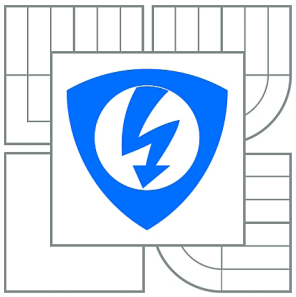
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

EXPLORING BRAIN NETWORK CONNECTIVITY THROUGH HEMODYNAMIC MODELING

ZKOUMÁNÍ KONEKTIVITY MOZKOVÝCH SÍTÍ POMOCÍ HEMODYNAMICKÉHO MODELOVÁNÍ

DIZERTAČNÍ PRÁCE

DOCTORAL THESIS

AUTOR PRÁCE — Ing. MARTIN HAVLÍČEK

AUTHOR

ŠKOLITEL — Prof. Ing. JIŘÍ JAN, CSc.

ACADEMIC ADVISOR

KONZULTANTI — Prof. M.S. M.A. VINCE D. CALHOUN, Ph.D.

CONSULTANTS — Prof. MUDr. MILAN BRÁZDIL, Ph.D.

BRNO 2011

Abstrakt

Zobrazení funkční magnetickou rezonancí (fMRI) využívající "blood-oxygen-level-dependent" efekt jako indikátor lokální aktivity je velmi užitečnou technikou k identifikaci oblastí mozku, které jsou aktivní během percepce, kognice, akce, ale také během klidového stavu. V poslední době také roste zájem o studium konektivity mezi těmito oblastmi, zejména v klidovém stavu.

Tato práce předkládá nový a originální přístup k problému nepřímého vztahu mezi měřenou hemodynamickou odezvou a její příčinou, tj. neuronálním signálem. Zmíněný nepřímý vztah komplikuje odhad efektivní konektivity (kauzálního ovlivnění) mezi různými oblastmi mozku z dat fMRI. Novost prezentovaného přístupu spočívá v použití (zobecněné nelineární) techniky slepé dekonvoluce, což dovoluje odhad endogenních neuronálních signálů (tj. vstupů systému) z naměřených hemodynamických odezev (tj. výstupů systému). To znamená, že metoda umožňuje "data-driven" hodnocení efektivní konektivity na neuronální úrovni i v případě, že jsou měřeny pouze zašumělé hemodynamické odezvy. Řešení tohoto obtížného dekonvolučního (inverzního) problému je dosaženo za použití techniky nelineárního rekurzivního Bayesovského odhadu, který poskytuje společný odhad neznámých stavů a parametrů modelu.

Práce je rozdělena do tří hlavních částí. První část navrhuje metodu k řešení výše uvedeného problému. Metoda využívá odmocninové formy nelineárního kubaturního Kalmanova filtru a kubaturního Rauch-Tung-Striebelova vyhlazovače, ovšem rozšířených pro účely řešení tzv. problému společného odhadu, který je definován jako simultánní odhad stavů a parametrů sekvenčním přístupem. Metoda je navržena především pro spojitě-diskrétní systémy a dosahuje přesného a stabilního řešení diskretizace modelu kombinací nelineárního (kubaturního) filtru s metodou lokální linearizace. Tato inverzní metoda je navíc doplněna adaptivním odhadem statistiky šumu měření a šumů procesu (tj. šumů neznámých stavů a parametrů). První část práce je zaměřena na inverzi modelu pouze jednoho časového průběhu; tj. na odhad neuronální aktivity z fMRI signálu.

Druhá část generalizuje navrhovaný přístup a aplikuje jej na více časových průběhů za účelem umožnění odhadu parametrů propojení neuronálního modelu interakce; tj. odhadu efektivní konektivity. Tato metoda představuje inovační stochastické pojetí dynamického kauzálního modelování, což ji činí odlišnou od dříve představených přístupů. Druhá část se rovněž zabývá metodami Bayesovského výběru modelu a navrhuje techniku pro detekci irelevantních parametrů propojení za účelem dosažení zlepšeného odhadu parametrů.

Konečně třetí část se věnuje ověření navrhovaného přístupu s využitím jak simulovaných tak empirických fMRI dat, a je významných důkazem o velmi uspokojivých výsledcích navrhovaného přístupu.

Klíčová slova: Efektivní konektivita, fMRI, neuronální, hemodynamické modelování, nelineární kubaturní Kalmanův filtr, vyhlazovač, odhad parametrů, Bayesovský výběr modelu, spojitě-diskrétní systémy, adaptivní filtrace, variační Bayes.

Abstract

Functional magnetic resonance imaging (fMRI) utilizing the blood-oxygen-level-dependent (BOLD) effect as an indicator of local activity is a very useful technique to identify brain regions that are active during perception, cognition, action, and also during rest. Currently, there is a growing interest to study connectivity between different brain regions, particularly in the resting-state.

This thesis introduces a new and original approach to problem of indirect relationship between observed hemodynamic response and its cause represented by neuronal signal, as this indirect relationship complicates the estimation of effective connectivity (causal influence) between different brain regions from fMRI data. The novelty of this approach is in (generalized nonlinear) blind-deconvolution technique that allows estimation of the endogenous neuronal signals (system inputs) from measured hemodynamic responses (system outputs). Thus, it enables a fully data-driven evaluation of effective connectivity on neuronal level, even though only fMRI hemodynamic responses are observed. The solution to this difficult deconvolution (model inversion) problem is obtained through a nonlinear recursive Bayesian estimation framework for joint estimation of hidden model states and parameters.

This thesis is divided into three main parts. The first part proposes a method to solve the above mentioned inversion problem. The method uses a square-root form of a nonlinear cubature Kalman filtering and cubature Rauch-Tung-Striebel smoothing extended to a joint estimation problem defined as a simultaneous estimation of states and parameters in a sequential manner. The method is designed particularly for continuous-discrete systems and obtains an accurate and stable solution to model discretization by combining nonlinear (cubature) filtering with local linearization. Moreover, the inversion method is equipped with the adaptive estimation of measurement, state, and parameter noise statistics. The first part of the thesis is focused only on the single time course model inversion; i.e. estimation of neuronal signal from fMRI signal.

The second part generalizes the proposed approach and applies it to multiple fMRI time courses in order to enable the estimation of coupling parameters of a neuronal interaction model; i.e. estimation of effective connectivity. This method represents a novel stochastic treatment of dynamic causal modeling, which makes it distinct from any previously introduced approach. The second part also deals with methods for Bayesian model selection and proposes a technique for detection of irrelevant connectivity parameters to achieve improved performance of parameter estimation.

Finally, the third part provides a validation of the proposed approach by using both simulated and empirical fMRI data, and demonstrates robust and very good performance.

Keywords: Effective connectivity, fMRI, neuronal, hemodynamic modeling, nonlinear cubature Kalman filter, smoother, sequential parameter estimation, Bayesian model selection, pruning, continuous-discrete systems, adaptive filtering, variational Bayes.

Bibliografická citace

HAVLÍČEK, M. *Zkoumání konektivity mozkových sítí pomocí hemodynamického modelování.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011. 172 s. Vedoucí dizertační práce Prof. Ing. Jiří Jan, CSc.

Declaration

This thesis titled "*Exploring Brain Network Connectivity through Hemodynamic Modeling*" is the result of my own original work, where it draws on the work of others. This is always acknowledged at the appropriate points in the text.

Prohlášení

Prohlašuji, že dizertační práce na téma „Zkoumání konektivity mozkových sítí pomocí hemodynamického modelování“ je výsledkem mé vlastní originální práce, za použití odborné literatury a dalších informačních zdrojů. Tyto jsou vždy uvedeny na patřičných místech v textu a v seznamu literatury.

V Albuquerque dne …… .....……

Ing. Martin Havlíček

Acknowledgements

This thesis has arisen with a contribution or a help of many people to whom I would like to express my special gratitude.

First of all, I would like to thank to my advisor, Jiří Jan from Brno University of Technology and Milan Brázdil from St. Anne Hospital for giving me a chance to work on this project and supporting any of my ideas and plans. Then I would like to thank to Michal Mikl and Radek Mareček from St. Anne Hospital for sharing their enthusiasm for fMRI, especially during the first year of my Ph.D. program.

Second, I would like to express my special gratitude to Vince Calhoun, who enabled my fruitful stay in Albuquerque at the University of New Mexico and the Mind Research Network Institute and became my mentor during last three years. It was great to be here.

Most of the presented work would not be possible to achieve without important help of Karl Friston from University College London, who was very kind and patient to explain and discuss huge number of questions that I had, while trying to understand the dynamic causal modeling and its estimation. Thank you Karl. You totally rock!

I am also grateful for personal discussions that I had with Steve Smith from Oxford University, Alard Roebroeck from Maastricht University and Pedro Valdés-Sosa from Cuban Neuroscience Center. Their great experience helped me to shape my opinions about the analysis of effective connectivity.

Many more people helped me directly or indirectly, knowingly or unknowingly. I would like to thank to all my colleagues from MIAlab, especially to Elena Allen, Eric Erhardt, Eswar Damaraju, Sergey Plis, Siddharth Khullar, Lei Wu and Mohammad Arbabsrihani. But also to my colleagues and staff from the Dept. of Biomedical engineering in Brno, especially to Vratislav Harabiš, Michal Bartoš, Jiří Gazárek and Martin Lamoš. Thanks for being great friends both at and outside of work.

Finally, I would like to thank to my family, friends and especially to Lei, for their love and support.

Thank you all!

Martin Havlíček

Albuquerque, October 2011

List of publications

Peer-reviewed journal articles:

- HAVLICEK, M., FRISTON, K.J., JAN, J., BRAZDIL, M., CALHOUN, V.D. *Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering*. NeuroImage, 2011, vol. 56. no. 4, pp. 2109-2128.
- HAVLICEK, M., JAN, J., BRAZDIL, M., CALHOUN, V.D. *Dynamic Granger causality based on Kalman filter for evaluation of functional network connectivity in fMRI data*. NeuroImage, 2010, vol. 53, no. 1, pp. 65-77.

Reviewed conference articles and abstracts:

- HAVLICEK, M., FRISTON, K.J, JAN, J., BRAZDIL, M., CALHOUN, V.D. *Effective connectivity in fMRI resting state data via blind deconvolution*. In 17th Annual Meeting of the Organization for Human Brain Mapping. Quebec: Organization for Human Brain Mapping, 2011.
- HAVLICEK, M., JAN, J., BRAZDIL, M., CALHOUN, V.D. *Estimation of neuronal signal from hemodynamic responses*. In Proc. of 33st IEEE EMBS conf. 2011. Boston: IEEE EMBS, 2011, pp. 8122-8125.
- HAVLICEK, M., JAN, J., BRAZDIL, M., CALHOUN, V.D. *Deconvolution of neuronal signal from hemodynamic response*. In Proceedings of 2011 International Conference on Acoustics, Speech, and Signal Processing. Prague: IEEE ICASSP, 2011, pp. 617-620.
- HAVLICEK, M., JAN, J., BRAZDIL, M., CALHOUN, V.D. *Nonlinear estimation of BOLD signal based on cubature particle filter*. In Proceedings of Biosignal 2010: Analysis of Biomedical Signals and Images. Brno, Czech Republic: Brno University of Technology, 2010. pp. 328-332.
- HAVLICEK, M., JAN, J., PLIS, S., BRAZDIL, M., CALHOUN, V.D. *Dynamic Granger causality with embedded hemodynamic model*. In 16th Annual Meeting of the Organization for Human Brain Mapping. Barcelona: Organization for Human Brain Mapping, 2010. (Talk).

- HAVLICEK, M., JAN, J., CALHOUN, V.D., BRAZDIL, M. *Extended time-frequency Granger causality for evaluation of functional network connectivity in event-related FMRI data*. In Proc. of 31st IEEE EMBS conf. 2009. Minneapolis: IEEE EMBS, 2009. pp. 4440-4443. (Talk)
- HAVLICEK, M., JAN, J., CALHOUN, V., BRAZDIL, M., MIKL, M. *Evaluation of functional network connectivity in event-related fMRI data based on ICA*. In Proc. WC 2009. Munich: IFMBE, 2009. pp. 716-719.
- HAVLICEK, M., JAN, J., MIKL, M., BRAZDIL, M. *Comparison of analytic methods for evaluation of functional and effective brain connectivity from fMRI data*. In Proceedings of Biosignal 2008: Analysis of Biomedical Signals and Images. Brno, Czech Republic: Brno University of Technology, 2008, pp. 38-47.

Journal articles under review:

- ARBABSHIRANI, M.R., HAVLICEK, M., KIEHL, K., PEARLSON, G., CALHOUN, V.D. Functional network connectivity during rest and task conditions: A comparative study. Submitted to Human Brain Mapping, August 2011.
- DUBEAU, S., HAVLICEK, M., BEAUMONT, E., FERLAND, G., LESAGE, F., POULIOT, P. Does neurovascular deconvolution reflect the true neuronal inputs? Submitted to IEEE Transactions in Medical Imaging, August 2011.

Contents

Abstract ... **ii**

Acknowledgements ... **v**

List of publications ... **vi**

1. Introduction ... **1**

1.1 Functional integration in the brain ... 1

1.2 fMRI signal ... 4

1.3 Hemodynamic model ... 7

1.4 Nonlinear hemodynamic response modeling overview ... 11

1.5 Scope and contributions of this thesis ... 13

1.6 Thesis organization ... 16

2. Estimation of neuronal signal from fMRI data ... **18**

2.1 Introduction ... 18

2.2 Probabilistic inference ... 19

2.3 Gaussian assumed density filtering and smoothing ... 26

2.3.1 Cubature integration rules ... 29

2.3.2 Statistical linear regression ... 31

2.3.3 Cubature Kalman Filter ... 34

2.3.4 Cubature Rauch-Tung-Striebel smoother ... 37

2.3.5 Square-root representation ... 38

2.4 Sequential parameter estimation ... 42

2.4.1 Maximum *a posteriori* estimate ... 43

2.4.2 Joint estimation ... 46

2.5 Hybrid continuous-discrete state-space models ... 49

2.5.1 Local linearization scheme ... 52

2.5.2 Square-root cubature Kalman filter with local linearization ... 54

2.6 Adaptive estimation of noise statistics ... 56

2.6.1 Variational Bayesian estimation of measurement noise ... 58

2.7 Algorithm for estimation of neuronal signal ... 63

2.8 Chapter summary ... 64

3. Modeling brain network connectivity ... **67**

3.1 Introduction ... 67

3.2 Stochastic dynamic causal modeling 70
3.2.1 Neuronal interaction model 73
3.2.2 Structured priors on coupling parameters 74
3.3 Model Selection 77
3.3.1 Bayes factor 79
3.3.2 *Post hoc* Bayesian model selection 82
3.3.3 Network pruning 85
3.4 Chapter summary 89
4. Validation and application of the method 91
4.1 Introduction 91
4.2 Single time course model inversion 92
4.2.1 Hemodynamic model inversion and estimation of neuronal signal 92
4.2.2 Identifiability of hemodynamic model parameters 98
4.3 Stochastic DCM 101
4.3.1 Inversion of sDCM 102
4.3.2 Effect of noisy data 108
4.3.3 Effect of data sampling period 112
4.3.4 Effect of interregional variability of HRF 114
4.3.5 Third region influence 117
4.3.6 Larger networks 121
4.3.7 Analysis of empirical fMRI data 125
4.4 Chapter summary 130
5. Conclusions and future work 132
A. Toy examples 138
A.1 Performance of local linearization scheme 138
A.2 Performance of joint estimation scheme 141
A.3 Performance of measurement noise estimation 144
B. SCKS toolkit 147
B.1 SCKS estimation toolkit for Matlab® 147
Bibliography 148

List of figures

Figure 1.1 The functional and effective connectivity 2
Figure 1.2 Illustration of MRI data 5
Figure 1.3 The diagram of physiological process underlying BOLD signal 6
Figure 1.4 Generation and receiving process of MR signal 7
Figure 1.5 Diagram of nonlinear hemodynamic model 8
Figure 1.6 Hemodynamic response to the stimulus (Dirac impulse). 10
Figure 2.1 Schematic diagrams of probabilistic inference 20
Figure 2.2 Diagram of the recursive Bayesian filter under Gaussian assumption 22
Figure 2.3 Illustrative example of cubature rules approximation 31
Figure 2.4 Illustration of cubature points propagation during time update of CKF 35
Figure 2.5 Illustration of joint filtering scheme 47
Figure 3.1 Difference between directed acyclic and directed cyclic graph 69
Figure 3.2 Schematic illustration of stochastic DCM 72
Figure 3.3 Illustration of optimal model fit 79
Figure 3.4 Model space and adjacency matrices 82
Figure 3.5 Illustration of the pruning procedure 87
Figure 4.1 Log-normal probability density function 95
Figure 4.2 Results of single fMRI time course model inversion (part 1.) 96
Figure 4.3 Results of single fMRI time course model inversion (part 2.) 97
Figure 4.4 Results of single fMRI time course model inversion (part 3.) 97
Figure 4.5 The effect of parameter change on output hemodynamic response 99
Figure 4.6 Results of hemodynamic model identifiability analysis 101
Figure 4.7 Example of simulated connectivity network 104
Figure 4.8 Results of sDCM model inversion based on Monte Carlo simulations 105
Figure 4.9 Results of sDCM model inversion using BIC for model selection 106
Figure 4.10 Visualization of the estimation process during sDCM model inversion. . 107
Figure 4.11 Demonstration of BOLD signals with different SNR levels 108
Figure 4.12 Results of SNR experiment (part 1.) 110
Figure 4.13 Results of SNR experiment (part 2.) 111
Figure 4.14 Results of sampling period experiment 114

Figure 4.15 Demonstration of different levels of hemodynamic response variability 115
Figure 4.16 Results of interregional variability experiment 116
Figure 4.17 Two common connections involving the influence of the third region.... 118
Figure 4.18 Demonstration of missing region problem 119
Figure 4.19 Results of missing region experiment 120
Figure 4.20 Results of lager network experiment (part 1.) 122
Figure 4.21 Results of lager network experiment (part 2.) 124
Figure 4.22 Selected regions for sDCM analysis and their associated time courses... 126
Figure 4.23 Results of empirical data analysis (part 1.) 127
Figure 4.24 Results of empirical data analysis (part 2.) 128
Figure 4.25 Results of empirical data analysis (part 3.) 129
Figure A.1 Change of motion trajectory for the turn rate $\omega = 4.5°$ 139
Figure A.2 Results of air-traffic-control experiment 140
Figure A.3 Results of Lorenz attractor model inversion..... 142
Figure A.4 Estimated parameter of Lorenz attractor model 143
Figure A.5 Simulation scenario of the Range-Only Tracking 145
Figure A.6 Results of the measurement noise variance estimation 146

List of tables

Table 1.1 Hemodynamic model parameters for 1.5 T field strength. 11
Table 3.1 Bayes factor and posterior model probability classification range. 80
Table 4.1 Selected regions for sDCM analysis.range. 125

List of algorithms

Algorithm 1. The cubature Kalman filter (CKF) - additive noise 35
Algorithm 2. The cubature Rauch-Tung-Striebel smoother - additive noise 37
Algorithm 3. The square-root cubature Kalman filter (SCKF) - additive noise 40
Algorithm 4. The square-root cubature RTS smoother - additive noise 41
Algorithm 5. Joint estimation of states and parameters by CKF. 48
Algorithm 6. Time update step for continuous-discrete SCKF. 55
Algorithm 7. CKF with adaptive estimation of measurement noise covariance. 61
Algorithm 8. Estimation of neuronal signal by CKS. 63
Algorithm 9. Network pruning 88

Abbreviations and symbols

Abbreviations

BOLD	blood oxygenation level dependent
BF	Bayes factor
BMS	Bayesian model selection
CKF	cubature Kalman filter
DAG	directed acyclic graph
DCM	dynamic causal modeling
EM	expectation maximization
FIM	Fisher information matrix
fMRI	functional magnetic resonance imaging
GCM	Granger causality modeling
HRF	hemodynamic response function
LL	local linearization
MAP	maximum *a posteriori*
ML	maximum likelihood
MMSE	minimum mean square error
ROI	region of interest
RTS	Rauch-Tung-Striebel smoother
RMSE	root-mean-square error
SCKF	square-root cubature Kalman filter
SCKS	square-root cubature Kalman smoother
SEM	structural equation modeling

Symbols

a, b, c	scalars
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	matrices
$\mathrm{diag}(\mathbf{A})$	diagonal of matrix $\mathbf{A}$
$\mathrm{diag}(a_1, \ldots, a_n)$	diagonal matrix with values $a_1, \ldots, a_n$
$\mathrm{Tria}(\mathbf{A})$	matrix triangularization
$\exp(.)$	exponential or matrix exponential
$\mathbf{A}^T$	matrix transpose

$\mathbf{A}^{-1}$	matrix inverse
$\mathbf{I}$	identity matrix
$p(\mathbf{x})$	probability density of random variable $\mathbf{x}$
$p(\mathbf{x}\|\mathbf{y})$	conditional probability density of $\mathbf{x}$ given $\mathbf{y}$
$\mathbf{x} \sim p(\mathbf{x})$	random variable $\mathbf{x}$ has the distribution $p(\mathbf{x})$
dt	differential of time variable t
$d\mathbf{x}$	differential of vector $\mathbf{x}$
$\frac{d\mathbf{f}(t)}{dt}$	derivative of function $\mathbf{f}(t)$ with respect to t
$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$	partial derivative of function $\mathbf{f}(.)$ with respect to $\mathbf{x}$
$\mathbb{R}_n$	n-dimensional space of real numbers
$\int_{\mathbb{R}_n} \mathbf{f}(\mathbf{x})\, d\mathbf{x}$	integration of function $\mathbf{f}(.)$ over n-dimensional variable $\mathbf{x}$
$\mathbb{E}[\mathbf{x}]$	expectation of $\mathbf{x}$
$\mathbb{E}[\mathbf{x}\|\mathbf{y}]$	expectation of $\mathbf{x}$ given $\mathbf{y}$
$\hat{\mathbf{x}}$	estimate of $\mathbf{x}$
$\mathbf{f}(.)$	process model function
$\mathbf{h}(.), \mathbf{g}(.)$	observation model functions
$\mathbf{x}(t)$	continuous state variable at time instance t
$\mathbf{x}_t$	discrete state variable at time step t
$\mathbf{y}_t$	discrete observations/measurement at time step t
$\mathbf{y}_{1:t}$	set containing the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_t$
$\boldsymbol{\theta}_t$	vector of model parameters at time step t

Chapter 1

Introduction

1.1 Functional integration in the brain

Functional magnetic resonance imaging (fMRI) utilizing the blood oxygenation level dependent (BOLD) effect as an indicator of local activity is a very useful technique to identify brain regions that are active during perception, cognition, action, but also during rest. The present research interest that dominates in fMRI neuroimaging community can be summarized by quoting Karl Friston[1]: *"A great deal of brain mapping is focused on functional segregation and the localization of function. Functional localization implies that a function can be localized in a cortical area, whereas segregation suggests that a cortical area is specialized with some aspects of perceptual or motor processing, and that this specialization is anatomically segregated within the cortex. The cortical infrastructure supporting a single function may involve many specialized areas whose union is mediated by the functional integration among them. In this view, functional segregation is only meaningful in the context of functional integration and vice versa."*. Since it is generally believed that human cognitive functions emerge from dynamic interactions of brain networks [2], it is not surprising that in the last decade there has been an increasing interest in identifying relationships among brain regions in order to better understand functional integration. This has lead to the formulation of connectivity analysis methods that attempt to identify associated brain regions and their interactions.

There are two distinct concepts of investigating brain network connectivity (integration) in fMRI data. First, there is a *functional connectivity,* which refers to correlated structures (or any other information theoretic measure) in the data such that

[1] [1] K. Friston, "Functional and effective connectivity: a review," *Brain Connectivity,* vol. 1, 2011.

brain areas can be grouped into interacting networks. This is usually accessed either by a pair-wise correlation (or a coherence in frequency domain) between a region of interest (ROI) and the rest of the brain [3] or by a multivariate approach such as independent component analysis (ICA) [4, 5]. Second, there is *effective connectivity*, which refers to the influence that one neural system exerts over another, either at the synaptic or population level [6]. In other words, effective connectivity moves beyond statistical dependency of functional connectivity, onto measures of directed (causal) influence. This is accessed through models of interactions, which try to explain observed dependences (functional connectivity). In addition, there is a principal difference between these two concepts regarding the questions they are able to address. Critically, effective connectivity enables to distinguish between a correlation and a causation. Just because two events correlate does not mean that one has caused the other. The Latin term for such an error is: "*Non cause pro causa*", which literally means: "Not a cause for the cause". This is important, because some correlations seen in fMRI data might be meaningless, which complicates an interpretation of the results and can lead researchers to their wrong conclusions.

Evaluation of effective connectivity often requires the definition of a structural model, i.e. an assembly of brain regions (nodes) among which the causal influence is assessed. In this work, the main interest rests upon the effective connectivity

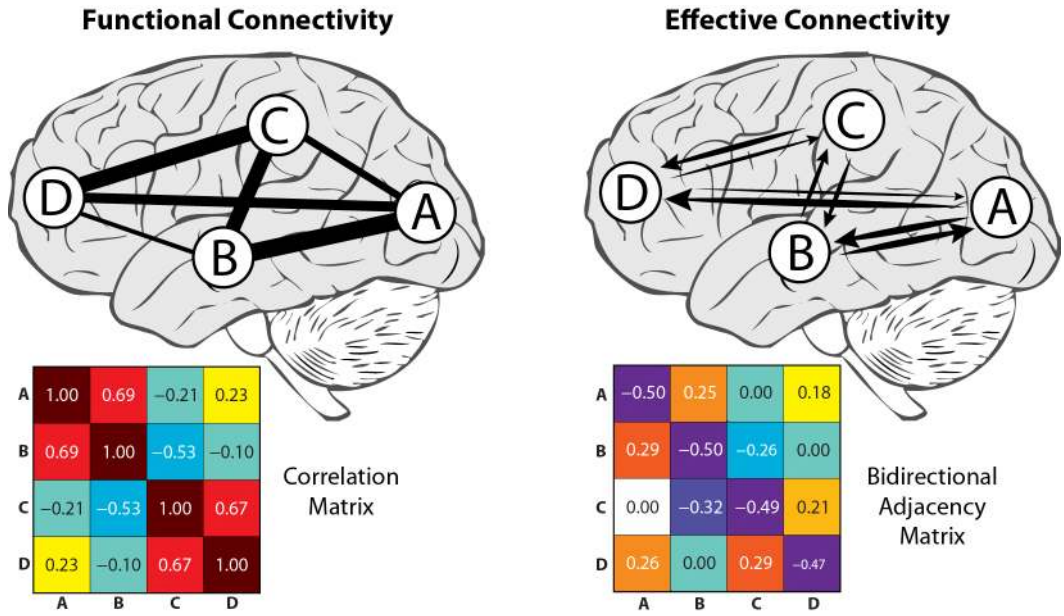


Figure 1.1 The functional and effective connectivity. The conceptual illustration of functional connectivity (left) and effective connectivity (right) with corresponding connectivity matrix representation.

framework, where an overview of the most common methods is provided in the introduction to Chapter 3.

