributions: A suggestion for a calibration-based approach," *Bayesian Analysis*, vol. 7, no. 2, pp. 411–438, 2012.
- [37] V. Sečkárová, "A note on weighted combination methods for probability estimation," in *Proceedings of 3rd Int. Workshop Scalable Decis. Making*, pp. 2013–23, 2013.
- [38] V. Sečkárová, *Cross-entropy Based Combination of Discrete Probability Distributions for Distributed Decision Making*. PhD thesis, Charles University, Prague, 2015.
- [39] T. Nagata and H. Sasaki, "A multi-agent approach to power system restoration," *IEEE Transactions on Power Systems*, vol. 17, no. 2, pp. 457–462, 2002.
- [40] H. B. Kim and A. Wolisz, "Load balancing for centralized wireless networks," in *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 3, pp. 1875–1879, 2005.
- [41] P. A. Jarvis, S. R. Wolfe, F. Y. Enomoto, R. A. Nado, and M. Sierhuis, "A centralized multi-agent negotiation approach to collaborative air traffic resource management planning," in *IAAI*, pp. 1787–1792, 2010.
- [42] M. Khamis, W. Gomaa, H. El-Shishiny, *et al.*, "Multi-objective traffic light control system based on Bayesian probability interpretation," in *Proceedings of 15th International Conference on Intelligent Transportation Systems*, pp. 995–1000, IEEE, 2012.
- [43] J. Xu, C. Tekin, S. Zhang, and M. van der Schaar, "Distributed multi-agent online learning based on global feedback," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2225–2238, 2015.

- [44] C. Wong, W. Chan, and P. Kam, "A Student t-mixture autoregressive model with applications to heavy-tailed financial data," *Biometrika*, vol. 96, no. 3, pp. 751–760, 2009.
- [45] S. Ganapathy, S. Thomas, P. Motlicek, and H. Hermansky, "Applications of signal analysis using autoregressive models for amplitude modulation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 341–344, 2009.
- [46] P. Lesage, F. Glangeaud, and J. Mars, "Applications of autoregressive models and time–frequency analysis to the study of volcanic tremor and long-period events," *Journal of Volcanology and Geothermal Research*, vol. 114, no. 3, pp. 391–417, 2002.
- [47] M. Prokopenko, "Learning autoregressive model order for multi-step ahead predictions," in *AI conference*, pp. 43–50, World Scientific Publishing, 1995.
- [48] S. Azizi and A. Quinn, "Hierarchical fully probabilistic design for deliberator-based merging in multiple participant systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2016.
- [49] S. Azizi and A. Quinn, "Fully probabilistic design for merging knowledge of multi-agent system," in *Annual Lero Workshop, Athlone*, 2015. Best award paper.
- [50] S. Azizi and A. Quinn, "Fully probabilistic design of multi-agent system for cloud prediction," in *Lero Research Centre Launch*, 2015.
- [51] S. Azizi and A. Quinn, "Centralized decision making scheme in distributed processing systems using fully probabilistic design," in *Annual Lero Workshop*, 2016.
- [52] S. Azizi and A. Quinn, "Approximate Bayesian filtering using stabilized forgetting," in *Proceedings of 23rd European Signal Processing Conference (EUSIPCO)*, pp. 2711–2715, IEEE, 2015.
- [53] S. Azizi and A. Quinn, "A data-driven forgetting factor for stabilized forgetting in approximate Bayesian filtering," in *Proceedings of 26th IEEE Irish Signals and Systems Conference (ISSC)*, pp. 1–6, 2015.
- [54] A. Ertas and J. C. Jones, *The engineering design process*. Wiley New York, 1996.
- [55] M. Kárný, J. Böhm, T. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesar, *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, 2006.
- [56] L. Savage, *Foundations of Statistics*. Wiley, NY, 1954.
- [57] J. Q. Smith, *Bayesian decision analysis: principles and practice*. Cambridge University Press, 2010.
- [58] D. M. Kilgour and Y. Gerchak, "Elicitation of probabilities using competitive scoring rules," *Decision Analysis*, vol. 1, no. 2, pp. 108–113, 2004.