Following the above mentioned definition of effective connectivity, it is desirable to detect causal influences among different brain regions at the neuronal (synaptic) level. This desire automatically raises an important question: Considering that the BOLD signal offers only a very indirect measure of neuronal activation, is it possible to evaluate effective connectivity at the neuronal level from fMRI data? In fMRI we measure hemodynamic responses, which reflect changes in blood flow and blood oxygenation that follow neuronal activation. Crucially, the form of this hemodynamic response can vary across subjects and different brain regions [7, 8]. These facts seriously complicate the identification of effective connectivity from fMRI [9]. However, one can reasonably justify that if it is possible to remove the effect of this hemodynamic blurring and variation, we could still achieve the aim of identifying effective connectivity from fMRI data. By saying this, we should note that there might be differences also between neuronal dynamics assigned to different brain regions having distinct cognitive functions. These differences can be more significant when processing highly cognitive tasks [10]. Nevertheless, this sort of variability is commonly ignored with respect to accuracy of modeling allowed by temporal resolution of fMRI signal.

In general, the relationship between initial neuronal activation and our fMRI observations rests on a complex biophysiological dynamic process. If this process is known and well described, it can be approximated by mathematical modeling. Considering this model (see Section 1.3), there are several ways to perform mapping from observed data to estimated neuronal signals that interact among each other, where this is partly defined by the experimental design conducting the acquisition of the data. For example, in the case of an experiment with specific task (task data), we have prior knowledge of the stimulation paradigm (i.e. any kind of stimulus presented to the subject during a scanning session), which can be used as the definition of a driving exogenous input into the model. It is then possible to model the relationship between neuronal signals and observed responses by considering a deterministic model [11], and simply infer the model parameters to fit the data. This formulation is often unsatisfactory since unexpected contributions in the "real world", which deviate from

the model, disturb the considered dynamic phenomena so that the deterministic models have a little explanatory power concerning the dynamics [12]. Therefore, it is usually preferable to consider some additive randomness to the modeled process, which then represents stochastic modeling [13, 14]. This approach is more general and is expected to have much more explanatory power than the deterministic one [15]. In a related context, there are many fMRI studies, where the data are collected when the subject is at rest (resting-state data). In this case, there is no stimulation paradigm and therefore no exogenous input that can be used for modeling. It means that the neuronal signal, which generates observed hemodynamics, has purely endogenous character. Until very recently, this fact did not allow estimation of effective connectivity in resting-state fMRI data, unless there was a good reason to assume that the neuronal dynamics and measured hemodynamics are equivalent [16]. It is specifically the form of stochastic modeling, which enables the estimation of neuronal signals and their interactions without any prior knowledge of exogenous input that allows evaluation of effective connectivity even in resting-state data [15, 17-19]. Critically, the inversion of such a stochastic model leads to a blind deconvolution[2] problem, which is described as estimating the unknown input to a dynamic system, given output data, when the model of the system contains unknown parameters [20].

1.2 fMRI signal

The fMRI signal indirectly reflects the activity of neuronal populations in the brain with excellent spatial resolution (millimeters down to hundreds of micrometers at high field strength), with a good temporal resolution (seconds down to hundreds of milliseconds) and the whole spatial coverage of the brain. Both the spatial and temporal resolution in fMRI are potentially higher than in neuroimaging methods using radioactive tracers, such as positron emission tomography (PET). In contrast to electro-encephalography (EEG) and magneto-encephalography (MEG) recordings, fMRI has a poorer temporal resolution, but it still holds its primacy in providing great spatial information about the brain function. Although fMRI is possible with few different imaging techniques, the blood oxygenation level dependent (BOLD) contrast

[2] A note on terminology is needed here: although convolution is usually defined as a linear operation, the term deconvolution is generally used in reference to the inversion of nonlinear (generalized) convolution models (i.e. restoration); we adhere to this convention.

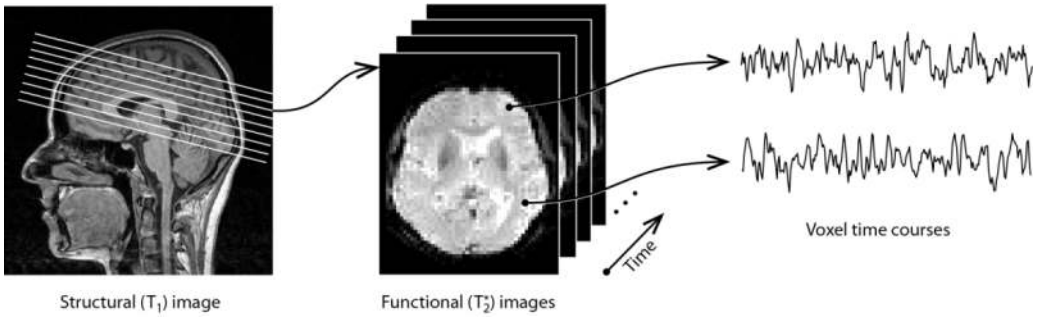


Figure 1.2 Illustration of MRI data. On the left side is the T_1-weighted anatomical image and in the middle slice of functional T_2^*-weighted image, which evolves in time. On the right side are then examples of time courses from voxels of functional data.

mechanism is employed in the great majority of cases [21, 22]. In this case, the fMRI signal is sensitive to local oxidative metabolism in the brain, which is necessary to fuel local neuronal activity [23]. The final fMRI signal is then a composite of complex chain of processes that can be classified into neuronal, physiological and physical processes [23-25].

At the neuronal level, a train of action potentials (spikes) in any single neuron arriving at the pre-synaptic terminal buttons induces neurotransmitter to release into the synaptic cleft (synaptic neuronal activity). As a consequence, excitatory and/or inhibitory electric potentials are originated in the dendrites of post-synaptic neurons due to the activation of ionic currents (I_c) that create an electrochemical disequilibrium in the cellular membrane. The neurotransmitters are rapidly terminated by a re-uptake mechanism in the astrocytes processes, while the electrochemical gradients are restored by ATP (adenosine triphosphate) transport mechanism. Hence, a metabolic and oxygen demand will appear in the neighborhood of the activated brain area [13] (see Figure 1.3).

From the perspective of physics, the MRI signal arises from the nuclei of tissue's hydrogen atoms (i.e. individual protons), which possess magnetic moment called spin. When placed to an external magnetic field (B_0), spins tend to align with it, gaining a lower energy state. The hydrogen nucleus will also experience a torque from the applied magnetic field which forces the magnetic moment to precess around the field direction. The frequency of this precession is directly determined by the strength of the applied magnetic field. At the macroscopic level, considering large (though still miscroscopic)

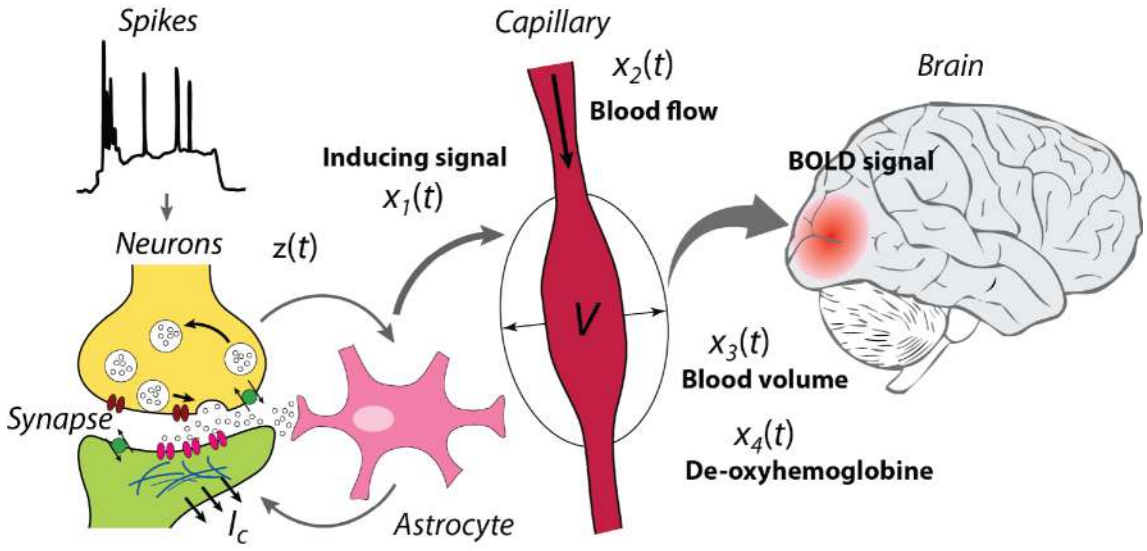


Figure 1.3 The diagram of physiological process underlying BOLD signal.

ensemble of hydrogen nuclei, there will be a net magnetic vector pointing in the direction of the applied magnetic field (stronger B_0, then more nuclei in lower energy state, results in stronger magnetic vector). If we apply a radio frequency (RF) pulse tuned to the precession frequency (known as Larmor frequency), some of nuclei are transferred to higher energy state, i.e. the mean magnetic vector is tilted and precesses around magnetic field vector. Here the nuclei spins precess in phase coherence. Then the tilted magnetic vector gradually returns to its equilibrium in a process called relaxation, emitting electromagnetic signal that generates current in receiver coils. It is the MR pulse sequence (series of RF pulses) and it parameters, which determines the magnetization preparation of the sample and the way the signal is subsequently acquired (see Figure 1.4). In fMRI, the gradient echo planar imaging (EPI) sequence is used because of its sensitivity to T_2^* relaxation time (relaxation in *xy*-plane associated with local inhomogenities in magnetic field), which allows measurement of BOLD signal [21]. In particular, neurons in the brain consume the oxygen, which is attached to hemoglobin molecules in the blood and the flow of blood continuously provides new oxygen to the neurons. This leads to the concentration increase of oxygenated blood in the capillaries surrounding the active brain area. The hemoglobin with oxygen has slightly diamagnetic properties, while without it it becomes paramagnetic, which means it distorts the applied magnetic field (by altering the local magnetic susceptibility). This local modulation of the magnetic field B_0 makes the hydrogen nuclei excited by RF pulse dephase faster, resulting in a shorter relaxation time constant T_2^*. Therefore, fMRI

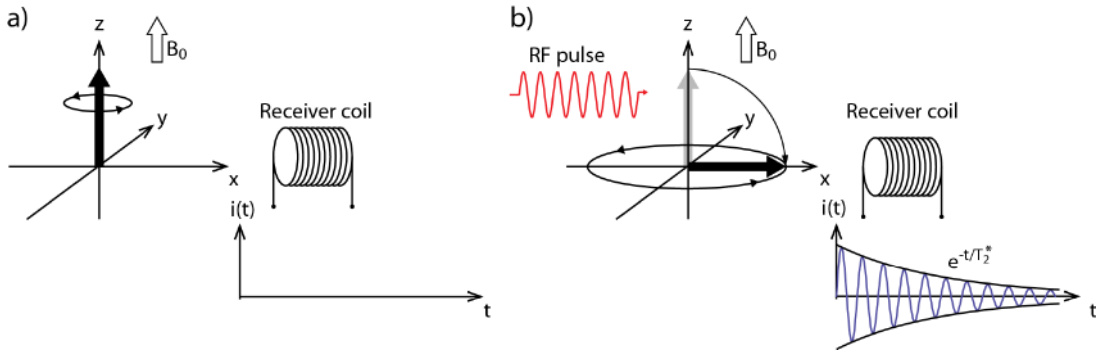


Figure 1.4 Generation and receiving process of MR signal. (a) There will be a small excess of hydrogen nuclei in the lower energy state and therefore a resultant magnetic vector will be pointing in the direction of B_0. (b) Energy can be supplied to the nuclei by applying a Radio Frequency (RF) pulse. The resultant magnetic vector is then tilted into the *xy*-plane and a current is induced in the receiver coil. Due to different relaxation processes, the *xy*-component of the magnetic vector, as well as the induced current in the receiver coil, will decay.

images will be brighter with a longer T_2^*, i.e. with a high oxygen concentration (more neuronal activation).

On the physiological level, increased neuronal activity in brain causes changes in the local cerebral flow (CBF), which increases much more than the cerebral metabolic rate of oxygen ($CMRO_2$), resulting in the decrease of local oxygen extraction fraction. With the increase of CBF, cerebral blood volume (CBV) increases as well. Because the local blood is more oxygenated, there is less deoxyhemoglobin present, the magnetic field distorts are reduced, and the local MR signal increases slightly [23].

Finally, it is known that the BOLD signal correlates more with afferent synaptic activity than with neuronal spikes [26].

1.3 Hemodynamic model

In this section, the hemodynamic model describing transformation from synaptic neuronal activity to measured BOLD signal is described.

In 1998, Buxton et al. [27] developed a mechanistically plausible model, so-called balloon model, of how evoked changes in blood flow are transformed into BOLD signal. A component of balloon model, namely the relationship between blood flow and blood volume, was then elaborated through the standard Windkessel theory [28]. This model describes the behavior of the post-capillary venous compartment by analogy to

an inflated balloon, determining a nonlinear dependence of the BOLD signal on blood volume and deoxyhemoglobin content. The complete form of the hemodynamic model was proposed by Friston [29], covering the coupling of synaptic activity and blood flow through simple linear dynamical model. The model flowchart is summarized in Figure 1.5. Here, the neural activity $u(t)$ (represented by exogenous input) causes an increase in vasodilatory signal $s(t)$ which is subject to auto-regulatory feedback. This flow-inducing signal is artificially designed to subsume many neurogenic and diffusive signal subcomponents. Blood flow $f(t)$ responds in proportion to this signal and causes changes in blood volume v(t) and deoxyhemoglobin content, $q(t)$. The dynamics of these four hemodynamic states are modeled by a set of differential equations:

$$\frac{ds(t)}{dt} = u(t) - \kappa s(t) - \chi(f(t) - 1) \tag{1.1}$$

$$\frac{df(t)}{dt} = s(t) \tag{1.2}$$

$$\frac{dv(t)}{dt} = \frac{1}{\tau}(f(t) - f_{out}(v,t)) \tag{1.3}$$

$$\frac{dq(t)}{dt} = \frac{1}{\tau}\left[f(t)E(f,t) - f_{out}(v(t))\frac{q(t)}{v(t)}\right] \tag{1.4}$$

In the first equation, τ_s and τ_f are parameters that determine dynamics of this component of the hemodynamic model. They represent the time constant for a signal decay (or elimination) and the time constant for the autoregulatory feedback from the blood flow, respectively. The later equation (1.3) says that volume changes reflect the difference between the inflow $f(t)$ to and the outflow $f_{out}(v,t)$ from venous compartment with a time constant τ. This constant expresses the mean transit time; i.e.

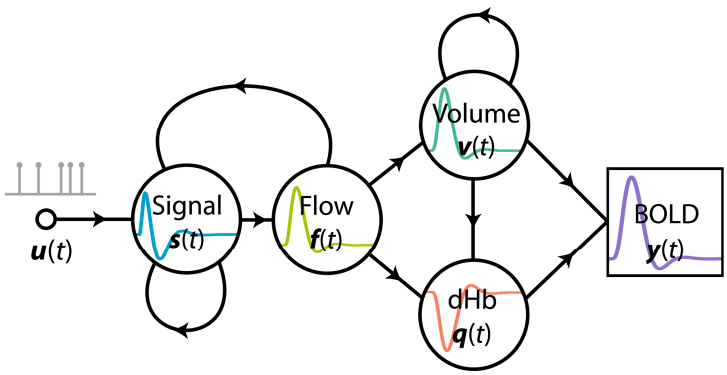


Figure 1.5 Diagram of nonlinear hemodynamic model. This diagram describes a relationship between stimulus, hemodynamic states and BOLD signal given by equations (1.1)-(1.4) and (1.7).

the average time it takes to traverse the venous compartment, or for that compartment to be replenished. The outflow is related to the blood volume through Grubb's exponent α (stiffness parameter):

$$f_{out}(v,t) = v(t)^{1/\alpha}, \tag{1.5}$$

where $1/\alpha = \gamma + \beta$. The parameters $\gamma = 2$ and $\beta > 1$ represent the laminar flow and the diminished volume reserve at high pressure, respectively. The value reported from animal studies $\alpha = 0.38 \pm 0.1$ seems to be very stable during steady-state stimulation [28]. The change in deoxyhemoglobin dq/dt reflects the delivery of deoxyhemoglobin into the venous compartment minus that expelled (outflow times concentration), where the relative oxygen extraction, E, is a function of flow and resting oxygen extraction fraction, E_0, by the capillary bed.

$$E(f,t) = \frac{1}{E_0}\left(1 - (1 - E_0)^{1/f(t)}\right) \tag{1.6}$$

The second term in (1.4) represents an important nonlinearity: the effect of flow on the state $q(t)$ is largely determined by the inflation of the balloon, resulting in the increase of $f_{out}(v,t)$ and in the clearance of deoxyhemoglobin. This effect depends upon the concentration of deoxyhemoglobin such that the clearance attained by the outflow will be severely attenuated when the concentration is low (e.g. during the peak response to prior stimulus).

Finally, the output or observed BOLD signal is expressed as a nonlinear function of blood volume and deoxyhemoglobin content:

$$y_t = V_0\left[k_1(1 - q_t) + k_2\left(1 - \frac{q_t}{v_t}\right) + k_3(1 - v_t)\right], \tag{1.7}$$

where V_0 and $k_1, \ldots, k_3$ are biophysical parameters: resting blood volume fraction and intravascular, concentration and extravascular coefficients, respectively, where the later ones are MR scanner dependent. All empirical prior values of hemodynamic model parameters (for 1.5 Tesla scanner) are summarized in Table 1.1. Although, a variety of extensions (and alternatives) have been proposed for both hemodynamic states and BOLD equations [23, 30-33]. The original version of the hemodynamic model (also one of the simplest) as described above will be considered in this work.

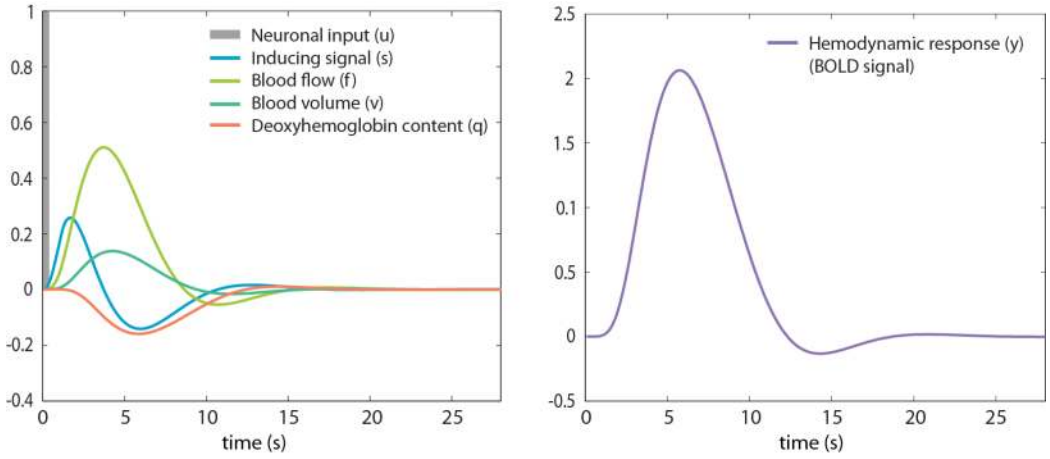


Figure 1.6 Hemodynamic response to the stimulus (Dirac impulse). The hemodynamic states on the left sides are generated by hemodynamic model and normalized to the zero baseline. On the right side is the typical (theoretical) hemodynamic response (the model output).

Notice that the model for BOLD signal (1.7) is already represented in discrete time with sampling period corresponding to the repeat time (TR), which is the time between repeated excitation of the same slice of the brain. Hereafter we use a notation of time index in subscript to denote discrete time samples. Together with physiological states $\mathbf{x}(t) = [s(t), f(t), v(t), q(t)]^T$, which evolve in continuous time, it forms a continuous-discrete model that has a state-space representation:

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t); \boldsymbol{\theta})dt + \sqrt{\mathbf{Q}}d\mathbf{w}(t) \tag{1.8}$$

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta}) + \mathbf{r}_t, \tag{1.9}$$

where function $\mathbf{f}(.)$ and $\mathbf{h}(.)$ substitute the nonlinear models for physiological states and BOLD signal, respectively, which are parameterized by the set of parameters $\boldsymbol{\theta} = \{\tau, \kappa, \chi, \varphi, \alpha, \varepsilon\}$, and the dynamics are caused by the exogenous input $\mathbf{u}(t)$. Additionally, one can also assume that the true model trajectories are disturbed by contribution of some random (stochastic) process $\mathbf{q}(t)$ and $\mathbf{r}_t$, which have here presumably simple additive form.

Table 1.1 Hemodynamic model parameters for 1.5 T field strength.

	Description	**Value**
Process equations:		
κ	Rate of signal decay	0.65 s^{-1}
χ	Rate of flow-dependent elimination	0.38 s^{-1}
τ	Hemodynamic transit time	0.98 s
α	Grubb's exponent	0.34
φ	Resting oxygen extraction fraction	0.32
Observation equation:		
V_0	Blood volume fraction	0.04
k_1	Intravascular coefficient	$4.3\vartheta_0 E_0 TE$
k_2	Concentration coefficient	$\varepsilon r_0 \varphi TE$
k_3	Extravascular coefficient	$1-\varepsilon$
ϑ_0	Frequency offset at the outer surface of the magnetized vessel	40.3 s^{-1}
E_0	Fixed resting oxygen extraction fraction	0.4
ε	Ratio of intra- and extravascular signal	1
r_0	Slope of intravascular relaxation rate	25 s^{-1}
TE	Time to echo	0.04 s

1.4 Nonlinear hemodynamic response modeling overview

In fMRI, the physiological mechanisms mediating the relationship between neuronal activation and vascular/metabolic systems have been studied extensively [34-36] and models of hemodynamic responses have been described at macroscopic level. The principal representative of these models is hemodynamic model, which was described above. This hemodynamic model is nonlinear in nature [37-39]. Therefore, to infer the hidden states and parameters of the underlying system, we require inversion methods that can handle these nonlinearities. In Friston et al. [29], the parameters of a hemodynamic model were estimated using a Volterra kernel expansion to characterize the hemodynamic response. Later, Friston et al. [40] introduced a Bayesian estimation framework to invert (i.e., fit) the hemodynamic model explicitly. This approach accommodated prior constraints on parameters and avoided the need for Volterra kernels. Subsequently, the approach was generalized to cover networks of coupled regions and to include parameters controlling the neuronal coupling (effective connectivity) among brain regions [11]. The Bayesian inversion of these models is

known as dynamic causal modeling (DCM) and is now used widely to analyze effective connectivity in fMRI and electrophysiological studies. These approaches to hemodynamic and causal models only account for noise at the level of the measurement; where this noise includes thermally generated random noise. However, there is an important contribution of physiological noise that represents stochastic fluctuations due to metabolic and vascular responses, which affect the hidden states of the system; furthermore, neuronal activity can show pronounced endogenous fluctuations [41, 42]. Motivated by this observation, Riera et al. [13] proposed a technique based on a fully stochastic model (i.e. including physiological noise) that used the local linearization filter (LLF) [43], which can be considered a form of extended Kalman filtering (EKF) [44] for continuous dynamic systems. Besides estimating hemodynamic states and parameters, this approach allows one to estimate the system's input, i.e. neuronal activity; by its parameterization via radial basis functions (RBFs). In Riera et al. [45], the number of RBFs was considered fixed *a priori*, which means that the solution has to lie inside a regularly distributed but sparse space (otherwise, the problem is underdetermined).

The hemodynamic response and hidden states of hemodynamic model possess strong nonlinear characteristics, which are prescient with respect to stimulus duration [39, 46]. This makes one wonder whether a linearization approach such as LLF can handle such strong nonlinearities. Johnston et al. [47] proposed particle filtering, a sequential Monte Carlo method, that accommodates true nonlinearities in the model. This approach was shown to be both accurate and robust, when used to estimate hidden physiologic and hemodynamic states; and was superior to LLF. Similarly, two-pass particle filtering, including a smoothing (backwards pass) procedure, was introduced by Murray et al. [48]. Another attempt to infer model parameters and hidden states used the unscented Kalman filter (UKF), which is more suitable for highly nonlinear problems [49]. Finally, Jacobson et al. [50] addressed inference on model parameters, using a Metropolis–Hastings algorithm for sampling their posterior distribution.

None of the methods mentioned above, except [13] with its restricted parameterization of the input, can perform a complete deconvolution of fMRI signals and estimate both hidden states and input; i.e. the neuronal activation, without knowing the input (stimulation function). Here, an important exception is the methodology

recently introduced by Friston et al. [14] called dynamic expectation maximization (DEM) and its generalizations: variational filtering [51] and generalized filtering [52]. DEM represents a variational Bayesian technique [53, 54], that is applied to models formulated in terms of generalized coordinates of motion.

This is the state of the art that covers the main contributions to nonlinear modeling of hemodynamic responses in fMRI data and inversion of such models. We purposely excluded our own contributions to this problem, since this is the main content of this thesis as listed below.

1.5 Scope and contributions of this thesis

In the introduction we have emphasized the general problem of estimating the effective connectivity among different brain regions from fMRI data. This problem stems from the fact that the BOLD signal is an indirect measure of the neuronal signal, and the shape of hemodynamic response function varies across different brain regions and also across subjects. In order to enable the identification of effective connectivity from fMRI data that is in the agreement with the true effective connectivity at the neuronal level, one has to solve the inverse ("deconvolution") problem. Moreover, we have also highlighted the methodological enrichment in considering stochastic representation of dynamic modeling as opposed to the limited deterministic one. Finally, we have mentioned the motivation to the inverse problem, where we do not have a prior knowledge of exogenous input, as it can be applied also to the resting-state data.

Although many attempts were already made in this direction as discussed above, there is still considerable room for improvement. Here, the main aim is to build a more accurate but less restrictive estimation approach that can be broadly applied to any fMRI data.

The successful solution to this inverse problem, i.e. successful estimation of the neuronal signal, requires the following:

1) An estimation framework that is able to handle the nonlinear characteristics of a hemodynamic model that couples neuronal activity to BOLD signal.
2) Fully stochastic modeling, since no model is completely able to catch the real world dynamics and that at any modeled physiological level, the likely

contribution of the noise has to be taken into account. Moreover, the endogenous neuronal signal can be recovered only by considering a stochastic modeling.