- [59] J. E. Bickel, "Some comparisons among quadratic, spherical, and logarithmic scoring rules," *Decision Analysis*, vol. 4, no. 2, pp. 49–65, 2007.

- [60] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [61] V. Šmídl and A. Quinn, *The variational Bayes method in signal processing*. Springer, 2006.
- [62] M. Kárný and T. V. Guy, "Fully probabilistic control design," *Systems & Control Letters*, vol. 55, no. 4, pp. 259–265, 2006.
- [63] I. Maherin and Q. Liang, "Multistep information fusion for target detection using UWB radar sensor network," *IEEE Sensors Journal*, vol. 15, no. 10, pp. 5927–5937, 2015.
- [64] S. G. Hall and J. Mitchell, "Combining density forecasts," *International Journal of Forecasting*, vol. 23, no. 1, pp. 1–13, 2007.
- [65] J. Homolová and M. Kárný, "Evaluation of Kullback-Leibler divergence," Tech. Rep. 2349, 2015.
- [66] D. F. Kerridge, "Inaccuracy and inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 184–194, 1961.
- [67] C. E. Shannon, "A mathematical theory of communication," *ACM Sigmoblie Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.
- [68] M. Kárný, "Automated preference elicitation for decision making," in *Decision Making and Imperfection*, pp. 65–99, Springer, 2013.
- [69] J. Van Campenhout and T. Cover, "Maximum entropy and conditional probability," *IEEE Transactions on Information Theory*, vol. 27, no. 4, pp. 483–489, 1981.
- [70] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 113–147, 1979.
- [71] D. Bosq, *Nonparametric statistics for stochastic processes: estimation and prediction*, vol. 110. Springer Science & Business Media, 2012.
- [72] Z. Artstein and R. A. Vitale, "A strong law of large numbers for random compact sets," *The Annals of Probability*, pp. 879–882, 1975.
- [73] S. Brooks, "Markov chain Monte Carlo method and its application," *Journal of the Royal Statistical Society: Series D (the Statistician)*, vol. 47, no. 1, pp. 69–100, 1998.
- [74] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.
- [75] J. W. Pratt, H. Raiffa, and R. Schlaifer, *Introduction to statistical decision theory*. MIT Press, 1995.
- [76] P. Bruneau, M. Gelgon, and F. Picarougne, "A low-cost variational-Bayes technique for merging mixtures of probabilistic principal component analyzers," *Information Fusion*, vol. 14, no. 3, pp. 268–280, 2013.

- [77] D. P. Bertsekas, *Dynamic programming and optimal control 3rd edition, volume II*. Athena Scientific., 2011.
- [78] M. Kárný, "On approximate fully probabilistic design of decision making strategies," in *Proceedings of the 3rd International Workshop on Scalable Decision Making, ECML/PKDD*, 2013.
- [79] J. S. Erkelens, *Autoregressive modeling for speech coding: Estimation, interpolation and quantization*. TU Delft, Delft University of Technology, 1996.
- [80] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 587–597, 2013.
- [81] B. Wang, X. Chen, and W. Chang, "A light-weight trust-based QoS routing algorithm for ad hoc networks," vol. 13, pp. 164–180, 2014.
- [82] Z. R. Zaidi and B. L. Mark, "Mobility estimation for wireless networks based on an autoregressive model," in *Global Telecommunications Conference, GLOBECOM*, vol. 6, pp. 3405–3409, IEEE, 2004.
- [83] J. W. Lichstein, T. R. Simons, S. A. Shriner, and K. E. Franzreb, "Spatial autocorrelation and autoregressive models in ecology," *Ecological monographs*, vol. 72, no. 3, pp. 445–463, 2002.
- [84] B. E. Hansen, "Threshold autoregression in economics," *Statistics and its Interface*, vol. 4, no. 2, pp. 123–127, 2011.
- [85] G. Hendranto, "Indrabayu," "a multichannel autoregressive model of rain attenuation on multiple radio links and its application in assessment of fade mitigation schemes in fixed wireless systems above 10 GHz," presented at the 28th general assembly int," *Union Radio Sci, New Delhi, India*, 2005.
- [86] W. Penny and S. Roberts, "Bayesian multivariate autoregressive models with structured priors," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 149, no. 1, pp. 33–41, 2002.