3) A robust approach to stochastic continuous-discrete modeling, because the causal chain of hemodynamic model is described in continuous time and, as required above, should also account for randomness.
4) An efficient framework for the estimation of model parameters in order to achieve a good fit of the model to the data and allow for diversity of hemodynamic responses across the brain. Additionally, this framework should preferably enable sequential modeling of conditional dependencies between parameters and modeled states.

These points define the topics that are addressed in this thesis (in the first half of the thesis in particular). When considering a suitable estimation framework that could possibly meet all the above requirements, the preference was to use and further develop new methods from the field of engineering, in the hope that their introduction to the society of computational neuroscience could raise the interest. Another important factor was to consider reasonable computational demands of the employed methods. A great deal of effort has been devoted to the introduction, description and motivation of using these methods. Specifically, we took an advantage and highlighted a recent development of new nonlinear cubature Kalman filter [55]. In this context, special attention is devoted to the joint state-parameter estimation problem. Another relevant part of the thesis describes an accurate discretization of continuous model based on local linearization scheme [56] and online Bayesian learning of measurement noise statistics.

In the latter part of the thesis, we generalize the inverse problem into the multivariate case, where multiple brain regions are involved and where the model of causal interactions at the neuronal level is considered. Critically, this introduces a new concept in evaluation of effective connectivity through stochastic dynamic causal modeling. This is accompanied by a description of the second level inference that is known as model selection. In particular, we discuss a Bayesian approaches to model selection based on different approximations of the marginal likelihood. Consequently, we introduce a simple algorithm for detection of irrelevant parameters in neuronal interaction model based on network pruning.

Finally, in the last part we validate the proposed method from different perspectives, and try to address questions, which presently dominate in the neuroscience community, regarding possible application of methods for analysis of effective connectivity.

This thesis proposes the following:

Contribution to computational neuroscience:

- A novel approach to the nonlinear modeling of hemodynamic signal, where the underlying neuronal signal is estimated from the measured BOLD time series. This approach performs a blind (nonlinear) deconvolution, where the model parameters, physiological states and mainly the endogenous input into the model (neuronal signal) are estimated from measured data.

- A new approach for the evaluation of effective connectivity based on stochastic dynamic causal modeling. This enables inversion of full connectivity models without knowing the driving input or having a hypothesis about the connectivity structure. This means that the *a priori* unknown model of neuronal interactions is learned from the data.

Contribution to engineering methodology:

- Formulation of combined use of cubature Kalman filtering and Rauch-Tung-Striebel smoothing in system identification for joint estimation of the hidden states, model parameters, and endogenous model input. The convergence is supported by an iterative scheme that automatically maximizes the log-likelihood.

- A new algorithm for estimation of continuous-discrete state-space models based on combination of the (square-root) cubature Kalman filter and local linearization scheme, which provides an accurate and stable discretization of a continuous model represented by stochastic differential equations.

- A new nonlinear adaptive Kalman filter for joint estimation problem, where the measurement noise covariance is effectively learned through a variational Bayesian approach, and parameter and state noise covariance are estimated by the Robbins-Monro stochastic approximation scheme.

1.6 Thesis organization

Beyond the introduction chapter that proceeded to this section, the content of this thesis is distributed into four chapter and two appendices. After each chapter a short summary is provided.

Chapter 2: Estimation of neuronal signal from fMRI data

The second chapter follows the general overview of the problem provided in the thesis introduction and provides the theoretical background for how to solve this problem. In particular, it introduces a novel approach for estimation of neuronal signals from fMRI data by utilizing a nonlinear, sequential estimation framework based on cubature Kalman filtering and smoothing. It also proposes the joint estimation of states and parameters, which is accompanied by adaptive estimation of process noise and measurement noise covariance matrices. Further, it introduces an estimation framework for continuous-discrete systems.

Chapter 3: Modeling brain network connectivity

The introduction to this chapter provides a more extensive overview of methodological advances for the evaluation of effective connectivity. This is followed by an extension of the estimation framework described in Chapter 2 to a multivariate case. In this sense, we describe a neuronal model of interaction, where the coupling parameters represent the weights of effective connectivity we wish to estimate. This chapter further addresses different approaches for Bayesian model selection and proposes improved estimation of coupling parameters based on network pruning.

Chapter 4: Validation and application of the method

The fourth chapter is focused on validation of the approaches introduced in Chapters 2 and 3. Here we demonstrate the estimation of the neuronal signal from a single fMRI time course and later also test the performance of full model inversion, which provides estimates of coupling parameters in neuronal interaction model, using multiple time courses of fMRI data.

Chapter 5: Conclusions and future work

This chapter contains concluding remarks on attained aims and results. It discusses their relevance and suggests a few directions for future research.

Appendix A: Toy examples

This appendix contains toy examples that demonstrate the performance of the proposed algorithm for continuous-discrete systems, the algorithm for sequential estimation of measurement noise covariance, and the algorithm for joint estimation of states and parameters.

Appendix B: SCKS toolkit

The last appendix gives basic information about the developed software toolkit for Matlab®, and provides a link where it can be downloaded.

Chapter 2

Estimation of neuronal signal from fMRI data

2.1 Introduction

The previous chapter provided a brief introduction to the nonlinear modeling of BOLD responses. We saw that the hemodynamic model naturally forms the state-space model, where the measured data are related to the subset of state-space variables (physiological states) by an observation equation. In this continuous-discrete time dynamic system, which represents a generative model of the BOLD signal, both state and observation equations are nonlinear and polluted by physiological and instrumental noise, respectively. In general, the estimation of the state of a continuous system from noisy discrete observations can be performed using the nonlinear filter theory, which is an extension of the original framework (Kalman filter theory) formulated to provide a sequential and computationally efficient solution to the linear filtering and prediction problems [57]. Finding the optimal nonlinear system identification method (i.e. the estimation of the model parameters and the trajectories of unobservable states) is an active research area.

It is the main goal of this chapter to introduce a new approach to this identification problem. In particular, we will first provide a short introduction to the probabilistic inference based on optimal recursive Bayesian solution. Since this solution is tractable only for linear systems, we will focus on very recent developments in nonlinear Kalman filtering based on efficient cubature integration rules [55]. This numerical tool called cubature Kalman filtering will serve as the cornerstone for further extensions and developments. Specifically, we will describe the cubature Rauch-Tung-Striebel smoother to obtain more accurate estimates of the state, including efficient square-root

implementation. Next, we propose the joint estimation framework to simultaneously infer the hidden states and model parameters. We will also introduce a new filtering approach for hybrid continuous-discrete systems based on an accurate discretization scheme of stochastic differential equations called local linearization combined with the above mentioned cubature integration rules. Consequently, in order to make the algorithm easily adaptable to the real data, we will discuss the extension to adaptive filtering through Bayesian estimation of the measurement noise covariance [55], and Robbins-Monro approximation of the parameter and state noise covariance matrices [58].

Finally, all these extensions and developments will be combined into one single algorithm, which will represent a new approach to estimation of neuronal signal from BOLD responses; i.e. the blind (nonlinear) deconvolution approach where all the hemodynamic states, the model parameters and mainly the input (neuronal signal) are estimated from observed BOLD responses.

Importantly, this chapter will deal only with a single time course modeling. The generalization to multivariate case, including inferring causal relationships, is left for the Chapter 3. From the perspective of Bayesian inference, we can also say that this chapter focuses only on the first level of inference, during which model parameters and unobserved trajectories of neuronal and physiological states are estimated. The second level of inference involving the model selection is described in Chapter 3.

2.2 Probabilistic inference

The problem of estimating the hidden states (causing data), parameters (influencing the dynamics of hidden states) and any non-controlled endogenous input to the system, in a situation when only observations are given, requires probabilistic inference. If we interpret our data through a dynamic state-space model (DSSM), then we are facing the sequential (recursive) probabilistic inference problem.

Assuming the first-order Markov process, a discrete dynamic state-space system is described by a pair of equations:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}; \boldsymbol{\theta}) + \mathbf{q}_{t-1} \tag{2.1}$$

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta}) + \mathbf{r}_t, \tag{2.2}$$

where the first equation represents the system (state) model, describing the evaluation of the states $\mathbf{x}_t$ as a function of time. Here, $\mathbf{q}_{t-1}$ is the process (state) noise that drives the dynamic system through an arbitrary (possibly nonlinear and time-varying) transition function $\mathbf{f}$, and $\mathbf{u}_t$ is the exogenous input to the system that is usually assumed known (though later in this thesis $\mathbf{u}_t$ will be considered unknown). The second equation represents the measurement (observation) model, where the measurement noise $\mathbf{r}_t$ corrupting the observation of the (hidden) states through arbitrary observation function $\mathbf{h}$. Both $\mathbf{f}$ and $\mathbf{h}$ can be parameterized using a set of parameters $\boldsymbol{\theta}$. In a Markovian setting, the current state $\mathbf{x}_t$ depends only on the immediate past state $\mathbf{x}_{t-1}$ through the state-transition distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$; i.e. conditional probability density. The observations $\mathbf{y}_t$ are conditionally independent, given the state, and are generated according to the observation likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$ [59]. Therefore, the dynamic state-space model, together with the known statistics of the noise (and the prior distribution of the system states), defines a probabilistic generative model of how system evolves over time and how we (partially or inaccurately) observe this hidden state.

The DSSM can be also interpreted as dynamic Bayesian network (DBN) [60] with directed edges connecting the (hidden) states. These edges directly model the temporal flow of information with the implied causality constraints. For example, the first-order Markov process and conditional observation independence of the DSSM, is modeled by the specific graphical structure and relationship of the directed edges and nodes called the directed acyclic graph (DAG) as depicted in Figure 2.1.

The optimal solution to the above inference problem is given by the recursive

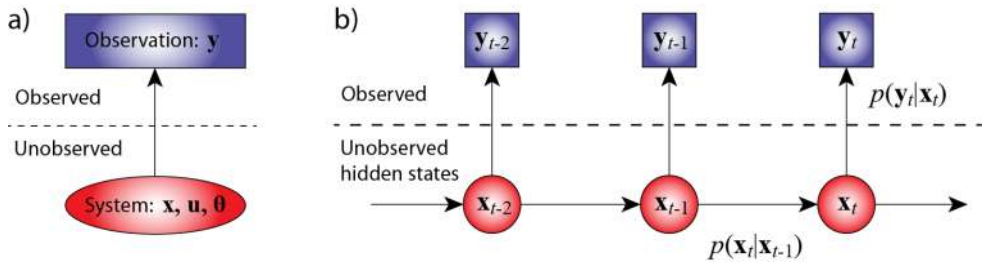


Figure 2.1 Schematic diagrams of probabilistic inference. (a) Given noisy observation y, what can we infer about system state, parameters or input. (b) Graphical model of a probabilistic dynamic state-space model. This representation is also known as a directed acyclic graph (DAG) in the graph theory field.

Bayesian estimation algorithm, which recursively updates the posterior density of the system state $p(\mathbf{x}_t|\mathbf{y}_t)$ as new observations arrive. Generally in Bayesian framework, the posterior density of the states $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ given all the observations $\mathbf{y}_{1:t}$, embodies the complete solution to the probabilistic inference problem. In other words, $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ contains all information necessary to calculate an optimal estimate of the state, such as the conditional mean:

$$\hat{\mathbf{x}}_{t|t} = \mathbb{E}[\mathbf{x}_t|\mathbf{y}_{1:t}] = \int_{\mathbb{R}^{n_x}} \mathbf{x}_t \, p(\mathbf{x}_t|\mathbf{y}_{1:t}) d\mathbf{x}_t, \tag{2.3}$$

and the covariance matrix, as a measure of accuracy of the estimate $\hat{\mathbf{x}}_{t|t}$:

$$\begin{aligned} \mathbf{P}_{t|t} &= \mathbb{E}\left[\left(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\right)\left(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\right)^T\right] \\ &= \int_{\mathbb{R}^{n_x}} \left(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\right)\left(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\right)^T p(\mathbf{x}_t|\mathbf{y}_{1:t}) d\mathbf{x}_t. \end{aligned} \tag{2.4}$$

In order to compute posterior density recursively (online) as new observations arrive, one make use of Bayes rule and dynamic state-space model, receiving recursive update:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{y}_{1:t}) &= \frac{p(\mathbf{y}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{y}_{1:t})} \\ &= \frac{p(\mathbf{y}_t, \mathbf{y}_{1:t-1}|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \\ &= \frac{p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \mathbf{x}_t)p(\mathbf{y}_{1:t-1}|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})p(\mathbf{y}_{1:t-1})} \\ &= \frac{p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})p(\mathbf{y}_{1:t-1})p(\mathbf{x}_t)}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})p(\mathbf{y}_{1:t-1})p(\mathbf{x}_t)} \\ &= \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})}. \end{aligned} \tag{2.5}$$

This relation can be decomposed into the two steps:

- *Time update:* The old posterior distribution at time $t-1$, $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$, is projected forward in time to yield predictive distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$. This involves multiplication of old posterior $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ with transient prior distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ obtained through (2.1), and integration of resulting joint distribution with respect to $\mathbf{x}_{t-1}$ [61]. This represents so called *Chapman-Kolmogorov equation*:

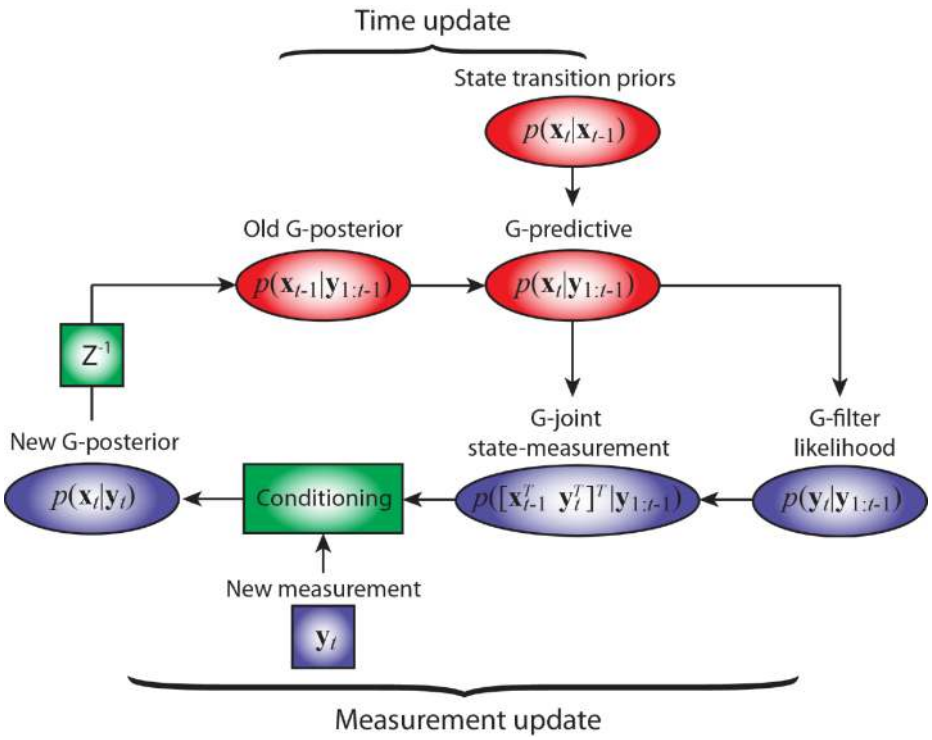


Figure 2.2 Diagram of the recursive Bayesian filter under Gaussian assumption. Here 'G-' stands for Gaussian. Based on [55].

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int_{\mathbb{R}^{n_x}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1}. \tag{2.6}$$

- *Measurement update:* The updated posterior distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ is then obtained as a multiplication of predicted posterior distribution and observation likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$, which incorporates the latest noisy measurement (obtained through (2.2)):

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{1}{C_t} p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) p(\mathbf{y}_t|\mathbf{x}_t), \tag{2.7}$$

where

$$C_t = p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int_{\mathbb{R}^{n_x}} p(\mathbf{y}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) d\mathbf{x}_t \tag{2.8}$$

is the normalizing constant. The summation over the time sequence of normalizing constants $C_{1:t}$ is in Bayesian terminology called the evidence, or marginal likelihood of the observations [62]. In our case, it will be represented by log likelihood:

$$\ln\left(p(\mathbf{y}_{1:t})\right) = \sum_{i=1}^{t} \ln(C_i). \tag{2.9}$$

Although this measure is commonly ignored during the first level of inference, i.e. during estimation of the states or parameters, it becomes very important in the second level of inference, where one wants to infer the most plausible model (model selection) [63].

The time update and measurement update are both carried out at every time step through the computation of the Bayesian model. In this way, we obtain posterior *filtering* distribution of the state $\mathbf{x}_t$ at time t given the history of the measurement up to the time step t, $p(\mathbf{x}_t|\mathbf{y}_{1:t})$. However, it is also possible to obtain the *smoothing* distribution $p(\mathbf{x}_t|\mathbf{y}_{1:T})$, where the posterior of the state is computed at the time step t after receiving the measurements up to time step T, where $T > t$. The difference between filter and smoother is that the optimal filter computes its estimates using only the measurements obtained before and at the time of the step t, whereas the optimal smoother uses also the future measurements for computing its estimates [64, 65], this generally provides more refined estimates of the states. Specifically, after obtaining the filtering posterior state distribution, the smoothing posterior is computed at each time step conditionally to all measurements up to the time step T. The backward recursive equation for Bayesian (fixed interval) smoothing is defined as:

- *Forward-Backward smoothing:* Given measurements up to time T $(> t)$, $\mathbf{y}_{1:T}$, and using Bayes' rule, the smoother density is factored as follows [66]:

$$\begin{aligned} p\left(\mathbf{x}_t|\mathbf{y}_{1:T}\right) &= \int_{\mathbb{R}^{n_x}} p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:T})\, d\mathbf{x}_{t+1} \\ &= \int_{\mathbb{R}^{n_x}} p\left(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}\right) p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:T})\, d\mathbf{x}_{t+1}. \end{aligned} \tag{2.10}$$

Due to the Markovian nature of the state-space model, given knowledge of $\mathbf{y}_t$ and $\mathbf{x}_{t+1}$, the state $\mathbf{x}_t$ is uncorrelated with future measurements $\mathbf{y}_{t+1:T}$. It means that $p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:T}) = p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})$, then we may express the smoothing density as:

$$p\left(\mathbf{x}_t|\mathbf{y}_{1:T}\right) = \int_{\mathbb{R}^{n_x}} p\left(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}\right) p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})\, d\mathbf{x}_{t+1}$$

$$= p(\mathbf{x}_t|\mathbf{y}_{1:t}) \int_{\mathbb{R}^{n_x}} \frac{p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}) p(\mathbf{x}_{t+1}|\mathbf{x}_t)}{p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})} d\mathbf{x}_{t+1}. \quad (2.11)$$

The above equation (2.11) says that the smoother has to perform two different passes: During forward filtering pass, it computes the posterior density $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ and the predictive density $p(\mathbf{x}_{t+1}|\mathbf{y}_t)$; during the backward smoothing pass, it recursively computes the smoothed density $p(\mathbf{x}_t|\mathbf{y}_{1:T})$ backward in time starting from $t = T$.

For linear and Gaussian dynamic systems, where **f** and **h** are linear functions and additive noise and state prior distributions are Gaussian, the solution to the filtering recursion is obtained by celebrated *Kalman filter* [67]. In this special case, the solution will be optimal in the minimum-mean-square-error (MMSE) sense, the maximum likelihood (ML) sense, and also in the maximum *a posteriori* (MAP) sense (derivation of Kalman filter from MAP perspective can be found in [59, 68]). The solution to the forward-backward smoothing is then obtained via *Rauch-Tung-Striebel smoother* (RTS) [69]; i.e. fixed interval Kalman smoother.

Unfortunately, in more realistic environment, which is nonlinear and possibly non-Gaussian, the optimal Bayesian recursion is intractable and an approximate solution must be used. Numerous approximation solutions to the recursive Bayesian estimation problem have been proposed over the last couple of decades, in a variety of fields. These methods can be grouped into the following four main categories:

- *Gaussian approximate methods:* These methods model the pertinent densities by Gaussian distributions, under assumption that a consistent minimum variance estimator (of the posterior state density) can be realized through the recursive propagation and updating of the first and second order moments of the true densities. Nonlinear filters that fall under this category are: a) the extended Kalman filter (EKF), which linearizes both the nonlinear process and measurement dynamics with a first-order Taylor expansion about current state estimate; b) the local linearization filter (LLF), which is similar to EKF, but the approximate discrete time model is obtained from piecewise linear discretization of nonlinear state equation; c) the unscented Kalman filter (UKF) [70], which chooses deterministic sample (sigma) points that capture the mean and covariance of a Gaussian density. When propagated through the nonlinear

function, these points capture the true mean and covariance up to a second-order of the nonlinear function; d) the divided difference filter (DDF) [71], which uses Stirling's interpolation formula. As with the UKF, DDF uses a deterministic sampling approach to propagate Gaussian statistics through the nonlinear function; f) the quadrature Kalman filter (QKF) [72], which uses the Gauss-Hermite numerical integration rule to calculate the recursive Bayesian estimation integrals, under a Gaussian assumption; g) the cubature Kalman filter (CKF), which is similar to UKF, but uses the spherical-radial integration rule; e) the Gaussian sum filters (GSF), which approximate both the predicted and posterior densities as a sum of Gaussian densities, where the mean and covariance of each Gaussian density is calculated using separate and parallel instances of EKF or UKF (or CKF).

- *Direct numerical integration methods:* these methods, also known as grid-based filters (GBF) or point-mass method, approximate the optimal Bayesian recursion integrals with large but finite sums over a uniform *n*-dimensional grid that covers the complete state-space in the area of interest. For even moderately high dimensional state-spaces, the computational complexity can become untenably large, which precludes any practical use of these filters [73].
- *Sequential Monte-Carlo (SMC) methods:* these methods (called particle filters) use a set of randomly chosen samples with associated weights to approximate the density [74]. Since the basic sampling dynamics (importance sampling) degenerates over time, the SMC method includes a re-sampling step. As the number of samples (particles) becomes larger, the Monte Carlo characterization of the posterior density becomes more accurate. However, the large number of samples often makes the use of SMC methods computationally prohibitive.
- *Variational Bayesian methods:* variational Bayesian methods approximate the true posterior distribution with a tractable approximate form. A lower bound on the marginal likelihood (evidence) of the posterior is then maximized with respect to the free parameters of this approximation [75].

The selection of suitable sub-optimal approximate solutions to the recursive Bayesian estimation problem represents a trade-off between global optimality on one hand and computational tractability (and robustness) on the other hand. In our case, the best criterion for sub-optimality is formulated as: "Do as best as you can, and not

more". Under this criterion, the natural choice is to apply the cubature Kalman filter [55]. The CKF is the closest known direct approximation to the Bayesian filter, which outperforms all other nonlinear filters in any Gaussian setting, including particle filters [55, 76, 77]. The CKF is numerically accurate, can capture true nonlinearity even in highly nonlinear systems, and it is easily extendable to high dimensional problems (the number of sample points grows linearly with the dimension of the state vector). Since the CKF belongs to the group of so-called Gaussian assumed density filters, which are considered as local approximation methods, we first introduce this more general framework and then move to the description of cubature approach.

2.3 Gaussian assumed density filtering and smoothing

In Gaussian optimal filtering framework, the filtering (smoothing) equations follow the assumption that filtering (smoothing) distributions are indeed Gaussian. The Gaussian approximation of the filtering distribution has then the form:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \approx \mathcal{N}\left(\mathbf{x}_t|\hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right), \tag{2.12}$$

where $\mathcal{N}\left(\mathbf{x}_t|\hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right)$ denotes the multivariate Gaussian distribution with mean $\hat{\mathbf{x}}_{t|t}$ and covariance $\mathbf{P}_{t|t}$.

Bayesian filtering

The dynamic state-space model (2.1)-(2.2) can be now adopted to nonlinear Bayesian filtering framework. Assuming that the process noise $\mathbf{q}_{t-1} \sim \mathcal{N}(0, \mathbf{Q}_{t-1})$ is zero mean and uncorrelated with the past measurements, the time update step (prediction) of the nonlinear filter can be obtained through calculating the following integrals that approximate the mean $\hat{\mathbf{x}}_{t|t-1}$ and error covariance matrix $\mathbf{P}_{t|t-1}$:

$$\begin{aligned}\hat{\mathbf{x}}_{t|t-1} &= \int_{\mathbb{R}^{n_x}} \mathbf{f}(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}; \boldsymbol{\theta})\, p(\mathbf{x}_{t-1}|\mathbf{y}_{t-1}) d\mathbf{x}_{t-1} \\ &= \int_{\mathbb{R}^{n_x}} \mathbf{f}(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}; \boldsymbol{\theta}) \times \mathcal{N}\left(\mathbf{x}_{t-1}|\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}\right) d\mathbf{x}_{t-1}\end{aligned} \tag{2.13}$$

$$\mathbf{P}_{t|t-1} = \int_{\mathbb{R}^{n_x}} \mathbf{f}(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}; \boldsymbol{\theta})\, \mathbf{f}^T(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}; \boldsymbol{\theta}) \times \mathcal{N}\left(\mathbf{x}_{t-1} | \hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}\right) d\mathbf{x}_{t-1} - \hat{\mathbf{x}}_{t|t-1}\hat{\mathbf{x}}_{t|t-1}^T + \mathbf{Q}_{t-1}, \tag{2.14}$$

where $\mathbf{Q}_{t-1}$ is the noise covariance matrix of process noise.

Next, during the measurement update step, the predicted estimate $\hat{\mathbf{y}}_{t|t-1}$ of the observation $\mathbf{y}_t$ is calculated. Again, under the assumption of zero mean measurement noise $\mathbf{r}_t \sim \mathcal{N}(0, \mathbf{R}_t)$, which is uncorrelated with both initial state $\mathbf{x}_0$ and the process noise $\mathbf{q}_t$ and also uncorrelated with the past observations, the prediction of current observation is given by:

$$\hat{\mathbf{y}}_{t|t-1} = \int_{\mathbb{R}^{n_x}} \mathbf{h}(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta}) \times \mathcal{N}\left(\mathbf{x}_t | \hat{\mathbf{x}}_{t|t-1}, \mathbf{P}_{t|t-1}\right) d\mathbf{x}_t. \tag{2.15}$$

The associated innovation covariance matrix has then the form:

$$\mathbf{P}_{yy,t|t-1} = \int_{\mathbb{R}^{n_x}} \mathbf{h}(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta})\, \mathbf{h}^T(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta}) \times \mathcal{N}\left(\mathbf{x}_t | \hat{\mathbf{x}}_{t|t-1}, \mathbf{P}_{t|t-1}\right) d\mathbf{x}_t - \hat{\mathbf{y}}_{t|t-1}\hat{\mathbf{y}}_{t|t-1}^T + \mathbf{R}_t, \tag{2.16}$$

where $\mathbf{R}_t$ is the covariance matrix of measurement noise. The cross covariance matrix between the state $\mathbf{x}_t$ and the observation $\mathbf{y}_t$ is:

$$\mathbf{P}_{xy,t|t-1} = \int_{\mathbb{R}^{n_x}} \mathbf{x}_t\, \mathbf{h}^T(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta}) \times \mathcal{N}\left(\mathbf{x}_t | \hat{\mathbf{x}}_{t|t-1}, \mathbf{P}_{t|t-1}\right) d\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1}\hat{\mathbf{y}}_{t|t-1}^T. \tag{2.17}$$

Although, each of these integral formulas represents different basic block of Bayesian filter, they all have one common feature, i.e. they are represented by product of nonlinear function with corresponding Gaussian function of known mean and covariance matrix. We will return to this property in the next Section 2.3.1.

Finally, in the last part of the measurement update step, one has to update the predictions from the time update to obtain filtering estimate of mean $\hat{\mathbf{x}}_{t|t}$ and error covariance matrix $\mathbf{P}_{t|t}$, which can be written similarly as for linear Kalman filter:

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right) \tag{2.18}$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{P}_{yy,t|t-1} \mathbf{K}_t^T \tag{2.19}$$

$$\mathbf{K}_t = \mathbf{P}_{xy,t|t-1} \mathbf{P}_{yy,t|t-1}^{-1}, \tag{2.20}$$

where the difference $\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right)$ in (2.19) is called the measurement innovation, or the residual. It reflects the discrepancy between the predicted measurement $\hat{\mathbf{y}}_{t|t-1}$ and actual measurement $\mathbf{y}_t$. The innovation is weighted by the optimal Kalman gain $\mathbf{K}_t$. This term minimizes the filtering error covariance $\mathbf{P}_{t|t}$ by weighting the innovation with respect to the prediction error covariance $\mathbf{P}_{t|t-1}$ [78, 79]. It is important to highlight the role of the error covariance matrix $\mathbf{P}_{t|t}$. It indicates how uncertain the state estimate $\hat{\mathbf{x}}_{t|t}$ is: a large values of covariance matrix indicate an inaccurate state estimate; the smaller the covariance, the larger the information content of the state estimate, i.e. the filter believes the prediction $\hat{\mathbf{x}}_{t|t-1}$ more, while the actual measurement $\mathbf{y}_t$ is trusted less.

Bayesian smoothing

The fixed-interval Gaussian assumed density smoother can be written in the form, where we first calculate the Gaussian integrals:

$$\hat{\mathbf{x}}_{t+1|t} = \int_{\mathbb{R}^{n_x}} \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta}) \times \mathcal{N}\left(\mathbf{x}_t | \hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right) d\mathbf{x}_t \tag{2.21}$$

$$\mathbf{P}_{t+1|t} = \int_{\mathbb{R}^{n_x}} \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta})\, \mathbf{f}^T(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta}) \times \mathcal{N}\left(\mathbf{x}_t | \hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right) d\mathbf{x}_t - \hat{\mathbf{x}}_{t+1|t} \hat{\mathbf{x}}_{t+1|t}^T + \mathbf{Q}_t, \tag{2.22}$$

$$\mathbf{P}_{t,t+1|t} = \int_{\mathbb{R}^{n_x}} \mathbf{x}_t\, \mathbf{f}^T(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta}) \times \mathcal{N}\left(\mathbf{x}_t | \hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right) d\mathbf{x}_t - \hat{\mathbf{x}}_{t|t} \hat{\mathbf{x}}_{t+1|t}^T. \tag{2.23}$$

Importantly, the terms $\hat{\mathbf{x}}_{t+1|t}$ and $\mathbf{P}_{t+1|t}$ are simply predicted mean and covariance from the Gaussian filter and the cross-covariance $\mathbf{P}_{t,t+1|t}$ can be computed during the filter pass as well. Consequently, the smoothing step follows, where the smoothing gain $\mathbf{A}_t$, the smoother mean $\hat{\mathbf{x}}_{t|T}^s$, and the error covariance $\mathbf{P}_{t|T}^s$ are computed:

$$\mathbf{A}_t = \mathbf{P}_{t,t+1|t} \mathbf{P}_{t+1|t}^{-1} \tag{2.24}$$

$$\hat{\mathbf{x}}_{t|T}^s = \hat{\mathbf{x}}_{t|t} + \mathbf{A}_t \left(\hat{\mathbf{x}}_{t+1|T}^s - \hat{\mathbf{x}}_{t+1|t}\right) \tag{2.25}$$

$$\mathbf{P}_{t|T}^s = \mathbf{P}_{t|t} - \mathbf{A}_t \left(\mathbf{P}_{t+1|t} - \mathbf{P}_{t+1|T}^s\right) \mathbf{A}_t^T. \tag{2.26}$$

This is again the equivalent to the linear version of RTS smoother.

2.3.1 Cubature integration rules

In this section, we describe numerical integration rules [55], which enable an efficient and accurate approximation of multidimensional weighted integral of the form:

$$I(\mathbf{f}) = \int_{\mathcal{D}} \mathbf{f}(\mathbf{x})\, w(\mathbf{x}) d\mathbf{x}, \tag{2.27}$$

where $\mathbf{f}(.)$ is an arbitrary, presumably nonlinear function, $\mathcal{D} \subseteq \mathbb{R}^{n_x}$ is a region of integration, and the known weighting function $w(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathcal{D}$. In the case of a Gaussian-weighted integral, $w(\mathbf{x})$ represents a Gaussian density. Besides the case when the function $\mathbf{f}(.)$ is linear, the solution to the above integral is difficult to obtain. Therefore, one uses numerical integration methods which allow its computation. The main goal of these methods is to find a set of points $\mathbf{x}_i$ and weights w_i that approximates the integral by a weighted sum of N function evaluations:

$$I(\mathbf{f}) \approx \sum_{i=1}^{N} w_i \mathbf{f}(\mathbf{x}_i). \tag{2.28}$$

In general, these integration methods can be based either on product rules or on non-product rules. As an example of product rule we mention Gauss-Hermite quadrature [80]. However, due to its higher computational complexity, which increases exponentially with the dimension of the state, this approach is not considered in this work. On the other hand, non-product rules include many Monte Carlo based methods [81-84], which suffer from curse of dimensionality as well. Nevertheless, there is one representative of non-product rules, which yields reasonable accuracy, requires small number of function evaluation and is easily extendable to high dimensions. This method applies the third-degree fully-symmetric cubature rule [55] to approximate *n*-dimensional Gaussian weighted integrals; i.e. integrals of the form $nonlinear\ function \times Gaussian\ density$. Critically, this cubature rule defines a way how to deterministically select a set of cubature points, and their corresponding weights, so that they completely capture the true mean and covariance of the prior random variable $\mathbf{x} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{P})$:

$$\int_{\mathbb{R}^{n_x}} \mathbf{f}(\mathbf{x}) \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}, \mathbf{P}) d\mathbf{x} \approx \sum_{i=1}^{2n} w_i \mathbf{f}(\boldsymbol{\mathcal{X}}_i), \tag{2.29}$$

where the weights are simply $w_i = \frac{1}{2n}$, with n equals the state dimension, and

$$\boldsymbol{\mathcal{X}}_i = \bar{\mathbf{x}} + \boldsymbol{\xi}_i\sqrt{\mathbf{P}}, \qquad i = 1, \dots, 2n. \tag{2.30}$$

This involves factorization of error covariance matrix $\mathbf{P} = \sqrt{\mathbf{P}}\sqrt{\mathbf{P}}^T$ and the elementary cubature points are:

$$\boldsymbol{\xi}_i = \begin{cases} \sqrt{n}\mathbf{e}_i, & i = 1, 2, \dots, n \\ -\sqrt{n}\mathbf{e}_i, & i = n+1, n+2, \dots, 2n. \end{cases} \tag{2.31}$$

Here $\mathbf{e}_i$ represents the *i*-th column vector, whose *i*-th entry is a unit and all other entries are zero. From this definition, it can be seen that the cubature points are distributed uniformly on a sphere centered at the origin, and their number increases linearly with the state dimension. Additionally, the points and weights of cubature rule are independent of the integrand $\mathbf{f}(\mathbf{x})$. It means they can be computed in advance and used during the execution of the nonlinear filter.

Example of cubature points approximation

An illustrative example of the effect of a non-linear transformation is shown in Figure 2.3. A bivariate normal distribution $\mathbf{x} = [r, \theta]^T \sim \mathcal{N}\left([80, 0.8]^T, \mathrm{diag}(40, 0.4)\right)$ is transformed through a nonlinear transformation:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}, \tag{2.32}$$

which corresponds to making a change of coordinates from radial to Cartesian. We first draw 10000 Monte Carlo (MC) samples and propagate them through this nonlinear function. As it can be seen in Figure 2.2b, the resulting distribution (in Cartesian plane) has a characteristic "banana-shape" and clearly does not resemble a Gaussian. The true mean and covariance (ellipse) of this distribution are indicated by blue color. The cubature points are drawn from the intersection points of unit circle and the axes (Figure 2.2a). When these four cubature points are propagated through the transformation, they approximatelly catch the charasteric "banana-shape" of the distribution and when the mean and covariance is calculated according to (2.34) and (2.35), indicated by red color, they almost exactly match the true mean and covarince based on MC samples.

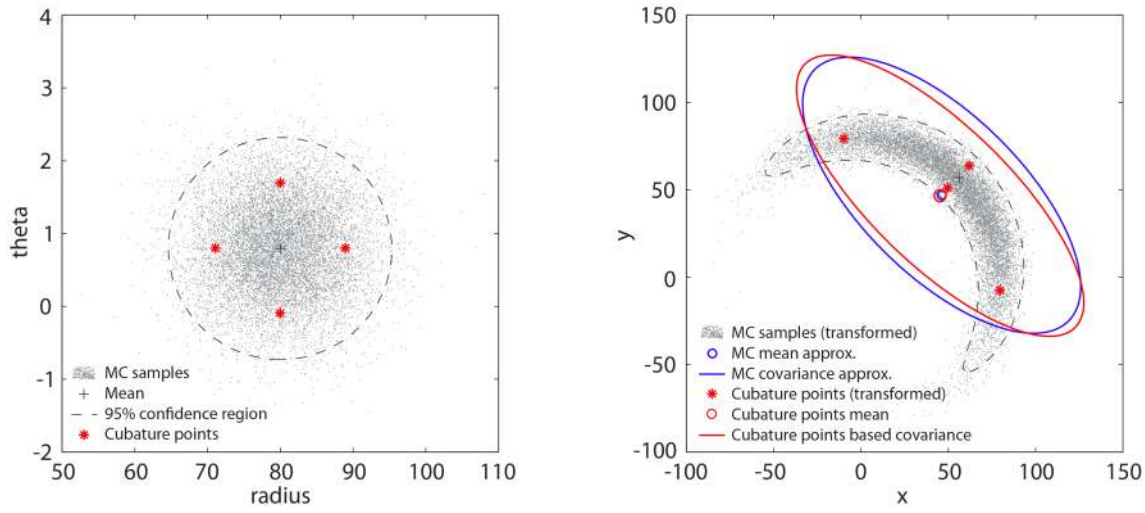


Figure 2.3 Illustrative example of cubature rules approximation. Here the cubature points (four for two states) are propagated through a nonlinear function to approximate the true mean and covariance.

2.3.2 Statistical linear regression

After an introduction of Gaussian approximations to Bayesian optimal recursion based on cubature integration rules (or generally based on any sigma points approach), it is enriching to mention its alternative interpretation from the perspective of statistical linear regression [85]. Specifically, we will show how the statistics computed via cubature integration rules can be used to linearize a nonlinear function in MMSE sense.

Consider a nonlinear function $\mathbf{y} = \mathbf{h}(\mathbf{x})$ evaluated in m points $(\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{Y}}_i)$, i.e.:

$$\boldsymbol{\mathcal{Y}}_i = \mathbf{h}(\boldsymbol{\mathcal{X}}_i), \quad i = 1, \dots, m, \tag{2.33}$$

where the points $\boldsymbol{\mathcal{X}}_i$ are chosen in the way that they capture certain statistical properties of $\mathbf{x}$ (such as the mean and covariance) through the sample based estimator of the form:

$$\bar{\mathbf{x}} = \sum_{i=1}^{m} w_i \boldsymbol{\mathcal{X}}_i \tag{2.34}$$

$$\mathbf{P} = \sum_{i=1}^{m} w_i (\boldsymbol{\mathcal{X}}_i - \bar{\mathbf{x}})(\boldsymbol{\mathcal{X}}_i - \bar{\mathbf{x}})^T. \tag{2.35}$$

Similarly, after propagation of sample points $\boldsymbol{\mathcal{X}}_i$ through nonlinear function $\mathbf{h}(.)$ we obtain a regression points $\boldsymbol{\mathcal{Y}}_i$, for which we define the following posterior statistics (mean, covariance and cross-covariance):

$$\bar{\mathbf{y}} = \sum_{i=1}^{m} w_i \boldsymbol{\mathcal{Y}}_i \tag{2.36}$$

$$\mathbf{P}_{yy} = \sum_{i=1}^{m} w_i (\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}})(\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}})^T \tag{2.37}$$

$$\mathbf{P}_{xy} = \sum_{i=1}^{m} w_i (\boldsymbol{\mathcal{X}}_i - \bar{\mathbf{x}})(\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}})^T. \tag{2.38}$$

Now, the objective is to find the linear regression of the form (2.39), which approximates a nonlinear function $\mathbf{y} = \mathbf{h}(\mathbf{x})$.

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{2.39}$$

Here $\mathbf{A}$ and $\mathbf{b}$ are a matrix and a vector, respectively, which are estimated by minimizing the sum of square errors:

$$\{\mathbf{A}, \mathbf{b}\} = \arg\min \sum_{i=1}^{m} w_i \boldsymbol{\varepsilon}_i^T \boldsymbol{\varepsilon}_i \tag{2.40}$$

with linearization error:

$$\boldsymbol{\varepsilon}_i = \boldsymbol{\mathcal{Y}}_i - (\mathbf{A}\boldsymbol{\mathcal{X}}_i + \mathbf{b}) \tag{2.41}$$

representing the deviations between the function values of the nonlinear and linearized function in the sampling points. The solution to (2.40) is given by [65]:

$$\mathbf{A} = \mathbf{P}_{xy}^T \mathbf{P}^{-1} \tag{2.42}$$

$$\mathbf{b} = \bar{\mathbf{y}} - \mathbf{A}\bar{\mathbf{x}}. \tag{2.43}$$

The linearization error has zero mean $\bar{\boldsymbol{\varepsilon}} = 0$ and covariance:

$$\begin{aligned}
\mathbf{P}_{\varepsilon} &= \sum_{i=1}^{m} w_i \boldsymbol{\varepsilon}_i \, \boldsymbol{\varepsilon}_i^T \\
&= \sum_{i=1}^{m} w_i [\boldsymbol{\mathcal{Y}}_i - (\mathbf{A}\boldsymbol{\mathcal{X}}_i + \mathbf{b})][\boldsymbol{\mathcal{Y}}_i - (\mathbf{A}\boldsymbol{\mathcal{X}}_i + \mathbf{b})]^T \\
&= \sum_{i=1}^{m} w_i [\boldsymbol{\mathcal{Y}}_i - \mathbf{A}\boldsymbol{\mathcal{X}}_i - \bar{\mathbf{y}} - \mathbf{A}\bar{\mathbf{x}}][\boldsymbol{\mathcal{Y}}_i - \mathbf{A}\boldsymbol{\mathcal{X}}_i - \bar{\mathbf{y}} - \mathbf{A}\bar{\mathbf{x}}]^T
\end{aligned}$$

$$\begin{aligned} &= \mathbf{P}_{yy} - \mathbf{AP} - \mathbf{PA}^T + \mathbf{APA}^T \\ &= \mathbf{P}_{yy} - \mathbf{APA}^T. \end{aligned} \tag{2.44}$$

Then the posterior statistics of **y** can be approximated by the following statistical regression form:

$$\bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{x}} + \mathbf{b} \tag{2.45}$$

$$\mathbf{P}_{yy} = \mathbf{APA}^T + \mathbf{P}_{\varepsilon}. \tag{2.46}$$

Notice that the linearization error covariance $\mathbf{P}_{\varepsilon}$ is added to the linearly propagated prior covariance $\mathbf{APA}^T$, to form the posterior covariance $\mathbf{P}_{yy}$. This implies that the more sever the nonlinearity is over the "uncertainty region" of **x**, the larger the linearization error and error covariance will be, and accordingly the normal linear approximation of $\mathbf{P}_{yy}$ and $\mathbf{APA}^T$ will be less accurate. The correction term $\mathbf{P}_{\varepsilon}$ thus needs to be large to compensate for it [59].

It is now possible to show how the cubature point approach (or any sigma point approach) makes the effective use of statistical linear regression, when used to approximate the mean and covariance within the Kalman filter framework. If we assume that the regression points are generated based on the (propagated) cubature points, and we substitute (2.36) into (2.43) and the result into (2.45), we obtain:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{A}\bar{\mathbf{x}} + \sum_{i=1}^{m} w_i \boldsymbol{\mathcal{Y}}_i - \mathbf{A}\bar{\mathbf{x}} \\ &= \sum_{i=1}^{2n} w_i \boldsymbol{\mathcal{Y}}_i. \end{aligned} \tag{2.47}$$

We can see that under this assumption, the posterior mean calculated by statistical linear regression is equivalent to the posterior mean calculated by using only the weighted average of cubature points. Similarly, if we substitute (2.37) into (2.44) and the result into (2.46), we obtain cubature-point based approximation of posterior error covariance

$$\begin{aligned} \mathbf{P}_{yy} &= \mathbf{APA}^T + \sum_{i=1}^{m} w_i (\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}})(\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}})^T - \mathbf{APA}^T \\ &= \sum_{i=1}^{2n} w_i (\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}})(\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}})^T, \end{aligned} \tag{2.48}$$

which is equivalent to the posterior error covariance calculated by statistical linear regression. The same can be done for the cross-covariance resulting in the form identical to (2.38). This form of the mean and covariance approximations is then utilized in the nonlinear Kalman filtering and smoothing algorithms as we will see in the next sections. Additionally, it is noteworthy that the linearized error covariance $\mathbf{P}_{\varepsilon}$ is never explicitly calculated during the cubature-point approach, but its effect is indirectly incorporated through the way, in which the posterior statistics are approximated.

2.3.3 Cubature Kalman Filter

The cubature Kalman filter [55] is a recursive, nonlinear and derivative free filtering algorithm, which computes the first two moments (i.e. mean and covariance) of all conditional densities by using the above described third-degree cubature integration rules. The application of cubature rules leads to an even number of equally weighted cubature points ($2n$ point, with n being dimensionality of the state vector), where all these points are distributed uniformly on an ellipsoid centered at the origin. In contrast, unscented Kalman filter (UKF) applies the unscented transform, which uses an odd number of sigma points ($2n + 1$), also distributed on ellipsoid but with non-zero center point. This center sigma point often receives more weighting power than other non-center points through additional scaling parameter κ. It is exactly the inclusion of parameter κ, which causes lower performance of UKF compared to CKF. It is interesting to note that the original UKF filter will be equivalent to CKF if $\kappa = 0$ [66]. Unlike extended Kalman filter, CKF effectively approximates both the Jacobian and Hessian accurately (in statistically average sense) through its sigma point propagation, without the need to perform any analytic differentiation.

In order to evaluate the dynamic state-space model described by (2.1)-(2.2), the CKF includes two standard Kalman filter steps: a) a time update, after which the predicted density $p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \mathcal{N}(\hat{\mathbf{x}}_{t|t-1}, \mathbf{P}_{t|t-1})$ is computed; and b) a measurement update, after which the posterior density $p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \mathcal{N}(\hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t})$ is computed. The algorithm of CKF, for a discrete system with additive noise, is summarized below. Note that in case when the process noise and measurement noise are not purely additive as it is assumed here, it is desirable to augment the state vector also with the noise components [20, 59].

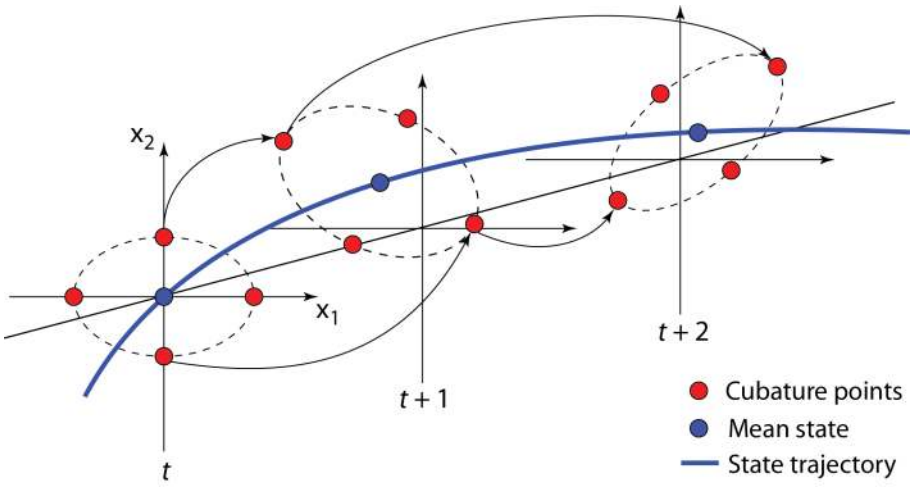


Figure 2.4 Illustration of cubature points propagation during time update of CKF. The cubature points in the two-dimensional state-space are propagated between time steps. The circles represent cubature points; the new cubature point set at time $t+1$ is computed by simply propagating the old cubature point set at time t through the process equation.

Algorithm 1. The cubature Kalman filter (CKF) - additive noise

- *Initialization:*

$$\hat{\mathbf{x}}_0 = \mathbb{E}[\mathbf{x}_0] \quad \text{and} \quad \mathbf{P}_0 = \mathbb{E}[(\mathbf{x}_0 - \hat{\mathbf{x}}_0)(\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T] \tag{2.49}$$

- *For* $t = 1, \dots, T$

  Time update:

  1. Factorize the state error covariance matrix:

$$\mathbf{P}_{t-1|t-1} = \mathbf{S}_{t-1|t-1}\mathbf{S}_{t-1|t-1}^T \tag{2.50}$$

  2. Evaluate the cubature points ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}_{i,t-1|t-1} = \mathbf{S}_{t-1|t-1}\boldsymbol{\xi}_i + \hat{\mathbf{x}}_{t-1|t-1} \tag{2.51}$$

  3. Propagate the cubature points through process model ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}^*_{i,t|t-1} = \mathbf{f}\left(\boldsymbol{\mathcal{X}}_{i,t-1|t-1}, \mathbf{u}_{t-1}; \boldsymbol{\theta}\right) \tag{2.52}$$

  4. Estimate the predicted state:

$$\hat{\mathbf{x}}_{t|t-1} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{X}}^*_{i,t|t-1} \tag{2.53}$$

  5. Estimate the predicted error covariance matrix:

$$\mathbf{P}_{t|t-1} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{X}}^*_{i,t|t-1}\boldsymbol{\mathcal{X}}^{*T}_{i,t|t-1} - \hat{\mathbf{x}}_{t|t-1}\hat{\mathbf{x}}^T_{t|t-1} + \mathbf{Q}_{t-1} \tag{2.54}$$

Measurement update:

6. Factorize the state error covariance matrix:

$$\mathbf{P}_{t|t-1} = \mathbf{S}_{t|t-1}\mathbf{S}_{t|t-1}^{T} \tag{2.55}$$

7. Evaluate the cubature points (i $= 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}_{i,t|t-1} = \mathbf{S}_{t|t-1}\boldsymbol{\xi}_i + \hat{\mathbf{x}}_{t|t-1} \tag{2.56}$$

8. Propagate the cubature points through measurement model ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{Y}}_{i,t|t-1} = \mathbf{h}\left(\boldsymbol{\mathcal{X}}_{i,t|t-1}, \mathbf{u}_t; \boldsymbol{\theta}\right) \tag{2.57}$$

9. Estimate the predicted measurement:

$$\hat{\mathbf{y}}_{t|t-1} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{Y}}_{i,t|t-1} \tag{2.58}$$

10. Estimate the innovation covariance matrix:

$$\mathbf{P}_{yy,t|t-1} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{Y}}_{i,t|t-1}\boldsymbol{\mathcal{Y}}_{i,t|t-1}^{T} - \hat{\mathbf{y}}_{t|t-1}\hat{\mathbf{y}}_{t|t-1}^{T} + \mathbf{R}_t \tag{2.59}$$

11. Estimate the cross-covariance matrix:

$$\mathbf{P}_{xy,t|t-1} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{X}}_{i,t|t-1}^{*}\boldsymbol{\mathcal{Y}}_{i,t|t-1}^{T} - \hat{\mathbf{x}}_{t|t-1}\hat{\mathbf{y}}_{t|t-1}^{T} \tag{2.60}$$

12. Estimate the Kalman gain:

$$\mathbf{K}_t = \mathbf{P}_{xy,t|t-1}\mathbf{P}_{yy,t|t-1}^{-1} \tag{2.61}$$

13. Estimate the updated state:

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right) \tag{2.62}$$

14. Estimate the updated error covariance matrix:

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{P}_{yy,t|t-1}\mathbf{K}_t^{T} \tag{2.63}$$

- Note that there are many types of matrix decomposition techniques that factorize the covariance matrix $\mathbf{P}$ into a form $\mathbf{P} = \mathbf{S}\mathbf{S}^{T}$. For example, the Cholesky decomposition, the eigenvector decomposition or the singular value decomposition (SVD). We prefer to apply SVD, since it is the most robust algorithm to factorize a covariance matrix especially when the covariance becomes nearly singular. The SVD of $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^{T}$, where $\mathbf{U}$ is matrix containing

eigenvectors and **D** is a diagonal matrix of eigenvalues. The square-root of the matrix **P** is then given by $\mathbf{S} = \mathbf{U}\sqrt{\mathbf{D}}\mathbf{U}^T$.