- [87] G. Rosen, "Comparison of autoregressive measures for DNA sequence similarity," in *IEEE International Workshop on Genomic Signal Processing and Statistics*, pp. 1–4, 2007.
- [88] M. Nascimento, T. Sáfiadi, F. F. e Silva, and A. C. C. Nascimento, "Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach," *Bioinformatics*, vol. 28, no. 15, pp. 2004–2007, 2012.
- [89] H. Cheng, P.-N. Tan, J. Gao, and J. Scripps, "Multistep-ahead time series prediction," pp. 765–774, Springer, 2006.
- [90] J. Šindelár, "Construction of multi-step ahead predictions in a normal BVAR (p) model using Monte Carlo sampling," in *Proceedings of the 10th International PhD workshop on Systems and Control*.

- [91] M. Kárný, "Approximate Bayesian recursive estimation," *Information Sciences*, vol. 285, pp. 100–111, 2014.
- [92] H. Leahu *et al.*, "On the Bernstein-von Mises phenomenon in the Gaussian white noise model," *Electronic Journal of Statistics*, vol. 5, pp. 373–404, 2011.
- [93] P. Bickel, B. Kleijn, *et al.*, "The semiparametric Bernstein-von Mises theorem," *The Annals of Statistics*, vol. 40, no. 1, pp. 206–237, 2012.
- [94] S. Kay, "Recursive maximum likelihood estimation of autoregressive processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 1, pp. 56–65, 1983.
- [95] M. L. Vis and L. L. Scharf, "A note on recursive maximum likelihood for autoregressive modeling," *IEEE transactions on signal processing*, vol. 42, no. 10, pp. 2881–2883, 1994.
- [96] L. J. Fitzgibbon, D. L. Dowe, and F. Vahid, "Minimum message length autoregressive model order selection," in *Proceedings of International Conference on Intelligent Sensing and Information Processing*, pp. 439–444, IEEE, 2004.
- [97] P. J. Brockwell, R. A. Davis, and A. A. Trindade, "Asymptotic properties of some subset vector autoregressive process estimators," *Journal of multivariate analysis*, vol. 90, no. 2, pp. 327–347, 2004.
- [98] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*. Springer Science & Business Media, 2006.
- [99] E. B. Andersen, "Asymptotic properties of conditional maximum-likelihood estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 283–301, 1970.
- [100] A. D. Lanterman, "Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection," *International Statistical Review*, pp. 185–212, 2001.
- [101] H. Akaike, "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting," *Biometrika*, vol. 66, no. 2, pp. 237–242, 1979.
- [102] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [103] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 190–195, 1979.
- [104] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of statistics*, pp. 416–431, 1983.
- [105] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Transactions on Information theory*, vol. 30, no. 4, pp. 629–636, 1984.
- [106] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 240–265, 1987.

- [107] D. L. Dowe, *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence: Papers from the Ray Solomonoff 85th Memorial Conference, Melbourne, Australia, Nov. 2011*, vol. 7070. Springer, 2013.
- [108] V. Smidl and A. Quinn, "Mixture-based extension of the AR model and its recursive Bayesian identification," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3530–3542, 2005.
- [109] M. Zhang, D. Ye, Q. Bai, D. Sutanto, and K. Muttaqi, "A hybrid multi-agent framework for load management in power grid systems," in *Advances in Practical Multi-Agent Systems*, pp. 129–143, Springer, 2010.
- [110] P. Zand, A. Dilo, and P. Havinga, "D-msr: A distributed network management scheme for real-time monitoring and process control applications in wireless industrial automation," *Sensors*, vol. 13, no. 7, pp. 8239–8284, 2013.
- [111] M. Kárný and R. Herzallah, "Scalable harmonization of complex networks with local controllers," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. PP, no. 99, pp. 1–11, 2015.
- [112] R. L. Winkler, "The consensus of subjective probability distributions," *Management Science*, vol. 15, no. 2, pp. B–61, 1968.
- [113] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.
- [114] D. Lindley, "Reconciliation of probability distributions," *Operations Research*, vol. 31, no. 5, pp. 866–880, 1983.