2.3.4 Cubature Rauch-Tung-Striebel smoother

The same approximation principles that were used in cubature Kalman filter can be applied also during the backward pass of the Rauch-Tung-Striebel (RTS) smoother, yielding the cubature RTS smoother. The backward pass is used for computing suitable corrections to the forward filtering results to obtain the smoothing solution $p(\mathbf{x}_t, \mathbf{y}_{1:T}) = \mathcal{N}\left(\hat{\mathbf{x}}_{t|T} | \hat{\mathbf{x}}_{t|T}^S, \mathbf{P}_{t|T}^S\right)$. Because the filtering and smoothing estimates of the last time step T are the same, we make $\hat{\mathbf{x}}_{T|T}^S = \hat{\mathbf{x}}_{T|T}$, $\mathbf{P}_{T|T}^S = \mathbf{P}_{T|T}$. This means the recursion can be used for computing the smoothing estimates of all time steps by starting from the last step $t = T$ and proceeding backward to the initial step $t = 0$. To accomplish this, all estimates of $\hat{\mathbf{x}}_{0:T}$ and $\mathbf{P}_{0:T}$ from the forward pass have to be stored to be later reused during the backward pass. Note that we will use an abbreviation CKS to refer to the forward run of cubature Kalman filter followed by the backward run of the cubature RTS smoother.

Algorithm 2. The cubature Rauch-Tung-Striebel smoother - additive noise

- *Initialization:*

$$\hat{\mathbf{x}}_{T|T}^S = \hat{\mathbf{x}}_{T|T} \quad \text{and} \quad \mathbf{P}_{T|T}^S = \mathbf{P}_{T|T} \tag{2.64}$$

- *For* $t = T - 1, \dots, 0$

1. Factorize the error covariance:

$$\mathbf{P}_{t|t} = \mathbf{S}_{t|t}\mathbf{S}_{t|t}^T \tag{2.65}$$

2. Evaluate the cubature points (i = 1, … , m):

$$\boldsymbol{\mathcal{X}}_{i,t|t} = \mathbf{S}_{t|t}\boldsymbol{\xi}_i + \hat{\mathbf{x}}_{t|t} \tag{2.66}$$

3. Propagate the cubature points through process model ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}_{i,t+1|t}^* = \mathbf{f}\left(\boldsymbol{\mathcal{X}}_{i,t|t}, \mathbf{u}_t; \boldsymbol{\theta}\right) \tag{2.67}$$

4. Estimate the predicted state:

$$\hat{\mathbf{x}}_{t+1|t} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{X}}^{*}_{i,t+1|t} \tag{2.68}$$

5. Estimate the predicted error covariance matrix:

$$\mathbf{P}_{t+1|t} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{X}}^{*}_{i,t+1|t}\, \boldsymbol{\mathcal{X}}^{*T}_{i,t+1|t} - \hat{\mathbf{x}}_{t+1|t}\hat{\mathbf{x}}^{T}_{t+1|t} + \mathbf{Q}_t \tag{2.69}$$

6. Estimate the cross-covariance matrix:

$$\mathbf{P}_{t,t+1|t} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{X}}_{i,t|t}\, \boldsymbol{\mathcal{X}}^{*T}_{i,t+1|t} - \hat{\mathbf{x}}_{t|t}\hat{\mathbf{x}}^{T}_{t|t} \tag{2.70}$$

7. Estimate the smoother gain:

$$\mathbf{G}_t = \mathbf{P}_{t,t+1|t}\mathbf{P}^{-1}_{t+1|t} \tag{2.71}$$

8. Estimate the smoothed state:

$$\hat{\mathbf{x}}^{s}_{t|T} = \hat{\mathbf{x}}_{t|t} + \mathbf{G}_t\left(\hat{\mathbf{x}}^{s}_{t+1|T} - \hat{\mathbf{x}}_{t+1|t}\right) \tag{2.72}$$

9. Estimate the smoothed error covariance:

$$\mathbf{P}^{s}_{t|T} = \mathbf{P}_{t|t} - \mathbf{G}_t\left(\mathbf{P}_{t+1|t} - \mathbf{P}^{s}_{t+1|T}\right)\mathbf{G}^{T}_t \tag{2.73}$$

- Note that the steps 1.-4. are not really needed for evaluation of the RTS smoother. These estimates are equivalent to the predictions from the time update step of the filter. Therefore, one can store the results of (2.52) and (2.53), and use them during the backward run of the smoother. This means that there is no need to propagate the states (cubature) points again through the nonlinear process model.

2.3.5 Square-root representation

In the Kalman filtering (and smoothing) framework it is important to preserve positive definitiveness and symmetry of the state covariance matrix during the entire recursion [55]. The cubature Kalman filter and smoother involves several numerically sensitive operations such as matrix square-rooting (2.50), matrix inversion (2.61), and substraction of two positive-definite matrices (2.63), which may violate the desired properties of covariance matrix and cause the filter to diverge [86]. Additionally, in each recursion cycle of cubature Kalman filter, one of the most costly operations is the

calculation of the matrix square-root of the state covariance in order to form set of cubature points. Therefore, it is natural to seek a solution, which would eliminate these drawbacks. Fortunately, it is possible to design a square-root version of cubature Kalman filter and smoother, which will preserve the positive (semi)definitiveness and symmetry of the covariance matrix, and hence improve the numerical stability especially during the state covariance update [87, 88]. In square-root version of CKF and CKS algorithm, the matrix square-rooting operations $\mathbf{P} = \mathbf{S}\mathbf{S}^T$ are avoided and the square-root covariance matrix $\mathbf{S}$ is propagated directly. This is achieved by utilizing the following powerful techniques of linear algebra:

- *QR decomposition:* The QR decomposition performs triangularization for covariance update. It factorizes the matrix $\mathbf{X}^T$ into an orthogonal matrix $\mathbf{Q}$ and upper triangular matrix $\mathbf{R}$ such that $\mathbf{X}^T = \mathbf{Q}\mathbf{R}$, and $\mathbf{X}\mathbf{X}^T = \mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R} = \mathbf{R}^T\mathbf{R} = \mathbf{S}\mathbf{S}^T$, where the resulting square-root (lower triangular) matrix is $\mathbf{S} = \mathbf{R}^T$. We use an abbreviation $\mathrm{Tria}(.)$ to denote QR decomposition of matrix $\mathbf{X}$ where only lower triangular matrix $\mathbf{S}$ is returned[3].
- *Efficient least-squares:* The least-squares method is used to compute the Kalman filter and smoother gain. If we substitute the innovation covariance matrix in (2.61) by its square-root representation we get the following expression:

$$\mathbf{K}_t\left(\mathbf{S}_{yy,t}\mathbf{S}_{yy,t}^T\right) = \mathbf{P}_{xy,t}. \tag{2.74}$$

  Since $\mathbf{S}_{yy,t}$ is square and lower triangular matrix, efficient forward substitution algorithm can be used to compute $\mathbf{K}_t$ directly without the need for a matrix inversion [59]. We use the symbol "/" to represent the matrix right-divide operator, which applies the forward substitution algorithm (as it is a common notation in Matlab®), see (2.87) and (2.94).

Although, the algorithms are now free of square-rooting operations, one still needs to calculate the square-root form of the state covariance matrix and also of the process and measurement noise covariance matrices during the initialization of the Kalman filter. This can be done by using the Cholesky factorization or preferably by the SVD decomposition as noted in Algorithm 1. Finally, the square-root versions of CKF and

[3] Note that matrices $\mathbf{Q}$ and $\mathbf{R}$ have nothing in common with process and measurement noise covariance matrices for which we use the same notation.

CKS, abbreviated as SCKF and SCKS, are summarized in Algorithm 3 and Algorithm 4, respectively.

Algorithm 3. The square-root cubature Kalman filter (SCKF) - additive noise

- *Initialization:*

$$\hat{\mathbf{x}}_0 = \mathbb{E}[\mathbf{x}_0], \quad \mathbf{S}_0 = \sqrt{\mathbb{E}[(\mathbf{x}_0 - \hat{\mathbf{x}}_0)(\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T]}, \quad \mathbf{S}_{q,t} = \sqrt{\mathbf{Q}_t}, \quad \text{and} \quad \mathbf{S}_{r,t} = \sqrt{\mathbf{R}_t} \tag{2.75}$$

- *For* $t = 1, \dots, T$

Time update:

1. Evaluate the cubature points ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}_{i,t-1|t-1} = \mathbf{S}_{t-1|t-1}\boldsymbol{\xi}_i + \hat{\mathbf{x}}_{t-1|t-1} \tag{2.76}$$

2. Propagate the cubature points through process model ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}^*_{i,t|t-1} = \mathbf{f}\left(\boldsymbol{\mathcal{X}}_{i,t-1|t-1}, \mathbf{u}_{t-1}; \boldsymbol{\theta}\right) \tag{2.77}$$

3. Estimate the predicted state:

$$\hat{\mathbf{x}}_{t|t-1} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{X}}^*_{i,t|t-1} \tag{2.78}$$

4. Estimate the predicted square-root error covariance matrix:

$$\mathbf{S}_{t|t-1} = \mathrm{Tria}\left(\left[\mathbf{X}_{t|t-1}, \mathbf{S}_{q,t}\right]\right) \tag{2.79}$$

with weighted and centered matrix:

$$\mathbf{X}_{t|t-1} = \frac{1}{\sqrt{m}}\left[\boldsymbol{\mathcal{X}}^*_{1,t|t-1} - \hat{\mathbf{x}}_{t|t-1}, \dots, \boldsymbol{\mathcal{X}}^*_{m,t|t-1} - \hat{\mathbf{x}}_{t|t-1}\right] \tag{2.80}$$

Measurement update:

5. Evaluate the cubature points ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}_{i,t|t-1} = \mathbf{S}_{t|t-1}\boldsymbol{\xi}_i + \hat{\mathbf{x}}_{t|t-1} \tag{2.81}$$

6. Propagate the cubature points through measurement model ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{Y}}_{i,t|t-1} = \mathbf{h}\left(\boldsymbol{\mathcal{X}}_{i,t|t-1}, \mathbf{u}_t; \boldsymbol{\theta}\right) \tag{2.82}$$

7. Estimate the predicted measurement:

$$\hat{\mathbf{y}}_{t|t-1} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{Y}}_{i,t|t-1} \tag{2.83}$$

8. Estimate the square-root of innovation covariance matrix:

$$\mathbf{S}_{yy,t|t-1} = \text{Tria}([\mathbf{Y}_{t|t-1} \quad \mathbf{S}_{r,t}]) \tag{2.84}$$

with weighted and centered matrix:

$$\mathbf{Y}_{t|t-1} = \frac{1}{\sqrt{m}}\left[\boldsymbol{y}_{1,t|t-1} - \hat{\mathbf{y}}_{t|t-1}, \dots, \boldsymbol{y}_{m,t|t-1} - \hat{\mathbf{y}}_{t|t-1}\right] \tag{2.85}$$

9. Estimate the cross-covariance matrix:

$$\mathbf{P}_{xy,t|t-1} = \mathbf{X}_{t|t-1}\mathbf{Y}_{t|t-1}^{T} \tag{2.86}$$

10. Estimate the Kalman gain:

$$\mathbf{K}_t = \left(\mathbf{P}_{xy,t|t-1}/\mathbf{S}_{yy,t|t-1}^{T}\right)/\mathbf{S}_{yy,t|t-1} \tag{2.87}$$

11. Estimate the updated state:

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right) \tag{2.88}$$

12. Estimate the updated square-root of error covariance matrix:

$$\mathbf{S}_{t|t} = \text{Tria}([\mathbf{X}_{t|t-1} - \mathbf{K}_t\,\mathbf{Y}_{t|t-1} \quad \mathbf{K}_t\mathbf{S}_{r,t}\,]) \tag{2.89}$$

Algorithm 4. The square-root cubature RTS smoother - additive noise

- *Initialization:*

$$\hat{\mathbf{x}}_{T|T}^{S} = \hat{\mathbf{x}}_{T|T} \quad \text{and} \quad \mathbf{S}_{T|T}^{S} = \mathbf{S}_{T|T} \tag{2.90}$$

- *For* $t = T-1, \dots, 0$

1. Compute the predicted error covariance matrix:

$$\mathbf{S}_{t+1|t} = \text{Tria}([\mathbf{X}_{t+1|t} \quad \mathbf{S}_{q,t}]), \tag{2.91}$$

where $\mathbf{X}_{t+1|t}$ is the weighted centered matrix, which was stored during a time update step of SCKF pass (2.80).

2. Estimate the predicted error covariance matrix:

$$\mathbf{X}_{t|t} = \frac{1}{\sqrt{m}}\left[\boldsymbol{\mathcal{X}}_{1,t|t} - \hat{\mathbf{x}}_{t|t}, \dots, \boldsymbol{\mathcal{X}}_{m,t|t} - \hat{\mathbf{x}}_{t|t}\right], \tag{2.92}$$

with $\boldsymbol{\mathcal{X}}_{i,t|t} = \mathbf{S}_{t|t}\boldsymbol{\xi}_i + \hat{\mathbf{x}}_{t|t}$, $i = 1{:}m$, where both $\hat{\mathbf{x}}_{t|t}$ and $\mathbf{S}_{t|t}$ are posterior estimates stored during the SCKF pass (2.88)-(2.89).

3. Estimate the predicted error covariance matrix:

$$\mathbf{P}_{t,t+1|t} = \mathbf{X}_{t|t}\mathbf{X}_{t+1|t}^{T} \tag{2.93}$$

4. Estimate the predicted error covariance matrix:

$$\mathbf{G}_t = \left(\mathbf{P}_{t,t+1|t}/\mathbf{S}_{t+1|t}^{T}\right)/\mathbf{S}_{t+1|t} \tag{2.94}$$

5. Estimate the predicted error covariance matrix:

$$\hat{\mathbf{x}}_{t|T}^{s} = \hat{\mathbf{x}}_{t|t} + \mathbf{G}_t\left(\hat{\mathbf{x}}_{t+1|T}^{s} - \hat{\mathbf{x}}_{t+1|t}\right) \tag{2.95}$$

6. Estimate the predicted error covariance matrix:

$$\mathbf{S}_{t|T} = \text{Tria}([\mathbf{X}_{t|t} - \mathbf{G}_t\,\mathbf{X}_{t|t-1} \quad \mathbf{G}_t\,\mathbf{S}_{t+1|T} \quad \mathbf{G}_t\,\mathbf{S}_q]) \tag{2.96}$$

- Derivation of the square-root cubature RTS smoother is the same as the one based on the unscented transform, which can be found in [89]. Alternative (but similar) solution is described in [66].

2.4 Sequential parameter estimation

Up to this point, we have assumed the parameters $\boldsymbol{\theta}$ occurring in the model equations to be known. However, it is very often the main aim of many analyses to determine the set of parameters so that they describe the model being the best fit to the measured data. Therefore, it will be an objective of this section to describe a parameter estimation framework, which well suits to the so far introduced sequential estimation theory.

Parameter estimation, which is very often called system identification, involves learning a nonlinear mapping:

$$\mathbf{y}_t = \mathbf{g}(\mathbf{x}_t; \boldsymbol{\theta}), \tag{2.97}$$

where $\boldsymbol{\theta}$ is the set of unknown parameters, which parameterize the (possibly nonlinear) mapping function $\mathbf{g}(.)$, i.e. mathematical model, between the input $\mathbf{x}_t$ and the output $\mathbf{y}_t$. Note, that in our case the function $\mathbf{g}(.)$ subsumes both hidden part of hemodynamic model and observation BOLD equation. Although many iterative optimization algorithms are available to perform parameter estimation from data available in bathes, where these include also the well known expectation maximization (EM) algorithm [90,

91], the interest of this thesis is to enable a sequential (online) estimation of the model parameters. In this sense, the (nonlinear) Kalman filtering framework may be used to estimate the parameters by considering a new state-space representation:

$$\begin{aligned} \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \mathbf{o}_{t-1} \\ \mathbf{y}_t &= \mathbf{g}(\mathbf{x}_t; \boldsymbol{\theta}_t) + \mathbf{r}_t, \end{aligned} \tag{2.98}$$

which can be regarded as a special case of general state estimation, where the parameters now represent the hidden states. Here $\boldsymbol{\theta}_{t-1}$ corresponds to a stationary process with identity state transition matrix, driven by process noise $\mathbf{o}_{t-1} \sim \mathcal{N}(0, \mathbf{O}_{t-1})$; i.e. Gaussian random walk model. The output of measurement equation, $\mathbf{y}_t$, then provides a nonlinear observation on $\boldsymbol{\theta}_t$. Clearly, one can use the same algorithms of nonlinear cubature Kalman filter/smoother, which were discussed in previous sections, with the above state-space system and perform parameter estimation in a sequential manner.

Nonlinear Kalman filters have a long history in application to neural networks in order to estimate coupling weights among single neurons. It is a common experiance that Bayesian filter-based training algorithms converge more rapidly than the gradient descent method. They are also well-suited to handle noisy and nonstationary training data. Especially, the introduction of sigma-point Kalman filters such as UKF or CKF led to significant improvement in obtaining robust globally optimal estimates in situations, where the standard EKF is very likely to get stuck in a non-optimal local minimum [59]. Additionally, the second-order information encoded in the filter estimated error covariance can often be used to prune the network weights. This is because the convergence goes along with the shrinkage of the parameter error covariance towards zero (this also forces the Kalman gain towards zero); i.e. filter becomes more confident about its predictive estimates.

In the following section, we will motivate the sequential parameter estimation from maximum *a posteriori* (MAP) perspective.

2.4.1 Maximum *a posteriori* estimate

Though the Kalman filter is mostly derived as a MMSE estimator [61, 67, 78], it has its interpretation also from MAP perspective [68]. Especially, it allows for interesting

analysis of the recursive Bayesian algorithm, when applied to parameter estimation. However, it is important to mention that the MAP interpretation of Kalman filter is constrained only to the environment, where the prior and posterior random variables are Gaussian, whereas MMSE interpretation of the optimal estimate is valid also for non-Gaussian environment [59]. To develop a sequential MAP learning procedure for parameter estimation, we use the state-space model (2.98) and Bayes rule to express the posterior distribution of parameters $\boldsymbol{\theta}_t$ conditioned on all of the observation $\mathbf{y}_{1:t}$ as:

$$p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})}. \tag{2.99}$$

Because the denominator of (2.99) is not a function of $\boldsymbol{\theta}_t$, the MAP estimate can be obtained by maximizing the numerator with respect to $\boldsymbol{\theta}_t$:

$$\widehat{\boldsymbol{\theta}}_t^{MAP} = \arg\max_{\theta}\big(p(\mathbf{y}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})\big), \tag{2.100}$$

which is equivalent to choosing $\boldsymbol{\theta}_t$ that minimizes negative logarithm of the numerator in (2.99), i.e.:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_t^{MAP} &= \arg\min_{\theta}\big(-\ln\big(p(\mathbf{y}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})\big)\big)\\ &= \arg\min_{\theta}\big(-\ln\big(p(\mathbf{y}_t|\boldsymbol{\theta}_t)\big) - \ln\big(p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})\big)\big)\\ &= \arg\min_{\theta}\big(J(\boldsymbol{\theta}_t)\big),\end{aligned} \tag{2.101}$$

where

$$J(\boldsymbol{\theta}_t) = -\ln\big(p(\mathbf{y}_t|\boldsymbol{\theta}_t)\big) - \ln\big(p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})\big) \tag{2.102}$$

is called the posterior log-likelihood function. Since we already mentioned that under the MAP interpretation of Kalman filter, all densities are assumed to be Gaussian, we can substitute the probability densities in (2.102), which represent the observation likelihood and prior density of system parameters, respectively, by:

$$p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) = \frac{1}{\sqrt{(2\pi)^{n_\theta}\big|\mathbf{P}_{\theta,t|t-1}\big|}}\exp\left(-\frac{1}{2}\big(\hat{\boldsymbol{\theta}}_{t|t}-\hat{\boldsymbol{\theta}}_{t|t-1}\big)^T\mathbf{P}_{\theta,t|t-1}^{-1}\big(\hat{\boldsymbol{\theta}}_{t|t}-\hat{\boldsymbol{\theta}}_{t|t-1}\big)\right) \tag{2.103}$$

$$p(\mathbf{y}_t|\boldsymbol{\theta}_t) = \frac{1}{\sqrt{(2\pi)^{n_y}|\mathbf{R}_t|}}\exp\left(-\frac{1}{2}\big(\mathbf{y}_t-\hat{\mathbf{y}}_{t|t-1}\big)^T\mathbf{R}_t^{-1}\big(\mathbf{y}_t-\hat{\mathbf{y}}_{t|t-1}\big)\right), \tag{2.104}$$

which results in the following form of the posterior log-likelihood function:

$$J(\boldsymbol{\theta}_t) = \frac{1}{2}\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right)^T \mathbf{R}_t^{-1}\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right) + \frac{1}{2}\left(\hat{\boldsymbol{\theta}}_{t|t} - \hat{\boldsymbol{\theta}}_{t|t-1}\right)^T \mathbf{P}_{\theta,t|t-1}^{-1}\left(\hat{\boldsymbol{\theta}}_{t|t} - \hat{\boldsymbol{\theta}}_{t|t-1}\right) + c. \tag{2.105}$$

Here, $\widehat{\boldsymbol{\theta}}_{t|t-1}$ is the prior estimate of the parameters (i.e. before the new observation is incorporated), $\mathbf{P}_{\theta,t|t-1}$ is its error covariance, both obtained during the time update step of Kalman filter. The constant c accounts for normalizing terms in the Gaussian density functions. The MAP estimate is now found by substituting (2.105) into (2.101) and solving for the minimum value of $\boldsymbol{\theta}_t$. This involves taking the derivative of (2.105) with respect to $\boldsymbol{\theta}_t$ and solving for zero, which results (after some matrix manipulation) in the standard Kalman update equation of the parameter estimate:

$$\widehat{\boldsymbol{\theta}}_{t|t}^{MAP} = \widehat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right), \tag{2.106}$$

with Kalman gain:

$$\mathbf{K}_t = \mathbf{P}_{\theta y,t|t-1}\left(\mathbf{P}_{yy,t|t-1} + \mathbf{R}_t\right)^{-1}. \tag{2.107}$$

Similarly also the update for parameter error covariance can be derived as:

$$\mathbf{P}_{\theta,t|t} = \mathbf{P}_{\theta,t|t-1} - \mathbf{K}_t\left(\mathbf{P}_{yy,t|t-1} + \mathbf{R}_t\right)\mathbf{K}_t^T. \tag{2.108}$$

For a complete derivation, the reader is referred to [68].

Although this brief derivation assumed linear mapping function, which is generally not the case, it was shown in [59] that if we approximate the nonlinear mapping function $\hat{\mathbf{y}}_{t|t-1} = \mathbf{g}(\widehat{\boldsymbol{\theta}}_{t|t-1})$ by statistically linearized form (2.45), as implicitly applied for example by cubature-point approach, the MAP derivation of nonlinear Kalman filter can be carried out (in similar manner) as well. Of course, the success of this approximation strategy is based on the assumption that the statistical linearization of the $\mathbf{g}(.)$, over the probabilistic spread (uncertainty) of the underlying parameter (state) random variable $\boldsymbol{\theta}$, is a good approximation; i.e. that errors are relatively small and normally distributed. This condition is met if the set of sampling (cubature) points is constructed correctly, which is automatically guaranteed by the third-degree cubature integration rule that accurately captures the first two statistical moments even in highly nonlinear models. This implies that the CKF parameter estimation algorithm is

equivalent to a maximum posterior likelihood estimate of the underlying parameters under a Gaussian posterior (and noise distribution) assumption.

Moreover, it is also known that the measurement update step of statistically linearized filters such as CKF or UKF, when applied to sequential parameter estimation, is equivalent to the modified Gauss-Newton method for solving a nonlinear least-squares problem [59, 92, 93]. Specifically, it was shown that the nonlinear least-squares problem and the maximum posterior likelihood problem as defined above optimize the same objective (cost) function. This gives rise to many interesting properties of this sequential parameter estimator. Mainly, it says that the recursively updated parameter (state) estimate of covariance matrix (as calculated by CKF parameter estimation filter) is equivalent to the recursively (online) calculated inverse of the expected Fisher information matrix (FIM) used by the Gauss-Newton equivalent optimization step [94]. In other words, the inverse of FIM at the maximum likelihood estimate is equal to the Cramer-Roe lower bound (CRLB) of the estimated variance. Importantly, the CRLB can be used for evaluating the performance of suboptimal nonlinear filtering methods [95], while the FIM has a large utilization in analysis of parameter space, parameter reduction, and model identifiability [96].

2.4.2 Joint estimation

It is very often the case that both hidden states of dynamic process and model parameters are unknown and have to be inferred from the measured data. Moreover, there might be even a situation, where one wants to estimate also the unknown input into the system. This special case of system identification can be consider as a blind (nonlinear) deconvolution problem, which is described as estimating the unknown input to a dynamic system, given output data, when the model of the system contains unknown parameters. As it was discussed above, the nonlinear cubature Kalman framework is a well suited approach to robust parameter estimation. What we should add now is the fact that because of so-called natural condition of control [55], it is possible to generate the input $\mathbf{z}_t$ using the state prediction $\hat{\mathbf{x}}_{t|t-1}$. In our case it means that if we augment the state vector by the process representing the input, e.g. similar to the one used for parameter estimation (2.98) with input noise $\mathbf{v}_t \sim \mathcal{N}(0, \mathbf{V}_t)$, we can estimate the system input together (jointly) with the states. By saying this, we can go

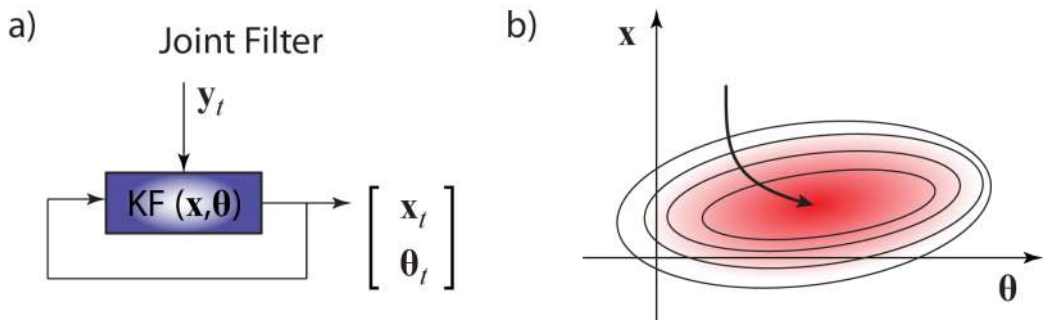


Figure 2.5 Illustration of joint filtering scheme. (a) Both model states and parameters are estimated simultaneously in the augmented form of the state vector. (b) Joint state-parameter optimization space.

one step further (towards blind deconvolution scheme) and attempt to solve a dual estimation problem, where under consideration of a nonlinear dynamic system, the system states $\mathbf{x}_t$, the parameters $\boldsymbol{\theta}_t$ and the input $\mathbf{z}_t$, are estimated simultaneously from the observed noisy signal $\mathbf{y}_t$.