- [115] R. Olfati-Saber, E. Franco, E. Frazzoli, and J. S. Shamma, "Belief consensus and distributed hypothesis testing in sensor networks," in *Networked Embedded Sensing and Control*, pp. 169–182, Springer, 2006.
- [116] J. A. Benediktsson and P. H. Swain, "Consensus theoretic classification methods," *IEEE Transaction on System, Man, Cybernetics A*, vol. 22, no. 4, pp. 688–704, 1992.
- [117] 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.
- [118] K. Dedecius and V. Sečkárová, "Dynamic diffusion estimation in exponential family models," *IEEE Signal Processing Letters*, vol. 20, no. 11, pp. 1114–1117, 2013.
- [119] E. Camponogara, D. Jia, B. H. Krogh, and S. Talukdar, "Distributed model predictive control," *IEEE Control Systems*, vol. 22, no. 1, pp. 44–52, 2002.
- [120] J. Maestre, D. Muñoz De La Peña, and E. Camacho, "Distributed model predictive control based on a cooperative game," *Optimal Control Applications and Methods*, vol. 32, no. 2, pp. 153–176, 2011.
- [121] R. T. Clemen, "Combining overlapping information," *Management Science*, vol. 33, no. 3, pp. 373–380, 1987.

- [122] I. Albert, S. Donnet, C. Guihenneuc-Jouyaux, S. Low-Choy, K. Mengersen, J. Rousseau, *et al.*, "Combining expert opinions in prior elicitation," *Bayesian Analysis*, vol. 7, no. 3, pp. 503–532, 2012.
- [123] R. T. Clemen and T. Reilly, "Correlations and copulas for decision and risk analysis," *Management Science*, vol. 45, no. 2, pp. 208–224, 1999.
- [124] S. Kay, Q. Ding, and D. Emge, "Joint pdf construction for sensor fusion and distributed detection," in *13th IEEE Conference on Information Fusion*, pp. 1–6, 2010.
- [125] S. Kay, Q. Ding, B. Tang, and H. He, "Probability density function estimation using the eef with application to subset/feature selection," *IEEE Transactions on Signal Processing*, vol. 64, no. 3, pp. 641–651, 2016.
- [126] J. M. Bates and C. W. Granger, "The combination of forecasts," *Journal of the Operational Research Society*, vol. 20, no. 4, pp. 451–468, 1969.
- [127] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow, *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.
- [128] R. A. Jacobs, "Methods for combining experts' probability assessments," *Neural computation*, vol. 7, no. 5, pp. 867–888, 1995.
- [129] A. Carvalho and K. Larson, "A consensual linear opinion pool," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 2518–2524, 2012.
- [130] V. Zarnowitz, "The new ASA-NBER survey of forecasts by economic statisticians," *The American Statistician*, vol. 23, no. 1, pp. 12–16, 1969.
- [131] M. H. DeGroot and J. Mortera, "Optimal linear opinion pools," *Management Science*, vol. 37, no. 5, pp. 546–558, 1991.
- [132] D. Bolger and B. Houlding, "Deriving the probability of a linear opinion pooling method being superior to a set of alternatives," *Reliability Engineering & System Safety*, vol. 158, pp. 41–49, 2017.
- [133] C. Genest and J. V. Zidek, "Combining probability distributions: A critique and an annotated bibliography," *Statistical Science*, pp. 114–135, 1986.
- [134] C. Genest and K. J. McConway, "Allocating the weights in the linear opinion pool," *Journal of Forecasting*, vol. 9, no. 1, pp. 53–73, 1990.
- [135] S. Wu, Y. Bi, X. Zeng, and L. Han, "Assigning appropriate weights for the linear combination data fusion method in information retrieval," *Information Processing & Management*, vol. 45, no. 4, pp. 413–426, 2009.
- [136] T. Heskes, "Selecting weighting factors in logarithmic opinion pools," *Advances in neural information processing systems*, pp. 266–272, 1998.

- [137] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [138] J. Geweke and G. Amisano, "Optimal prediction pools," *Journal of Econometrics*, vol. 164, no. 1, pp. 130–141, 2011.
- [139] A. Quinn and M. Kárný, "Learning for nonstationary Dirichlet process," *International Journal of Adaptive Control and Signal Processing*, vol. 21, no. 10, pp. 827–855, 2007.