It should be noted that by the input $\mathbf{z}_t$ we mean an endogenous input (or signal), which might be different from the exogenous input $\mathbf{u}_t$. In the context of fMRI, the input $\mathbf{u}_t$ is presented to the subject, whereas $\mathbf{z}_t$ reflects the actual neuronal response, which might (or might not) reflect the exogenous stimulus. In other words, there is always some endogenous activity present in brain even in the absence of any external stimuli, i.e. at rest.

A general theoretical and algorithmic framework for dual Kalman filter based estimation was presented in [68], [59]. This framework encompasses two main approaches, namely joint estimation and dual estimation. In the dual filtering approach, two Kalman filters are run simultaneously (in an iterative fashion) for a state and a parameter estimation. At every time step, the current estimate of the parameters $\boldsymbol{\theta}_t$ is used in the state filter as a given (known) input and likewise, the current estimate of the state $\hat{\mathbf{x}}_t$ is used in the parameter filter. This results in a step-wise optimization within the joint state-parameter space. On the other hand, in the joint filtering approach, the unknown system state and parameters are concatenated into a single higher-dimensional joint state vector, $\boldsymbol{\mu}_t = [\mathbf{x}_t^T, \mathbf{z}_t^T, \boldsymbol{\theta}_t^T]^T$. This results in a smoothed convergence in the joint state-parameter space (see Figure 2.5). There is a prevalent opinion that the performance of joint estimation scheme is superior to dual estimation scheme [59, 68,

97]. Therefore, the joint estimation framework based on cubature Kalman filtering and smoothing is considered in this work.

The state-space model for joint estimation scheme is then formulated as:

$$\boldsymbol{\mu}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \\ \boldsymbol{\theta}_t \end{bmatrix} = \begin{bmatrix} \mathbf{f}(\mathbf{x}_{t-1}, \boldsymbol{\theta}_{t-1}, \mathbf{z}_{t-1}) \\ \mathbf{z}_{t-1} \\ \boldsymbol{\theta}_{t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{q}_{t-1} \\ \mathbf{v}_{t-1} \\ \mathbf{o}_{t-1} \end{bmatrix} \tag{2.109}$$

$$\mathbf{y}_t = \mathbf{h}(\boldsymbol{\mu}_t) + \mathbf{r}_t. \tag{2.110}$$

Since the joint filter concatenates the state and parameter variables into a single state vector, it effectively models the cross-covariances between the state, input and parameters estimates:

$$\mathbf{P}_t = \begin{bmatrix} \mathbf{P}_{x,t} & \mathbf{P}_{xz,t} & \mathbf{P}_{x\theta,t} \\ \mathbf{P}_{zx,t} & \mathbf{P}_{z,t} & \mathbf{P}_{z\theta,t} \\ \mathbf{P}_{\theta x,t} & \mathbf{P}_{\theta z,t} & \mathbf{P}_{\theta,t} \end{bmatrix}. \tag{2.111}$$

This full covariance structure allows the joint estimation framework not only to deal with uncertainty about parameter and state estimates (through the cubature-point approach), but also to model the interaction (conditional dependences) between the states and parameters, which generally provides better estimates [59, 68]. Note that since the parameters are estimated simultaneously with the states, the convergence of parameter estimates depends also on the length of the time series.

Algorithm 5. Joint estimation of states and parameters by CKF.

- *Initialization:*
  - Create augmented state vector, error covariance matrix and process noise covariance matrix:

$$\hat{\boldsymbol{\mu}}_0 = \mathbb{E}[\boldsymbol{\mu}_0] = [\mathbf{x}_0^T, \mathbf{z}_0^T, \boldsymbol{\theta}_0^T]^T \quad \text{and} \quad \mathbf{P}_0 = \mathbb{E}[(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0)(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0)^T],$$

$$\boldsymbol{\mu}_0 = [\mathbf{x}_0^T, \mathbf{z}_0^T, \boldsymbol{\theta}_0^T]^T, \quad \mathbf{P}_0 = \begin{bmatrix} \mathbf{P}_{x,0} & 0 & 0 \\ 0 & \mathbf{P}_{z,0} & 0 \\ 0 & 0 & \mathbf{P}_{\theta,0} \end{bmatrix} \quad \text{and} \quad \mathbf{W}_0 = \begin{bmatrix} \mathbf{Q}_{x,0} & 0 & 0 \\ 0 & \mathbf{V}_{z,0} & 0 \\ 0 & 0 & \mathbf{O}_{\theta,0} \end{bmatrix}$$

- *For* $t = 1, \dots, T$

  Time update:

  1. Calculate the set of cubature points based on previous step estimates; i.e. $\hat{\boldsymbol{\mu}}_{t-1|t-1}$ and $\mathbf{P}_{t-1|t-1}$.

2. Estimate the predicted joint state vector $\hat{\boldsymbol{\mu}}_{t|t-1}$ and predicted error covariance matrix $\mathbf{P}_{t|t-1}$ after propagating cubature points through state equation (2.109).

Measurement update:

1. Calculate the set of cubature points based on time update predictions, $\hat{\boldsymbol{\mu}}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$.
2. Estimate the predicted measurement and $\hat{\mathbf{y}}_{t|t-1}$ the innovation covariance matrix $\mathbf{P}_{yy,t|t-1}$ after propagation cubature points through measurement equation (2.110).
3. Estimate the joint posterior estimates of the mean $\hat{\boldsymbol{\mu}}_{t|t}$ and error covariance matrix $\mathbf{P}_{t|t}$.

Log-likelihood:

4. Calculate the log-likelihood at time step t:

$$\mathcal{L}_t = -\frac{n_y}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{P}_{yy,t|t-1}\right| - \frac{1}{2}\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right)^T \mathbf{P}_{yy,t|t-1}^{-1}\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right) \quad (2.112)$$

- Note: Although this description is provided for non-square-root version of CKS, the same step are followed also for the SCKS. In that case, the square-root covariance matrices are considered. Additionally, the implementation of cubature RTS smoother (or its square-root version) is very straightforward. Simply the augmented (joint) state vector and corresponding error covariance matrix are employed.

2.5 Hybrid continuous-discrete state-space models

In previous sections, we have considered the state-space model to be described in a discrete time, however, in many practical situations, the process equation of state-space model is derived from underlying physics of a continuous dynamic system, and is expressed in the form of a set of differential equations. But still, the measurements $\mathbf{y}_t$ are acquired by digital devices; i.e. they are available at discrete time points ($t = 1,2,\dots,T$). Therefore, we have a model with a continuous process equation and a

discrete measurement equation. The stochastic representation of this state-space model, with additive noise, can be formulated as:

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), t)dt + \sqrt{\mathbf{Q}}d\mathbf{w}(t) \tag{2.113}$$

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, t) + \mathbf{r}_t, \tag{2.114}$$

where $\mathbf{x}(t)$ is the state of the system at time t; $\mathbf{f}(.)$ is a known nonlinear drift function; $\mathbf{w}(t)$ denotes standard Brownian motion that is independent of $\mathbf{x}(t)$; and $\mathbf{Q}$ is a known diffusion matrix. The process equation is the simplest form of Itô's stochastic differential equation [98]. The system is observed through the noisy measurements in discrete time intervals (discrete times are denoted as subscripts).

The recursive Bayesian solution to the above continuous-discrete model is very similar to the one described in Section 2.2. The only difference appears during the time update step of Kalman filter, where the old posterior density is propagated trough the process equation (2.113). In this case, the probability density of the state at time t obeys the *Fokker-Plank equation* (FPE), also called *Kolmogorov's forward equation* [99, 100]:

$$\begin{aligned}\frac{\partial p(\mathbf{x}(t)|\mathbf{y}_{1:t})}{\partial t} = &-\frac{\partial p(\mathbf{x}(t)|\mathbf{y}_{1:t})}{\partial \mathbf{x}(t)}\mathbf{f}(\mathbf{x}(t), t) - p(\mathbf{x}(t)|\mathbf{y}_{1:t})\mathrm{tr}\left(\frac{\partial \mathbf{f}(\mathbf{x}(t), t)}{\partial \mathbf{x}(t)}\right) \\ &+ \frac{1}{2}\mathrm{tr}\left(\mathbf{Q}(t)\frac{\partial^2 p(\mathbf{x}(t)|\mathbf{y}_{1:t})}{\partial^2 \mathbf{x}(t)}\right),\end{aligned} \tag{2.115}$$

which describes the evolution of probability density between the measurement time instants. The exact solution to FPE is available only for linear Gaussian system represented by the time update of Kalman-Bucy filter [101]. In other cases, the FPE has to be approximated. There are two main groups of methods that attempt to approximate the time step of continuous-discrete filter. Methods in the first group compute the conditional density by explicitly solving FPE. These cover many numerical methods such as finite element methods [102, 103], finite difference methods [104], particle methods [105], or simulation approaches using Markov chain Monte Carlo approximation [106]. Main disadvantage of these methods is that their computational complexity increases exponentially with the dimension of the state vector. On the other hand, the second group involves methods, which compute a finite number of summary statistics in terms of conditional moments after discretizing the continuous time process equation using the Euler or higher order Runge-Kutta methods [107]. Continuous-discrete forms of nonlinear Kalman filters that were mentioned in Section 2.2 fall under

this group [64, 108, 109]. From the concept of this work, it is obvious that our interest will be in the approaches that fit into the second group. However, the aim will not be to develop an approximate nonlinear filter by approximating continuous time filter equations (as it is often the case), but rather to use the standard discrete time filtering equations for an approximate discrete time model of the original continuous time dynamical system [56]. This automatically puts high demand on accuracy of the method that discretizes the model represented by stochastic differential equations.

The simplest and the most common method is to apply the Euler approximation to the stochastic differential equation over time interval $[t, t+\delta)$:

$$\mathbf{x}_{t+\delta} = \mathbf{x}_t + \delta \mathbf{f}(\mathbf{x}_t) + \sqrt{\mathbf{Q}}\mathbf{w}_t, \tag{2.116}$$

where $\mathbf{w}_t = (\mathbf{w}(t+\delta) - \mathbf{w}(t))$ is the standard Gaussian noise, $\mathbf{w}_t \sim \mathcal{N}(0, \delta\mathbf{I})$, and is interpreted as the time-derivative of the Brownian motion. This approximate discrete equation (2.116) can be considered as a discrete time counterpart of the Itô's process equation (2.113). If we seek a good discrete time approximation of the continuous stochastic dynamical model for nonlinear and Gaussian multivariate process $\mathbf{x}(t)$, we need to find a discretization scheme that satisfies the following criteria [110]:

- The model should be consistent, i.e.:

$$\frac{\mathbf{x}_{t+\delta} - \mathbf{x}_t}{\delta} \to \mathbf{f}(\mathbf{x}_t) \quad \text{for} \quad \delta \to 0. \tag{2.117}$$

- The trajectory of $\mathbf{x}_t$ should coincide with the true trajectory of $\mathbf{x}(t)$ at the discrete time points $t, t+\delta, t+2\delta, \ldots,$ at least when $\mathbf{f}(.)$ is linear.
- The discrete time model should preserve the qualitative characteristics of the continuous time model; i.e. zero points of $\mathbf{f}(\mathbf{x})$ and the Jacobian are preserved.

Unfortunately, all of very common discretization schemes, such as Euler (2.116) and Runge-Kutta methods, well known in the field of numerical analysis, do not satisfy the second and the third condition (see [56] for the review).

For a discrete model the computationally stable time discretization method is given by [111]:

$$\mathbf{x}_{t+\delta} = \exp(\mathbf{J}_x \delta)\, \mathbf{x}_t, \tag{2.118}$$

where $\exp(\mathbf{J}_x\delta) = \sum_{i=0}^{\infty} \frac{\delta^i}{i!}\mathbf{J}_x^i$ represents a matrix exponential of $\mathbf{J}_x$ being Jacobian matrix of the process $\mathbf{f}(\mathbf{x})$. This approach will be computationally stable and ergodic also for stochastic process:

$$\mathbf{x}_{t+\delta} = \exp(\mathbf{J}_x\delta)\,\mathbf{x}_t + \sqrt{\mathbf{Q}}\mathbf{w}_t, \tag{2.119}$$

but unfortunately it is not consistent in nonlinear models. Nevertheless, this simple method can be considered as a cornerstone for so called local linearization (LL) scheme [56], which satisfies all the above criteria and is also consistent when applied to nonlinear models.

2.5.1 Local linearization scheme

This approach intuitively assumes the nonlinear function $\mathbf{f}(.)$ to be locally approximately linear with respect to the process $\mathbf{x}(t)$. In order to derive this discretization process, the equation (2.113) is first considered in a deterministic context, i.e. only the following differential equation is considered:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}\big(\mathbf{x}(t)\big). \tag{2.120}$$

We assume an appropriate approximation that the process is linear on a small time interval $[t, t+\delta)$, where its Jacobian $\mathbf{J}_x$ is constant. Here $\mathbf{J}_x$ represents the Jacobian of $\mathbf{f}(\mathbf{x})$, which is given by:

$$\mathbf{J}_x = \left.\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_t} \tag{2.121}$$

From this assumption, the following relation for the process $\mathbf{x}(s)$ on the interval $t \le s < t+\delta$ is valid:

$$\frac{d^2\mathbf{x}(s)}{ds^2} = \mathbf{J}_x \frac{d\mathbf{x}(s)}{ds}. \tag{2.122}$$

If this is integrated on the interval $[t, t+\tau)$, where $0 \le \tau < \delta$, one obtains:

$$\begin{aligned} \frac{d\mathbf{x}(t+\tau)}{dt} &= \exp(\mathbf{J}_x\tau)\,\frac{d\mathbf{x}(t)}{dt} \\ &= \exp(\mathbf{J}_x\tau)\,\mathbf{f}\big(\mathbf{x}(t)\big). \end{aligned} \tag{2.123}$$

By integrating this again with respect to τ on $[0, \delta)$, it results in the following relation between $\mathbf{x}_t$ and $\mathbf{x}_{t+\delta}$:

$$\mathbf{x}_{t+\delta} = \mathbf{x}_t + \mathbf{J}_x^{-1}[\exp(\mathbf{J}_x\delta) - \mathbf{I}]\mathbf{f}(\mathbf{x}_t). \tag{2.124}$$

This represent the local linearization equation corresponding to the first order linear approximation [56]. Additionally, it is possible to apply the LL scheme also to the random term of the state equation in (2.113) that follows multivariate normal distribution with zero mean vector and covariance matrix [12]:

$$\mathbf{Q}_{t+\delta} = \int_t^{t+\delta} \exp(\mathbf{J}_x\delta)\sqrt{\mathbf{Q}_t}\sqrt{\mathbf{Q}_t}^T \exp(\mathbf{J}_x\delta)^T \, dt\,. \tag{2.125}$$

Finally, there exists a simple algebraic expression [43, 112] that is especially useful for both deterministic and random part of (2.116), since the form (2.124) is not reliable in the case of ill-conditioned matrix $\mathbf{J}_x$. In order to do that, it is necessary to first rewrite (2.113) in the following (but equivalent) form:

$$\mathbf{x}_{t+\delta} = \mathbf{x}_t + \mathbf{l}(\mathbf{x}_t, \delta) + \sqrt{\mathbf{Q}}\mathbf{w}_t, \tag{2.126}$$

then the $\mathbf{l}(\mathbf{x}_t, \delta)$ is defined in the block matrix as:

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_t, \delta) & \mathbf{l}(\mathbf{x}_t, \delta) \\ 0 & 1 \end{bmatrix} = \exp(\mathbf{C}\delta), \tag{2.127}$$

where $\mathbf{L}(\mathbf{x}_t, \delta) = \exp(\mathbf{J}_x\delta)$ and

$$\mathbf{C} = \begin{bmatrix} \mathbf{J}_x & \mathbf{f}(\mathbf{x}_t) \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(n_x+1)\times(n_x+1)}. \tag{2.128}$$

Similarly, the covariance matrix of Gaussian process noise is obtained through [113]:

$$\begin{bmatrix} \mathbf{E} & \mathbf{G} \\ 0 & \mathbf{F} \end{bmatrix} = \exp(\mathbf{D}\delta), \tag{2.129}$$

with

$$\mathbf{D} = \begin{bmatrix} -\mathbf{J}_x & \sqrt{\mathbf{Q}_t}\sqrt{\mathbf{Q}_t}^T \\ 0 & \mathbf{J}_x^T \end{bmatrix} \in \mathbb{R}^{(2n_x)\times(2n_x)} \tag{2.130}$$

resulting in a process noise covariance matrix:

$$\mathbf{Q}_{t+\delta} = \mathbf{F}^T\mathbf{G}. \tag{2.131}$$

2.5.2 Square-root cubature Kalman filter with local linearization

In the previous section we have discussed the theoretical background of local linearization scheme as a method for discretization of stochastic differential equations. Now, we apply this scheme to the time-update step of nonlinear Kalman filter. Assuming a Gaussian approximation of the filtering distribution, we can obtain the predicted mean and the error covariance through the local linearization as:

$$\hat{\mathbf{x}}_{t-1+\delta|t-1} \approx \int_{\mathbb{R}^{n_x}} [\mathbf{x}_{t-1} + \mathbf{l}(\mathbf{x}_{t-1}, \delta)] \times \mathcal{N}\left(\mathbf{x}_{t-1} | \hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}\right) d\mathbf{x} \tag{2.132}$$

$$\begin{aligned} \mathbf{P}_{t-1+\delta|t-1} \approx \int_{\mathbb{R}^{n_x}} &\left([\mathbf{x}_{t-1} + \mathbf{l}(\mathbf{x}_{t-1}, \delta)] - \hat{\mathbf{x}}_{t-1+\delta|t-1}\right)\left([\mathbf{x}_{t-1} + \mathbf{l}(\mathbf{x}_{t-1}, \delta)]\right. \\ &\left. - \hat{\mathbf{x}}_{t-1+\delta|t-1}\right)^T \mathcal{N}\left(\mathbf{x}_{t-1} | \hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}\right) d\mathbf{x} + \mathbf{Q}_{t-1+\delta}, \end{aligned} \tag{2.133}$$

The next step is to numerically compute the integrals above, which can be done using the cubature integration rules.

In order to compute the predicted state estimate more accurately before receiving the measurement at time step t, several integration steps are usually performed within the time interval $[t-1, t]$; i.e. interval between available measurement samples. In this sense we will partially follow the cost-reduced approach introduced for continuous discrete systems [109].

Given the set of cubature points representing the posterior at time $t-1$, we propagate these points recursively through the noise-free nonlinear process model (2.124) up to K steps forward, where the size of intermediate integration step is $\delta = 1/K$. Similarly, we update also the process noise covariance according to (2.125). Subsequently, we compute the predicted state and error covariance matrix by using the set of cubature points from the last step K, and continue with the standard measurement update of CKF. It means that the predicted mean and covariance are computed only once per the time update step of CKF. This is summarized in Algorithm 6.

Notice that LL scheme involves calculation of Jacobian matrix. Once we combine LL with cubature integration rules, we have to compute this Jacobian matrix for each cubature point. In addition, we also have to compute matrix exponential. This represents a serious demand on computational cost. However, this cost is paid in favor of high accuracy and stability of the filter.

Algorithm 6. Time update step for continuous-discrete SCKF.

- *Start the time update step:*

  1. Evaluate the cubature points ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}_{i,t-1|t-1} = \mathbf{S}_{t-1|t-1}\boldsymbol{\xi}_i + \hat{\mathbf{x}}_{t-1|t-1} \tag{2.134}$$

- *Accumulate the state vector and the noise covariance over the intermediate steps ($k = 1, \dots, K$) within measurement interval $[t-1, t]$, with $\delta = 1/K$:*

  3. Propagate the cubature points through the process model ($i = 1, \dots, m$):

$$\boldsymbol{\mathcal{X}}^*_{i,t-1+k\delta|t-1} = \boldsymbol{\mathcal{X}}_{i,t-1+(k-1)\delta|t-1} + \mathbf{l}\left(\boldsymbol{\mathcal{X}}_{i,t-1+(k-1)\delta|t-1}, \delta\right) \tag{2.135}$$

  4. Based on the average Jacobian $\mathbf{J}_x = \frac{1}{m}\sum_i^m \mathbf{J}_{i,x}$, compute the process noise covariance matrix according to (2.129)-(2.131):

$$\mathbf{Q}_{t-1+k\delta} = \begin{cases} \mathbf{g}\left(\mathbf{Q}_{t-1+(k-1)\delta}, \delta\right) & for\ k = 1 \\ \mathbf{Q}_{t-1+(k-1)\delta} + \mathbf{g}\left(\mathbf{Q}_{t-1+(k-1)\delta}, \delta\right) & for\ k > 1 \end{cases} \tag{2.136}$$

- *If $k = K$ continue to evaluate the prediction:*

  5. Estimate the predicted state:

$$\hat{\mathbf{x}}_{t|t-1} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{X}}^*_{i,t-1+K\delta|t-1} \tag{2.137}$$

  6. Estimate the predicted square-root error covariance matrix:

$$\mathbf{S}_{t|t-1} = \mathbf{Tria}\left(\left[\mathbf{X}_{t|t-1}, \sqrt{\mathbf{Q}_{t-1+K\delta}}\right]\right) \tag{2.138}$$

  with weighted and centered matrix:

$$\mathbf{X}_{t|t-1} = \frac{1}{\sqrt{m}}\left[\boldsymbol{\mathcal{X}}^*_{1,t-1+K\delta|t-1} - \hat{\mathbf{x}}_{t|t-1}, \dots, \boldsymbol{\mathcal{X}}^*_{m,t-1+K\delta|t-1} - \hat{\mathbf{x}}_{t|t-1}\right] \tag{2.139}$$

- *Continue with the regular discrete time measurement update step.*

Although this cost-reduced formulation of CKF time update represents a stable algorithm to estimate the hidden states of continuous model, it seems to be not well suited for the joint estimation of both states and parameters. Specifically, if we use this scheme for joint estimation, we propagate the hidden states *K* times forward between observations, while the parameters (having discrete model) are propagated only once. This causes some instability into the performance of the joint filter, which is not desirable. Therefore, we suggest a slightly different strategy. Rather than performing

several integration steps through the process model between two observation samples, we linearly interpolate the observation sequence, so that with each integration step we have a new measurement available (now $K = \delta$), hence both state and parameter predictions are obtained at each step. This further improves the estimation of parameters, because we simply have more time points to achieve the convergence. In addition, this formulation is also well suited for a backward propagation through an RTS smoother.

Finally, we compared these two formulations of filter time update with a recently introduced approach that uses Itô-Taylor expansion of the order 1.5 to discretize the continuous system, and which was designed directly for CKF [109]. For details see Appendix A.3.

2.6 Adaptive estimation of noise statistics

The Kalman formulation of filtering problem assumes complete *a priori* knowledge of the process and measurement noise statistics. In most practical situations, these noise statistics are unknown or not known perfectly. When incorrect prior statistics are used to implement sequential filtering algorithm, it might result in suboptimal performance and possibly in filter divergence. Therefore, in the lack of system statistics knowledge, it is desirable to adaptively estimate the process noise and measurement noise statistics simultaneously with the system state. The adaptive filtering methods can be classified into four main categories: Bayesian estimation, maximum likelihood estimation, correlation methods and covariance matching methods. A brief surveys on these different approaches can be found in [100, 114-116]. Note that there are many other methods that perform offline estimation of noise statistics [117], including EM approach [90, 91]. However, since we already have the preference to perform joint estimation of the states and parameters, it is a logical choice to adaptively estimate the noise statistics as well. Also, some of the offline estimators assume multiple repetitions or very long recordings of measured data [117], which are not really available for our problem.

In standard Kalman filtering framework, all noise statistics are described by the first two statistic moments, i.e. by the mean and the covariance, where the mean is usually assumed to be zero. Therefore, the goal of adaptive filtering is to estimate the

covariances of process noise and measurement noise. It is important to mention that the estimated noise covariance can have a significant influence on the performance of the algorithm. By increasing the process noise covariance, the Kalman gain also increases, thereby producing bigger changes in the state updates. It means that more importance is placed on the most recent measurements. Additionally, as the Kalman gain increases, the process error covariance also increases and filter become less immune to noise and outliers. On the other hand, if the measurement noise covariance increases, the actual measurement is trusted less and vice versa [79].

In the case of parameter estimation, the amount of oscillations in the model prediction clearly depends on the value of the (parameter) process noise covariance. As a result, this covariance can be used as a regularization mechanism to control the smoothness of the prediction. A very common practice in neural networks is to approximate the parameter noise covariance by using simple annealing [88, 118]:

- *Annealing:* The process noise is injected artificially based on process error covariance matrix:

$$\mathbf{O}_t = (\lambda_\theta^{-1} - 1)\mathbf{P}_{t,\theta} \tag{2.140}$$

  where $\lambda_\theta \in (0,1]$ is called the forgetting factor [118]. This allows to put exponentially decaying weigh on past data. Typically, the choice of λ_θ being slightly less than unity works well for many problem.

Another already more sophisticated approach to recursive estimation of noise covariance is the Robbins-Monro (RM) stochastic approximation [111]:

- *Robbins-Monro stochastic approximation* [58]*:*

$$\mathbf{O}_t = \lambda_\theta \mathbf{O}_{t-1} + (\lambda_\theta^{-1} - 1)\mathbf{K}_t[\mathbf{y}_t - \mathbf{g}(\mathbf{x}_t; \widehat{\boldsymbol{\theta}}_t)][\mathbf{y}_t - \mathbf{g}(\mathbf{x}_t; \widehat{\boldsymbol{\theta}}_t)]^T \mathbf{K}_t^T \tag{2.141}$$

  The method assumes that the covariance of the Kalman update model should be consistent with the actual update model. Typically, $\mathbf{O}_t$ is also constrained to be a diagonal matrix, which implies an independence assumption on the parameters. Here, λ_θ has the same function as in the previous case. It is known that an RM approximation provides a very fast convergence and a low final MMSE [59].

Both RM and annealing help to escape poor local minima of the error surface. We made a choice to apply the RM approximation to the recursive estimation of parameter noise covariance $\mathbf{O}_t$ and also to approximate the process noise covariance matrix $\mathbf{Q}_t$, since it proved to have better convergence properties [59]. We should also note that we do not expect to estimate the exact process noise covariance of the dynamic model with RM approach. The aim here is to maintain some artificial level of randomness, which supports the convergence of the algorithm, and prevents the filter from becoming overconfident with the estimate (i.e. it avoids overfitting). By saying this, we consider the noise covariance of the input to be fixed. In this case, any attempt to adaptively estimate the input noise covariance led to the quick divergence of the filter. Nevertheless, as we will see in Chapter 4, there is no need to adaptively estimate the input noise covariance, because it can be easily defined *a priori*; i.e. proportional to the average peak to peak amplitude variation of the observations.

The most important part in adaptive filtering is accurate estimation of measurement noise statistics. For this particular task, we consider a recently introduced Bayesian approach [119] to recursive estimation of measurement noise covariance, which is suitable also to nonlinear filtering and is able to take the advantage of assumed Gaussian density filter such as CKF. In the next section, we describe the details of this approach.

2.6.1 Variational Bayesian estimation of measurement noise

The Bayesian estimation of the noise statistic is the most general one and the other approaches can often be interpreted as an approximation to the Bayesian approach. One way how to perform tractable approximation to posterior inference in Bayesian sense is to apply variational Bayesian (VB) approach [114, 116].

If the measurement noise covariance $\mathbf{R}_t$ is unknown then the goal of Bayesian filtering is to compute the joint posterior distribution $p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t})$ of the state $\mathbf{x}_t$ and of the covariance $\mathbf{R}_t$. We know already from the earlier discussion that the posterior is given by a product of observation likelihood and predictive distribution, which now takes a form:

$$p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t}) \approx p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{R}_t)p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t-1}). \tag{2.142}$$

In [119] they assumed that the probability distribution of noise variance is inverse-Gamma and so the filtering distribution at time $t-1$, and predictive distribution at time t can be approximated as a product of Gaussian and independent inverse-Gamma distributions, respectively [116]:

$$p(\mathbf{x}_{t-1}, \mathbf{R}_{t-1}|\mathbf{y}_{1:t-1}) = \mathcal{N}\left(\mathbf{x}_{t-1}|\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}\right) \times \prod_{i=1}^{d} iG\left(R_{i,\mathrm{t}-1|t-1}|\alpha_{i,\mathrm{t}-1|t-1}, \beta_{i,\mathrm{t}-1|t-1}\right) \tag{2.143}$$

$$p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t-1}) = \mathcal{N}\left(\mathbf{x}_t|\hat{\mathbf{x}}_{t|t-1}, \mathbf{P}_{t|t-1}\right) \times \prod_{i=1}^{d} iG\left(R_{i,t|t-1}|\alpha_{i,t|t-1}, \beta_{i,t|t-1}\right), \tag{2.144}$$

where the dynamic model for the measurement noise covariance maps an inverse-Gamma distribution at the previous step into inverse-Gamma distribution at current step.

It is further assumed that the measurement noise covariance has a form of diagonal matrix $\mathbf{R}_t = \mathrm{diag}(R_{1,t}, \dots, R_{d,t})$. The inverse-Gamma distribution is chosen because it represents the conjugate prior distribution for variance of Gaussian distribution (another common choice is to use inverse-Wishart distribution [120][[121]). These normal and inverse-Gamma densities in (2.144) are parameterized as follows:

$$\mathcal{N}\left(\mathbf{x}_t|\hat{\mathbf{x}}_{t|t-1}, \mathbf{P}_{t|t-1}\right) \propto \left|P_{t|t-1}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\hat{\mathbf{x}}_{t|t} - \hat{\mathbf{x}}_{t|t-1}\right)^T \mathbf{P}_{t|t-1}^{-1}\left(\hat{\mathbf{x}}_{t|t} - \hat{\mathbf{x}}_{t|t-1}\right)\right) \tag{2.145}$$

$$iG\left(R\middle|\alpha_{i,t}, \beta_{i,t}\right) \propto R^{-\alpha-1} \exp\left(-\frac{\beta}{R}\right), \tag{2.146}$$

where we skipped the constants not containing $\mathbf{x}$ or R.

At this point, in order to make the computation of posterior (2.142) tractable, it is possible to apply the VB approach. The VB usually applies the mean field approximation [75] that factorizes the posterior distribution as follows:

$$p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t}) \approx Q_x(\mathbf{x}_t) Q_R(\mathbf{R}_t), \tag{2.147}$$

where $Q_x(\mathbf{x}_t)$ and $Q_R(\mathbf{R}_t)$ are in our case the approximations of normal and inverse-Gamma densities, respectively. The VB approximation can now be formed by minimizing the Kullback-Leibler (KL) divergence between the separable approximations and the true posterior distribution:

$$\mathrm{KL}\big(Q_x(\mathbf{x}_t)Q_R(\mathbf{R}_t), p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t})\big) = \int Q_x(\mathbf{x}_t)Q_R(\mathbf{R}_t) \ln\left(\frac{Q_x(\mathbf{x}_t)Q_R(\mathbf{R}_t)}{p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t})}\right) d\mathbf{x}_t d\mathbf{R}_t. \tag{2.148}$$

By minimizing the KL divergence with respect to probability densities $Q_x(\mathbf{x}_t)$ and $Q_R(\mathbf{R}_t)$ in turn, while keeping the other one fixed, we get the following forms:

$$Q_x(\mathbf{x}_t) \propto \exp\left(\int \ln p(\mathbf{y}_t, \mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t-1})Q_R(\mathbf{R}_t)d\mathbf{R}_t\right) \tag{2.149}$$

$$Q_R(\mathbf{R}_t) \propto \exp\left(\int \ln p(\mathbf{y}_t, \mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t-1})Q_x(\mathbf{x}_t)d\mathbf{x}_t\right). \tag{2.150}$$

The integrals in the exponentials of (2.149) and (2.150) can be expanded as follows:

$$\begin{aligned}\int \ln p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t-1})Q_R(\mathbf{R}_t)d\mathbf{R}_t \\ = \frac{1}{2}\big(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\big)^T\langle\mathbf{R}_t^{-1}\rangle_R\big(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\big) \\ + \frac{1}{2}\big(\hat{\mathbf{x}}_{t|t} - \hat{\mathbf{x}}_{t|t-1}\big)^T\mathbf{P}_{t|t-1}^{-1}\big(\hat{\mathbf{x}}_{t|t} - \hat{\mathbf{x}}_{t|t-1}\big) + c_1\end{aligned} \tag{2.151}$$

$$\begin{aligned}\int \ln p(\mathbf{x}_t, \mathbf{R}_t|\mathbf{y}_{1:t-1})Q_x(\mathbf{x}_t)d\mathbf{x}_t \\ = -\sum_{i=1}^{d}\left[\left(\frac{3}{2} + \alpha_{i,t}\right)\ln\big(R_{i,t}\big) - \frac{\beta_{i,t}}{R_{i,t}}\right] \\ - \frac{1}{2}\sum_{i=1}^{d}R_{i,t}^{-1}\langle\big(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\big)_i^2\rangle_x + c_2,\end{aligned} \tag{2.152}$$

The result of the first integral can be immediately recognized as a form of posterior log-likelihood function (described in Section 2.4.1), which once solved for minimum value of $\mathbf{x}_t$ gives the Kalman update equations for the states (assuming cubature approximation). The expectation $\langle\cdot\rangle_R = \int(\cdot)Q_R(\mathbf{R}_t)d\mathbf{R}_t$ represents the estimate of measurement noise covariance matrix:

$$\langle\mathbf{R}_t^{-1}\rangle_R = \mathrm{diag}\big(\alpha_{1,t|t}/\beta_{1,t|t}, \dots, \alpha_{d,t|t}/\beta_{d,t|t}\big). \tag{2.153}$$

The update of parameters $\alpha_{i,t|t}$ and $\beta_{i,t|t}$ describing inverse-Gamma distribution are derived from the result of the second integral (2.152), where the approximate solution to expectation $\langle\cdot\rangle_x = \int(\cdot)Q_x(\mathbf{x}_t)d\mathbf{x}_t$ can be separately computed by Gaussian integration methods such as cubature rules:

$$\alpha_{t|t} = \alpha_{t|t-1} + \frac{1}{2} \tag{2.154}$$

$$\boldsymbol{\beta}_{t|t} = \boldsymbol{\beta}_{t|t-1} + \frac{1}{2}\int_{n_d} \left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right)\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right)^T \mathcal{N}\left(\mathbf{x}_t | \hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right). \tag{2.155}$$

Here we used simplification $\boldsymbol{\beta}_{t|t-1} = \mathrm{diag}\left(\beta_{1,t|t-1}, \dots, \beta_{d,t|t-1}\right)$ and α_{t-1} represents scalar that is the same for all measurements. In [119] they also suggested to update $\alpha_{t|t}$ and $\boldsymbol{\beta}_{t|t}$ iteratively to improve the estimation performance of the algorithm.

So far, we have described only the update of parameters α_t and $\boldsymbol{\beta}_t$. The dynamic model, which takes a part during the prediction step has to satisfy a condition that when this model is applied to inverse-Gamma distribution, it produces another inverse-Gamma distribution. This can be simply achieved:

$$\alpha_{t|t-1} = \rho\alpha_{t-1|t-1} \tag{2.156}$$

$$\boldsymbol{\beta}_{t|t-1} = \rho\boldsymbol{\beta}_{t-1|t-1}. \tag{2.157}$$

where $\rho \in (0,1]$ controls the assumed dynamics of the measurement noise, i.e. it is possible to model also nonstationary noise covariance. If $\rho = 1$ then we assume stationary covariance and $\rho < 1$ allow for higher time-fluctuations [119]. The summary of the resulting adaptive algorithm in the sense of CKF is presented bellow.

Algorithm 7. CKF with adaptive estimation of measurement noise covariance.

- *Initialization:*
  - Standard initialization as for CKF.
  - Initialization of inverse-Gamma distribution parameters; e.g. $\alpha_0 = 1$, $\boldsymbol{\beta}_0 = \mathbf{I}_d$, and ρ = 0.998.
- *For* $t = 1, \dots, T$

  Time update:

  1. Perform regular step of CKF time update, i.e. steps 1.-5. of Algorithm 1, to obtain predicted mean and predicted error covariance.
  2. Calculate parameters of predicted inverse-Gamma distribution:

$$\alpha_{t|t-1} = \rho\alpha_{t-1|t-1} \tag{2.158}$$

$$\boldsymbol{\beta}_{t|t-1} = \rho\boldsymbol{\beta}_{t-1|t-1}. \tag{2.159}$$

Measurement update:

3. Perform steps 6.-9. and 11. of Algorithm 1. to obtain propagated cubature point array $\boldsymbol{\mathcal{Y}}_{i,t|t-1}$ with associated mean response prediction $\hat{\mathbf{y}}_{t|t-1}$, and the cross-covariance matrix $\mathbf{P}_{xy,t|t-1}$.

- *Set* $\boldsymbol{\beta}_{t|t}^{0} = \boldsymbol{\beta}_{t|t-1}$ *and* $\alpha_{t|t} = \alpha_{t|t-1} + 1$, *and iterate the following steps K-times* $(k = 1, .., K)$:

4. Compute the measurement covariance matrix (its diagonal elements):

$$\mathbf{R}_t^k = \mathrm{diag}\left(\beta_{1,t|t}^{k-1}/\alpha_{1,t|t}, \dots, \beta_{d,t|t}^{k-1}/\alpha_{d,t|t}\right) \tag{2.160}$$

5. Estimate the innovation covariance matrix:

$$\mathbf{P}_{yy,t|t-1}^{k} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{Y}}_{i,t|t-1}\boldsymbol{\mathcal{Y}}_{i,t|t-1}^{T} - \hat{\mathbf{y}}_{t|t-1}\hat{\mathbf{y}}_{t|t-1}^{T} + \mathbf{R}_t^k \tag{2.161}$$

6. Estimate the Kalman gain:

$$\mathbf{K}_t = \mathbf{P}_{xy,t|t-1}\left[\mathbf{P}_{yy,t|t-1}^{k}\right]^{-1} \tag{2.162}$$

7. Estimate the updated state:

$$\hat{\mathbf{x}}_{t|t}^{k} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right) \tag{2.163}$$

8. Estimate the updated error covariance:

$$\mathbf{P}_{t|t}^{k} = \mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{P}_{yy,t|t-1}^{k}\mathbf{K}^{T} \tag{2.164}$$

9. Factorize the state error covariance matrix:

$$\mathbf{P}_{t|t}^{k} = \mathbf{S}_{t|t}^{k}\left[\mathbf{S}_{t|t}^{k}\right]^{T} \tag{2.165}$$

10. Evaluate the cubature points $(\mathrm{i} = 1, \dots, m)$:

$$\boldsymbol{\mathcal{X}}_{i,t|t} = \mathbf{S}_{t|t}^{k}\boldsymbol{\xi}_i + \hat{\mathbf{x}}_{t|t}^{k} \tag{2.166}$$

11. Propagate the cubature points through measurement model $(i = 1, \dots, m)$:

$$\boldsymbol{\mathcal{Y}}_{i,t|t}^{*} = \mathbf{h}\left(\boldsymbol{\mathcal{X}}_{i,t|t}, \mathbf{u}_t; \boldsymbol{\theta}\right) \tag{2.167}$$

12. Update beta parameter of inverse-Gamma distribution:

$$\boldsymbol{\beta}_{t|t}^{k} = \boldsymbol{\beta}_{t|t-1} + \mathbf{D}_{t|t}\mathbf{D}_{t|t}^{T} \tag{2.168}$$

with centered and weighted matrix:

$$\mathbf{D}_{t|t} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mathcal{Y}}^{*}_{i,t|t}\, \boldsymbol{\mathcal{Y}}^{*T}_{i,t|t} - \mathbf{y}_t \mathbf{y}_t^T \tag{2.169}$$

- *If* $k = K$, *set* $\beta_{t|t} = \beta^K_{t|t}$, $\hat{x}_{t|t} = \hat{x}^K_{t|t}$, $P_{t|t} = P^K_{t|t}$, *and continue with the next time step.*

- Note that similarly we can apply the VB estimation of the measurement noise covariance also to the square-root version of CKF. In that case, we use equations for SCKF and the iterative update of beta parameters will remain the same. The only difference will be in step 4., where we additionally calculate a square-root of diagonal elements of $\mathbf{R}^k_t$ to obtain $\mathbf{S}^k_{r,t}$ [120].

2.7 Algorithm for estimation of neuronal signal

In this section, we will put together all the items that were described so far in this chapter and will finally build the complete algorithm for estimation of neuronal signal from fMRI data. In this sense, we are introducing an algorithm that is able to solve a triple estimation problem, i.e. we jointly estimate not only the model states (including the endogenous input), the model parameters, but also hyperparameters that represent noise statistics.

Algorithm 8. Estimation of neuronal signal by CKS.

- *Initialization.*
- *Iterate the following* ($i = 1, \ldots, I$)*:*
  - Forward run of CKF ($t = 1, \ldots, T$):
    1. Time update of joint CKF including discretization of continuous part of process model by LL scheme:
       - Obtain predicted joint state vector $\hat{\boldsymbol{\mu}}_{t|t-1}$ and error covariance matrix $\mathbf{P}_{t|t-1}$.
    2. Measurement update (iterated K-times):
       - Obtain joint posterior estimate of state vector $\hat{\boldsymbol{\mu}}_{t|t}$ and error covariance matrix $\mathbf{P}_{t|t}$ at k-step.

- Update the estimate of measurement noise covariance matrix $\mathbf{R}_t$ through VB approach at k-step.

2. Estimate the state and parameter process noise covariance matrices, $\mathbf{Q}_t$ and $\mathbf{O}_t$, respectively, through Robbins-Monro approximation scheme.
3. Calculate the log-likelihood $\mathcal{L}_{t,i}$ at time step t.

o Backward run of cubature RTS smoother ($t = T, \dots, 0$) using the stored posterior estimates and predictions of forward run:

1. Obtain the smoothed posterior estimates of $\hat{\boldsymbol{\mu}}_{t|T}^{S}$ and $\mathbf{P}_{t|T}^{S}$

o Calculate the accumulated log-likelihood $\mathcal{L}_i = \sum_{t=1}^{T} \mathcal{L}_{t,i}$ and its difference with $\mathcal{L}_{i-1}$. If $(\mathcal{L}_i - \mathcal{L}_{i-1}) > 0$, set $\hat{\boldsymbol{\mu}}_0 = \left[\left[\hat{\mathbf{x}}_{1|T}^{S}\right]^T, \left[\hat{\mathbf{z}}_{1|T}^{S}\right]^T, \left[\sum_{t=1}^{T} \hat{\boldsymbol{\theta}}_t\right]^T\right]$ and $\mathbf{P}_0 = \mathbf{P}_{T|T}$, and proceed to the next iteration $(i + 1)$.

- *If* $(\mathcal{L}_i - \mathcal{L}_{i-1}) < threshold$ *or* $(i = I)$ *terminate the algorithm.*
- Note: Although this description is provided for non-square-root version of CKS, the same step are followed also for the SCKS. In that case, the square-root covariance matrices are considered.

2.8 Chapter summary

In this chapter we have introduced a complete methodological framework for the inversion problem consisting of the estimation of the neuronal signal from noisy fMRI data using the forward model of the BOLD signal [29]. The concept draws on the theory of recursive Bayesian filtering and smoothing, which allows one to perform probabilistic inference about the hidden variables (states and parameters) given noisy or incomplete observations. Since the optimal recursive solution for filtering and smoothing problem is tractable only for linear, Gaussian systems and the hemodynamic model describing a coupling between the neuronal signal and observed BOLD signal is nonlinear, we have pursued an approximate solution by using a Gaussian approximate method to model the probability densities.

In particular, we have proposed an approach based on the recently introduced cubature Kalman filtering [55] to the joint estimation problem where the hidden states and parameters are concatenated into a single joint state vector and estimated

simultaneously, yielding joint MAP estimates. This form allows us not to only accurately treat the dual uncertainty of the parameter and state estimates (by using cubature point approach), but also to accurately model the interaction (conditional dependence) between the states and parameters. To obtain more accurate estimates of hemodynamic states and mainly of neuronal signal, we have employed forward-backward smoothing, encompassing also the cubature points formulation of Rauch-Tung-Striebel smoother. The overall estimator performance is further enhanced by considering a square-root formulation that ensures a numerical stability during the recursion.

Next, because the states of hemodynamic model are represented by ordinary differential equations, we have introduced a novel continuous-discrete time representation of CKF that combines a statistical linearization with the local linearization approach for accurate and stable discretization of the process model. Additionally, this new algorithm is also suitable for joint estimation, where the states are propagated through the continuous model and the parameters are propagated through the discrete model.

The estimation framework would not be complete if we were restricted to the informed model inversion, where one assumes the noise statistic to be known, which is typically not the case. Therefore, we have proposed an adaptive estimation of state and parameter noise covariance matrix based on Robbins-Monro stochastic approximation scheme, and further adopted the recently introduced variational Bayesian approach for estimation of measurement noise variance [119] to the cubature point Kalman filter.

All these developments and extensions were combined to create an iterative optimization method, which maximizes the log-likelihood with each iteration and achieves a fast convergence. As a result, we have obtained a novel advanced approach to the estimation of the neuronal signal from the observed BOLD signal superior to what has been so far introduced in the field of neuroscience (comparison is provided in [20]).

This method and results have been presented at [120, 122], and published in [20], where also the comparison with other recently introduced approach is provided. Besides a general development of the methods, the Chapter 2 contains also detailed description

of all respective algorithms. Additional demonstration of the method performance is provided in Chapter 4.

Chapter 3

Modeling brain network connectivity

3.1 Introduction

In this chapter, we focus on modeling coupling among different brain regions (nodes) in terms of effective connectivity. In particular, we will introduce a direct generalization of the estimation framework described in the previous chapter to a multivariate case, where the main goal will be to infer the directional influence among different brain regions at the neuronal level. Before we do so, it will be useful to provide a short overview and motivation on methods that attempt to assess effective connectivity.

In effective connectivity, the neuronal states describe the activity of set of nodes that comprise a graph. The aim of analysis is to identify the directional (causal) influence of activated links in the graph. Importantly, these nodes are in fMRI defined as neural populations at macroscopic level, i.e. whole brain areas, whose activity is summarized by a time varying state vector.

In general, there are two streams of statistical causal modeling: one based on Bayesian dependency graphs or graphical models called structural causal modeling and the other based on causal influence over time [1]. Structural causal modeling is related to structural equation modeling (SEM) [123, 124] and uses graphical models in which direct causal links are encoded by directed edges. However, this approach has two limitations. First, it is restricted to discovering conditional independencies in directed acyclic graphs (DAG), i.e. it cannot deal with (cyclic) feedback loops (see Figure 3.1). This is a serious drawback because the brain works as a directed cyclic graph, where every brain region is connected reciprocally (at least polysynaptically) [15]. Second, the

estimation is completely based on the sample covariance matrix, i.e. it ignores time dynamics. Fortunately, the DAG restriction can be finessed by considering dynamics and temporal precedence within structural causal modeling. This is because the arrow of time can be used to convert a directed acyclic graph into a cyclic graph when the nodes are deployed over successive time points. This leads to SEM with time-lagged data described by autoregressive (AR) models, which are the ground for Granger causality modeling (GCM). The GCM approach is based on temporal precedence, i.e. *A* causes *B* if one reduces uncertainty about the future of *B* given the past of *A*. It is formulated in discrete time analysis framework, where the directionality is usually inferred directly from measured signals. Although GCM has become quite popular in the neuroimaging community during the last several years, there is an ongoing discussion to determine if the concept of temporal precedence is suitable for application to fMRI time series analysis. Main concerns are that GCM does not account for variability in hemodynamic response function across different brain regions [9, 125]; the measurement noise can reverse the estimation of causality direction [126]; and the coupling strengths are parameterized in terms of regression coefficients, which are not the true coupling parameters of effective connectivity. Additionally, the reliability of GCM degrades with the increase of sampling interval [127], which is important for fMRI because the sampling interval is quite large with respect to the time scale of neuronal events. Of course, there have been several attempts to partially eliminate some of these obstacles. For example, one can take the regression coefficients out of the picture of noisy measurement by mapping from observed data to hidden states (represented by AR coefficients) [128], or go even further and consider a generative model that maps the observations to neuronal states through the hemodynamic model and then use the estimated neuronal states for determining effective connectivity [129]. However, the later is suboptimal because it assumes that causal modeling between neuronal states is conditionally independent of the mapping from measured data [130].

As a result of these recent discussions [131-134], it is now clear that discovering effective connectivity should be based on state-space models of controllable (causal in the control theory sense) biophysical processes that have hidden neuronal states and possibly exogenous input [135]. Further, the optimal statistical procedure is to invert the complete generative model described by a set of state equations that quantify how the observed data are affected by the presence of causal links [130]. This possibly allows to

accommodate the conditional dependencies between parameters of the state equations, which are mapped to the observations [15]. If we now recall the definition of the effective connectivity as stated in Chapter 1, i.e. that effective connectivity refers to the influence that one neural system exerts over another, either at synaptic or population level, one realizes that the procedure mentioned above is a sensible choice and, at the moment, probably also the only choice suitable for application to fMRI data.

This reasoning has led to the development of dynamic causal modeling (DCM), which employs biophysically motivated generative model that relates the observed BOLD data to neuronal signal [11]. Here, the causal influence is defined as a physical influence, where changing influences causes changes in their consequences [130], and it is modeled by a continuous time dynamic state-space system. The original formulation of DCM requires knowledge of known exogenous input, assigned to some of the network nodes, which drives the dynamics of the system. In this case, all hidden states are treated deterministically and the random term is considered only at the level of observation equation. The coupling and hemodynamic parameters are inferred through variational Bayesian formulation of EM algorithm, which maximizes the model evidence [11]. In this scenario, DCM is seen as a hypothesis-based approach to understanding distributed neuronal architectures underlying observed brain responses. Then, different hypotheses (model candidates) represented by different networks (or graphs) are compared based on the model fit reflected in evidence, via Bayesian model selection (BMS) [136]. However, as we emphasized in Chapter 1, one can do better if the model accounts for randomness at all levels, including hidden states, i.e. it is

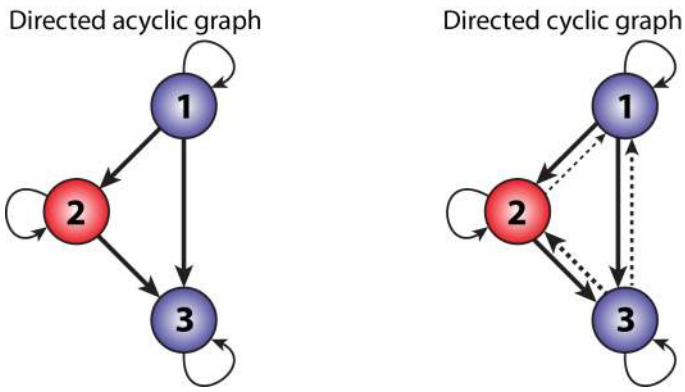


Figure 3.1 Difference between directed acyclic and directed cyclic graph. Note that cyclic graphs in our case allow for different strength of bidirectional connection; e.g. forward connection (solid line) can be stronger than the backward one (dashed line). Directed cyclic graph can be reduced to directed acyclic graph if any of two bidirectional connections are set (or estimated) to zero.

formulated as a fully stochastic system. Since we have introduced a fully stochastic scheme in the previous chapter for simultaneous estimation of neuronal signal (i.e. endogenous input) and model parameters for single time course, it is reasonable to think that the same scheme can be extended to multivariate case, which includes the modeling of neuronal interactions among different brain regions. In other words, we do not utilize any prior knowledge about the experimental causes of observed responses as required by deterministic DCM and introduce a stochastic DCM, which can be completely data-driven. This enables network discovery using both observed and unobserved responses during both activation based studies and (task-free) studies of autonomous or endogenous activity during the resting state [15]. In addition, because we jointly estimate both model parameters and neuronal signals in temporally sequential sense, i.e. we estimate the hidden states generating observed data, while properly accommodating endogenous inputs and model parameters, we implicitly assume that the uncertainty about the parameter estimates depends on uncertainty of hidden states (including endogenous inputs). This is more proper assumption compared to the deterministic DCM, which assumes that the uncertainty about parameters (after seeing data) does not depend on uncertainty about the states [15].

The rest of this chapter is structured as follows: First, we start with a state-space formulation of stochastic DCM where we focus on the form of the neuronal interaction model and its properties in terms of encoding a directional influences into the connectivity matrix. Second, we discuss different ways to perform model selection, i.e. how to compare different models and to infer whether some connections are likely to be present or not. Importantly, we show how to improve the system identification during the optimization process, by evaluating the importance of estimated coupling parameters.

3.2 Stochastic dynamic causal modeling

The stochastic DCM (sDCM) represents a straightforward extension to the deterministic DCM (dDCM), when it has the following properties: (i) it releases the need of known exogenous input; (ii) accounts for random process at hidden states level; (iii) and provides conditionally dependent estimates of the states and parameters.