- [140] A. Dogandzic and B. Zhang, "Distributed estimation and detection for sensor networks using hidden Markov random field models," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 3200–3215, 2006.
- [141] S. Frühwirth-Schnatter, *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.
- [142] F. E. Daum, "Nonlinear filters: beyond the Kalman filter," in *Aerospace and Electronic Systems Magazine, IEEE*, pp. 57–69, 2005.
- [143] V. Šmídl and F. Gustafsson, "Bayesian estimation of forgetting factor in adaptive filtering and change detection," in *IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 197–200, IEEE, 2012.
- [144] S. Sarkka and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *IEEE Transactions on Automatic Control*, vol. 54, no. 3, pp. 596–600, 2009.
- [145] S. Thrun, "Probabilistic algorithms in robotics," *Ai Magazine*, vol. 21, no. 4, p. 93, 2000.
- [146] T. Jiang, N. D. Sidiropoulos, and G. B. Giannakis, "Kalman filtering for power estimation in mobile communications," *IEEE Transactions on Wireless Communications*, vol. 2, no. 1, pp. 151–161, 2003.
- [147] K. Ramachandra, *Kalman filtering techniques for radar tracking*. CRC Press, 2000.
- [148] F. Daum, "Nonlinear filters: beyond the Kalman filter," *IEEE Aerospace and Electronic Systems Magazine*, vol. 20, no. 8, pp. 57–69, 2005.
- [149] G. Pappas and M. Zohdy, "Extended Kalman filtering and pathloss modeling for shadow power parameter estimation in mobile wireless communications," *Int. J. Smart Sens. Intell. Syst*, vol. 7, pp. 898–924, 2014.
- [150] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pp. 153–158, IEEE, 2000.
- [151] V. Smidl and A. Quinn, "Variational Bayesian filtering," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 5020–5030, 2008.

- [152] S. Saha, E. Özkan, F. Gustafsson, and V. Šmídl, "Marginalized particle filters for Bayesian estimation of Gaussian noise parameters," in *13th Conference on Information Fusion*, pp. 1–8, IEEE, 2010.
- [153] C. Nemeth, P. Fearnhead, and L. Mihaylova, "Sequential Monte Carlo methods for state and parameter estimation in abruptly changing environments," *IEEE Transactions on Signal Processing*, vol. 62, no. 5, pp. 1245–1255.
- [154] D. Xu, C. Shen, and F. Shen, "A robust particle filtering algorithm with non-Gaussian measurement noise using Student-t distribution," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 30–34, 2014.
- [155] R. Kulhavý and M. B. Zarrop, "On a general concept of forgetting," *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.
- [156] R. Kulhavý and F. Kraus, "On duality of regularized exponential and linear forgetting," *Automatica*, vol. 32, no. 10, pp. 1403–1415, 1996.
- [157] V. Šmídl and A. Quinn, "Bayesian estimation of non-stationary AR model parameters via an unknown forgetting factor," in *Proc. IEEE Workshop on Dig. Sig. Process.*, (Taos (New Mexico)), pp. 221–225, 2004.
- [158] S. S. Rao and S. S. Rao, *Engineering optimization: theory and practice*. John Wiley & Sons, 2009.
- [159] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.
- [160] L. M. de Carvalho, D. A. Villela, F. C. Coelho, and L. S. Bastos, "Choosing the weights for the logarithmic pooling of probability distributions," *arXiv preprint arXiv:1502.04206*, 2015.
- [161] S. Kay, Q. Ding, and M. Rangaswamy, "Sensor integration for classification," in *IEEE Conference on Signals, Systems and Computers*, pp. 1658–1661, 2010.
- [162] S. Kay, Q. Ding, and M. Rangaswamy, "Sensor integration by joint PDF construction using the exponential family," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 49, no. 1, pp. 580–593, 2013.
- [163] H. Kennedy, "Fusion of possibly biased location estimates using Gaussian mixture models," *Information Fusion*, vol. 13, no. 3, pp. 214–222, 2012.
- [164] S. G. Iyengar, P. K. Varshney, and T. Damarla, "A parametric copula based framework for multimodal signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1893–1896, 2009.
- [165] S. Kotz and S. Nadarajah, *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.