Similarly to dDCM, sDCM is formulated as a multiple-input multiple-output (MIMO) system that comprises *m* inputs and *l* outputs with one output per region. Unlike in dDCM, where inputs must correspond to causes, $\mathbf{u}_t$ (i.e. designed exogenous inputs), in sDCM the inputs can be treated as endogenous $\tilde{\mathbf{z}}_t$, i.e. they can be generated by the fMRI data, which makes sDCM data-driven approach. However, it does not mean that sDCM is limited only to this scenario. Importantly, as we will see in the following section, one can still define any exogenous input as in the case of dDCM and use the sDCM for testing different hypotheses that motivated the experimental design but with the fully stochastic treatment of the model. Also, the stochastic formulation of DCM can be always easily converted to deterministic one, by setting the process noise variances of hidden states to zero (or to very small values). In either case, DCM rests on a choice of neuronal model $\mathbf{f}^n(.)$ of interacting cortical regions, which is defined in continuous time. This neuronal model is further supplemented with a forward hemodynamic model (summarized by $\mathbf{f}^h(.)$ and $\mathbf{h}(.)$), which describes how neuronal or synaptic activity is transformed into a measured response $\mathbf{y}_t$. This complete generative model allows to estimate the neuronal model parameters $\boldsymbol{\theta}_t^n$ (i.e. effective connectivity) from observed data, where the parameters represent couplings among unobserved brain states (i.e. neuronal activity in different brain regions), but it also accounts for parameterization of the hemodynamic response, $\boldsymbol{\theta}_t^h$. Then, the state-space model can have the following joint form:

$$\boldsymbol{\mu}_t = \begin{bmatrix} \tilde{\mathbf{z}}_t \\ \tilde{\mathbf{x}}_t \\ \boldsymbol{\theta}_t^{n,h} \end{bmatrix} = \begin{bmatrix} \mathbf{f}^n(\tilde{\mathbf{z}}_{t-1}, \boldsymbol{\theta}_{t-1}^n, \mathbf{u}_{t-1}) \\ \mathbf{f}^h\left(\tilde{\mathbf{x}}_{t-1}, \boldsymbol{\theta}_{t-1}^h, \tilde{\mathbf{z}}_{t-1}\right) \\ \boldsymbol{\theta}_{t-1}^{n,h} \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{q}}_{t-1}^n \\ \tilde{\mathbf{q}}_{t-1}^h \\ \mathbf{o}_{t-1}^{n,h} \end{bmatrix} \tag{3.1}$$

$$\mathbf{y}_t = \mathbf{h}(\boldsymbol{\mu}_t) + \mathbf{r}_t, \tag{3.2}$$

where for simplicity we skipped the notation for multiple *l* regions; e.g. $\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_l]^T$, $\tilde{\mathbf{x}} = [\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_l]^T$, $\boldsymbol{\theta}^h = \left[\boldsymbol{\theta}_1^h, \dots, \boldsymbol{\theta}_l^h\right]^T$, etc. Further, we mark the variables that are obtained by discretization of the continuous process, using a local linearization approach, with tilde.

In summary, each of the *l* regions is described by one neuronal state $\tilde{\mathbf{z}}$, four hemodynamic states $\tilde{\mathbf{x}} = [\tilde{s}, \tilde{f}, \tilde{v}, \tilde{q}]$, and by a set of hemodynamic parameters $\boldsymbol{\theta}^h$ (see Section 1.3 for description of the hemodynamic model). Crucially, all regions are coupled together (with mutual influence) through the neuronal model, where the

strength of couplings is encoded by parameters $\boldsymbol{\theta}^{n}$. The neuronal model represents a bottom of the generative model, where the neuronal activities from different regions talk to each other. It is supposed that there is no influence or interaction between hemodynamic states of different regions; i.e. at the higher level of the generative model. Schematic representation of this model is depicted in Figure 3.2.

Since the hemodynamic model is the same as described in section 1.3, we will proceed with description of the neuronal model, which is the core of the causal interactions and effective connectivity in particular.

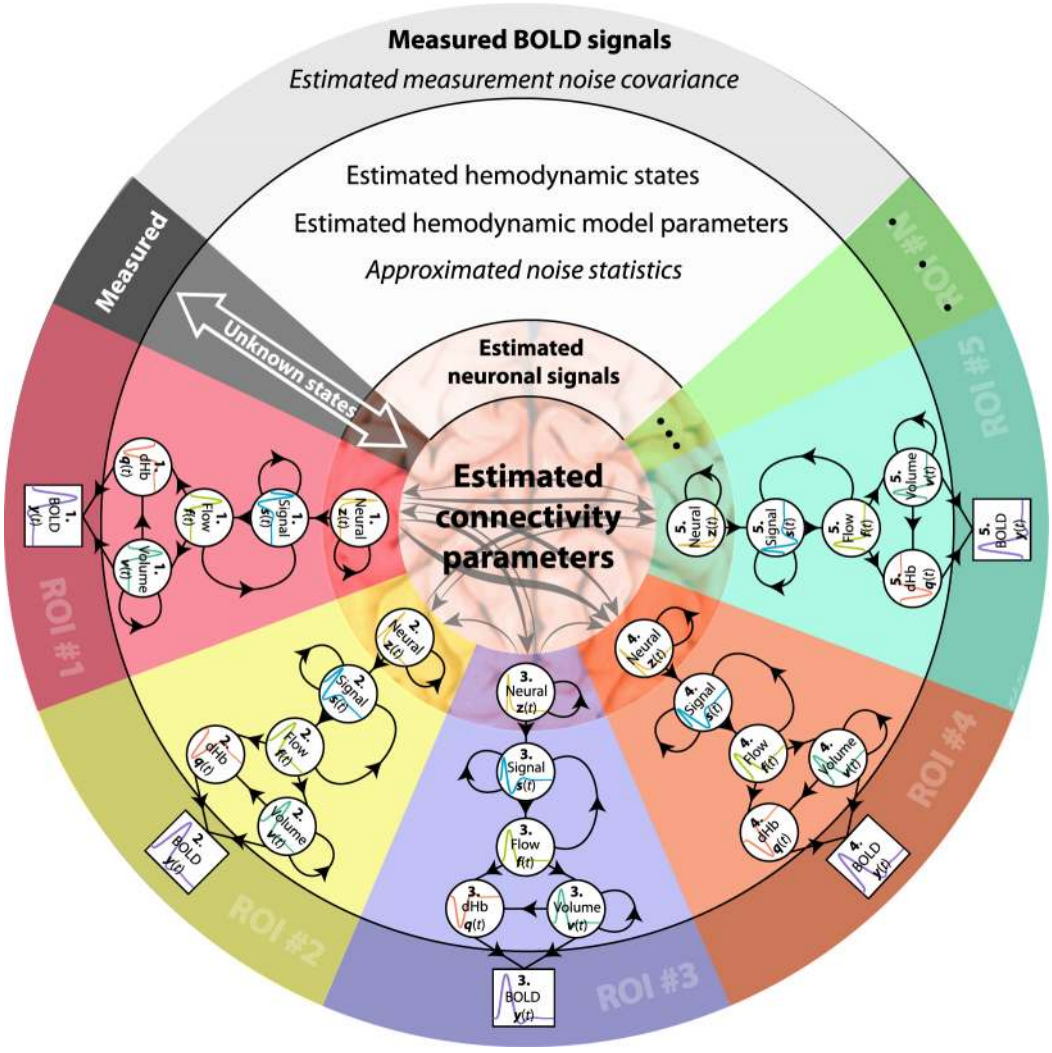


Figure 3.2 Schematic illustration of stochastic DCM. From measured BOLD signals associated with different brain regions we perform model inversion assuming hemodynamic model and neuronal model of interactions.

3.2.1 Neuronal interaction model

For fMRI data it is reasonable to define the model of neuronal interactions at macroscopic level, where one can study the whole brain dynamics and interactions between large-scale neuronal systems such as cortical regions. In this sense, it is common to consider a simple model of neuronal responses distributed over *l* nodes, where under the mean field assumption (see [137, 138]) the dynamics of one node are determined by the average activity of another. This is like assuming that each neuron in one node can see a sufficiently large number of neurons in another node to render the effective influence that is the same as the average over all neurons in the source node. As a result, only the slow dynamics are communicated among nodes, which means we can model distributed activity with a small number of macroscopic variables (e.g. one per node), whose time constants are greater than underlying fast fluctuations that are specific to each node. These fluctuations are continuous and can be represented by system noise. Therefore, the neuronal model can be described through simple linear stochastic differential equation:

$$\begin{aligned} d\mathbf{z}(t) &= \mathbf{f}(\mathbf{z}(t), \boldsymbol{\theta}^n(t))dt + \sqrt{\mathbf{Q}^n}d\mathbf{w}(t) \\ &= \mathbf{A}\mathbf{z}(t)dt + \sqrt{\mathbf{Q}^n}d\mathbf{w}(t), \end{aligned} \tag{3.3}$$

where

$$\mathbf{A} = \frac{\partial \mathbf{f}(\mathbf{z}, \boldsymbol{\theta}^n)}{\partial \mathbf{z}}, \tag{3.4}$$

is the connectivity matrix (the Jacobian), also called adjacency matrix, which represents the first-order connectivity among nodes [139]. The elements of this connectivity matrix are function of endogenous neuronal states and represent the unknown parameters which we want to estimate, $\boldsymbol{\theta}^n = \mathbf{A}$. One can also understand these coupling parameters of effective connectivity as a rate constants (with units s^{-1}) of neuronal population responses that have exponential nature (the solution of differential equation (3.3) is exponential function). In other words, a strong connection means an influence that is expressed quickly (or with a small time constant). Additionally, since the parameters $\boldsymbol{\theta}^n$ are estimated sequentially with the proposed model inversion scheme, it means that we are able to obtain time-varying parameters of effective connectivity, where the uncertainty about the parameters might change with time as well. Crucially, this continuous model allows estimating the cyclic directed graphs, i.e. it enables distinction

between forward and backward (feedback) connections. The neuronal interaction model is connected with the hemodynamic model (see Section 1.3) through (1.1), where the exogenous inputs $\mathbf{u}(t)$ are replaced with the endogenous activity $\mathbf{z}(t)$.

This is the simplest possible form of neuronal model that one can consider[4] but it is suitable for fMRI responses because it is generally assumed that fMRI signal scales with predominant frequency of neuronal activity [139, 140]. It should be noted that in the standard discrete DCM a more complex neuronal models are usually applied. The original neuronal model under DCM has a bilinear form [11], which models interactions between neuronal state **z** and a known input function **u,** i.e. experimentally controlled context variable. This enables context-dependent modulation of connection strengths, which can be thought of as a dynamic formulation of the so-called psycho-physiological interactions [6]. Later, two extensions of the neuronal state equation for DCM were introduced. First, one can extend the neuronal model with excitatory and inhibitory subpopulations in each region [141], allowing for an explicit description of intrinsic (between subpopulations) connectivity within a region. This is effectively modeled by two states per region, where by using positivity constrains, the model reflects the fact that extrinsic (inter-regional) connections of cortical areas are purely excitatory. Second, it is possible to account for nonlinear interactions among neuronal states, where the effective strength of connection between two regions is modulated by activity in a third region [142]. This nonlinear controlling allows modeling of various neurobiological processes, including attentional modulation, learning, and especially neuromodulation, i.e. mapping the modulatory influence to neuronal origin (state). These extended models are of high interest. Nevertheless, in order to make the concept more general and also understandable, we will consider only the simple linear form of neuronal model (3.3) in this work.

3.2.2 Structured priors on coupling parameters

In the following section, we discuss the basic properties of the adjacency matrix **A** in terms of the system stability and specification of priors on coupling parameters.

[4] Note that for data simulation we need to define the exogenous input $\mathbf{u}(t)$, where we consider the following model $d\mathbf{z}(t)/dt = \mathbf{f}(\mathbf{z}(t), \boldsymbol{\theta}^n(t), \mathbf{u}(t)) = \mathbf{A}\mathbf{z}(t) + \mathbf{C}\mathbf{u}(t)$. Matrix $\mathbf{C}$ is in this case represented by simple identity matrix.

In the absence of the input, the neuronal state must return to a stable mode. Mathematically, this means that the principal real eigenvalue of the square connectivity matrix **A** must be negative. In other words, we require the diagonal elements of **A** to be negative, i.e. these nodes have negative self-feedback, and if the overall feedback of the system is negative, then the system will tend to be stable. Further, we can factorize the connectivity matrix into a scalar σ and normalized couplings a_{ij}, such that the strength of connections among regions are relative to self-connections a_{ii}:

$$\mathbf{A} \to \sigma\mathbf{A} = \sigma \begin{bmatrix} -1 & a_{12} & \cdots & a_{ij} \\ a_{21} & -1 & & \\ \vdots & & \ddots & \\ a_{ij} & & & a_{ii} \end{bmatrix}. \tag{3.5}$$

This factorization enforces the same self-connection or temporal scaling σ in all regions. This assumption is supported by the fact that there is a minor reason to suppose that the neuronal dynamics, intrinsic to each region, will be very different (as opposed to the strong evidence that hemodynamics vary between regions) [11]. Also, the temporal scaling considered here is equal to $1/2$ s^{-1}, which is motivated by assumption that was described in the previous section that the time constants implicit in the neuronal model parameters are much longer (e.g. 100 to 10,000 ms) than the microscopic time constants (e.g. 1 to 100 ms). This is important because it suggests that priors on parameters in the effective coupling matrix **A** should allow for slow dynamics.

Above, we have defined the main structure of the adjacency matrix **A**. In fully Bayesian inversion scheme one could now simply define Gaussian priors on each of coupling parameters in order to ensure the proper structure [62]. For example, the priors on diagonal elements could be $p(a_{ii}) = \mathcal{N}(0, 10^{-3})$, where the small variance ensures that diagonal elements will remain very close to initial $-1/2$ during estimation, which will ensure that the system stays stable. Next, the priors on off-diagonal elements can be defined as $p(a_{ij}) = \mathcal{N}(0,2),\ i \neq j$, where the variance defines allowable range of the coupling parameters in adjacency matrix (assuming slow dynamics).

Unfortunately, our estimation scheme based on Kalman filter is not fully Bayesian, which means that we cannot define prior distribution, from which the parameters are sampled, directly. In Kalman filtering framework, we have to work with the transient priors, which are mainly defined through the initialization of the parameter error

covariance matrix (and regularized through the parameter noise covariance, which is responsible for the smoothness/perturbations). As it is already prompting by the name "transient", these priors are changing with time, i.e. the variance can be seen as a change in a parameter estimate that is allowed between time steps. Obviously, compared to the real prior variance, the variance of transient prior must be much smaller but still able to effectively cover (over the time) the allowed range. In this particular case we can initialize the parameter error covariance matrix (its diagonal elements) as:

$$\mathbf{P}_0^{\theta^n} = \begin{cases} 1 \cdot 10^{-2} & for\ off-diagonal\ elements\ of\ \mathbf{A} \\ 1 \cdot 10^{-4} & for\ diagonal\ elements\ of\ \mathbf{A} \end{cases} \tag{3.6}$$

The prior (initial) expectation on connectivity parameters will be then $\theta_{0,ij}^n = 0$ for off-diagonal elements and $\theta_{0,ij}^n = -1/2$ for diagonal ones. Since we consider different variances for different classes of parameters (based on their function), we will call these priors structured. Similarly, we can define the parameter noise covariance by using these structured priors, e.g. as $\mathbf{Q}_0^n = \mathbf{P}_0^{\theta^n} \cdot 10^{-3}$. The parameter noise covariance is then adaptively updated through the Robbins-Monro approximation (2.141). In addition, by defining the same prior variance on all off-diagonal elements, we allow for full-connectivity. It is then a goal of optimization scheme to estimate the correct parameter values, where we hope that during the convergence, the uncertainty on parameter estimates becomes smaller, i.e. their corresponding error (co)variances in $\mathbf{P}_t^{\theta^n}$ shrink.

Moreover, these prior means and variances can be effectively used to specify any model in terms of its adjacency matrix, which defines allowable connections or conditional dependencies among nodes. For example, if we initialize the prior variance on some connection a_{ij} to be zero (i.e. no uncertainty), we automatically force the posterior estimate to take the prior mean. Under zero mean prior, as in our case, a zero entry in the adjacency matrix thus prohibits an effective connectivity between respective nodes (regions). Conversely, if we allow finite variance on connection in the adjacency matrix, it means that this connection can have non-zero value of the posterior mean. By specifying the structured priors on elements of adjacency matrix we are ready to perform model inversion using SCKS.

Finally, these structured priors (both real and transient) will play an important role in the second level of inference, which is discussed in the next section. There we will be

comparing different model candidates, which will be defined by switching off subsets of connections within the full model using priors.

3.3 Model Selection

Until now, we have always considered only the first level of inference, where we fit the model m_i to the data **y**. Our model includes free parameters $\boldsymbol{\theta}$ and by fitting the model to the data we are inferring what values those parameters should probably take, given the data. As a result of this inference we have obtained the most probable posterior parameter estimates and the uncertainties on these estimates described by posterior error variances. Using Bayes' rule, we define the first level inference (model fitting) in the probabilistic sense as:

$$p(\boldsymbol{\theta}|\mathbf{y}, m_i) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, m_i)p(\boldsymbol{\theta}|m_i)}{p(\mathbf{y}|m_i)}. \tag{3.7}$$

In words:

$$Posterior = \frac{Likelihood \times Prior}{Evidence}. \tag{3.8}$$

The denominator on the right hand side of (3.7) represents normalizing constant known as evidence or marginal likelihood. The normalizing constant $p(\mathbf{y}|m_i)$ is ignored during the first level of inference. In our case, we obtain posterior estimates $p(\boldsymbol{\theta}|\mathbf{y}, m_i)$ of model parameters by using SCKS estimation scheme, where the estimates are optimal in both maximum likelihood and maximum *a posteriori* sense.

However, in the task where one wants to identify the connectivity couplings between nodes, which possibly involves many free parameters (depending on the size of network), we cannot be sure that the model m_i we have inverted is the best one. In other words, there are many possible models $m_i \in M$, where each model is defined by its unique structure (or adjacency matrix) of allowed connections between nodes. Therefore, after model fitting it is common (or necessary) to perform the second level of inference, represented by model comparison. At this level we wish to infer which model is the most plausible, given the data. This is the reason, why one usually considers a set of alternative model candidates, and for each of the model inversion is performed. In this case, the posterior probability of each model is given by:

$$p(m_i|\mathbf{y}) \propto p(\mathbf{y}|m_i)p(m_i), \tag{3.9}$$

where the data-dependent term on the right side of (3.9) is the marginal likelihood (evidence) that appeared already in (3.7), but was ignored. Assuming that there is no *a priori* belief that one model should be better than others, we usually assign equal priors $p(m_i)$ for all model candidates. This means that the models can be uniquely ranked by evaluating the marginal likelihood [63].

Unfortunately, the marginal likelihood is not straightforward to compute, since this computation involves integrating out the dependence on model parameters:

$$p(\mathbf{y}|m_i) = \int p(\mathbf{y}|\boldsymbol{\theta}, m_i)\, p(\boldsymbol{\theta}|m_i) d\boldsymbol{\theta}. \tag{3.10}$$

Therefore, the approximation to marginal likelihood (evidence) is generally considered. Critically, this approximation should represent a balance between the model fit and model complexity. It is not simply possible to choose the model that fits the data best, it is a known fact that more complex models can fit data better (especially if highly nonlinear models are assumed). This may result in overfitting, see Figure 3.3. This balance between model fit and complexity follows Occam's razor assertion that if two models fit equally well, the simpler model is likely to be better description of the reality. In this sense we can verbally express the evidence as:

$$Evidence(m_i) = Best\ fit\ likelihood(m_i) \times Complexity(m_i), \tag{3.11}$$

where the complexity term is usually known as Occam's factor (which is always less or equal to one), which scales with respect to the number of parameters (if number of parameters increases, Occam's factor decreases). Thus, the models with more parameters are automatically penalized. Note that this is very simplified interpretation of Occam's factor, which fits to our framework. In general, Occam's factor can represent much more than that (see [63]).

The following is a description of mathematical formulations and heuristics that are suitable for application to our estimation scheme in order to achieve successful model selection and an improved system identification.

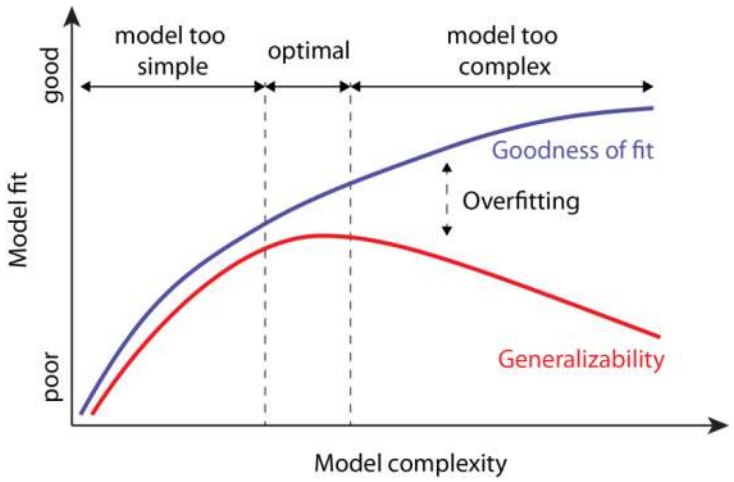


Figure 3.3 Illustration of optimal model fit. Relationship between goodness of fit (blue line) and generalizability (red line) as a function of model complexity. The *y*-axis represents any measure of goodness of fit (e.g. log-likelihood), where a larger value represents a better fit. The goal of model selection is to choose the model that generalizes the best across all model candidates.

3.3.1 Bayes factor

By following the Bayes' rule, models can be compared in Bayesian sense as a ratio of posterior model probabilities (i.e. posterior odds ratio), expressed as:

$$\frac{p(m_i|\mathbf{y})}{p(m_j|\mathbf{y})} = \frac{p(\mathbf{y}|m_i)}{p(\mathbf{y}|m_j)} \cdot \frac{p(m_i)}{p(m_j)} \tag{3.12}$$

The most right quantity in (3.12) is the prior odds ratio, if there is no prior preference for either model, the prior odds ratio will be equal to 1. Then the posterior odds ratio reduces to the ratio of marginal likelihood, which is called Bayes factor (B):

$$B_{ij} = \frac{p(\mathbf{y}|m_i)}{p(\mathbf{y}|m_j)}. \tag{3.13}$$

Now we know how to compare models between each other, but what we do not know yet is how to approximate the marginal likelihood. Fortunately, asymptotic approximation to the logarithm of marginal likelihood can be provided by Bayesian information criterion (BIC):

$$BIC_i = -2\mathcal{L}_i + n_{\theta,i} \log T, \tag{3.14}$$

where $\mathcal{L}_i$ is the log-likelihood obtained during model inversion, computed according to (2.112)[5] and $-2\mathcal{L}_i$ is known as deviance. The second term on the right side of (3.14) represents the approximation to the model complexity (Occam's factor), where $n_{\theta,i}$ is the number of free parameters considered in the model, and T is the number of observation samples; i.e. number of time points over which we have accumulated our log-likelihood $\mathcal{L}_i$. In [143] it is said that the difference of BIC for two models gives a rough approximation to minus twice the logarithm of the Bayes factor, which is easy to use and does not require evaluation of prior distribution.

$$2\ln(B_{ij}) = -(BIC_i - BIC_j). \tag{3.15}$$

The regular Bayes factor can be then obtained by conversion from log-space as:

$$B_{ij} = \exp\left(-\frac{1}{2}(BIC_i - BIC_j)\right). \tag{3.16}$$

The interpretation of Bayes factor is summarized in the following Table 3.1.

Table 3.1 Bayes factor and posterior model probability classification range.

$2\ln(B_{ij})$	B_{ij}	$p(m_i\|\mathbf{y})$	Evidence m_i vs. m_i
< 2	< 3	50-75%	Weak
-2 to 6	3 to 20	76-95%	Positive
6 to 10	20 to 150	96-99%	Strong
≥ 10	≥ 150	$\geq 99\%$	Very strong

Bayes factors are very flexible, they allow for multiple hypotheses to be compared simultaneously, and nested models are not required in order to make comparison. A common way how to interpret Bayes factors is by converting them to posterior model probabilities. For a finite set of competing models M, the posterior model probabilities in M are proportional to their Bayes factors with respect to a common model m_c:

$$p(m_i|\mathbf{y}) = \frac{B_{ic}}{\sum_{m_j \in M} B_{jc}}. \tag{3.17}$$

[5] Note that no constants should be discarded in calculating the log-likelihood. $\mathcal{L}_i$ is the accumulated likelihood over the time samples $\mathcal{L}_i = \sum_{t=1}^{T} \mathcal{L}_{t,i}$, where $\mathcal{L}_{t,i}$ for given model m_i is computed according to (2.112).

In our case, we consider m_c to be the model with the minimum value of BIC compared to all other models from the model space M. This assumption is correct if our model space M contains all possible models; i.e. all possible structures of adjacency matrix.

Many other criteria similar to BIC such as Akaike information criterion (AIC) are just as simple to calculate:

$$AIC_i = -2\mathcal{L}_i + 2n_{\theta,i} \tag{3.18}$$

Indeed, it is suggested that whenever BIC is presented for comparison, one should also calculate AIC. It is well-known that BIC tends to favor simpler models, whereas AIC tends to favor more complex models [143, 144]. Reporting both BIC and AIC may therefore be a crude form of sensitivity analysis: if these two criteria select the same model, one can be more confident about the result [145].

Although this approach to model selection seems to perform well, it has two limitations. First, the approximation to the log marginal likelihood based on BIC ignores the uncertainty about the model parameters. Second, it is clearly limited to relatively small size of network (or small number of coupling parameters). In our case, the former limitation does not seem to be a serious issue, mainly because we effectively estimate the measurement noise variance and also adaptively approximate parameter noise variance and hemodynamic state noise variance, which all contribute to the regularization of parameter estimate. This partially protects our model against overfitting. In this case, model selection criteria such as BIC can still provide a very sensible suggestion. Conversely, the later limitation represents a relevant disadvantage, since it starts to be a very common practice to search over a larger number of competing models. Therefore, we are interested in model selection strategies, which can compare a large number of models, but do not require to invert each model variant separately.

To emphasize the motivation for this, it is important to describe what sizes of model space we have in mind. For example, if we constrain ourselves to search a model space where only bidirectional connections are allowed (which does not mean that the couplings in both directions have the same strength), then we can calculate the number of possible models with respect to the number of nodes by using simple combinatorics $|\mathbf{A}(m_i)| = 2^{l(l-1)/2}$. For network containing four nodes we have 64 possible models, whereas for 6 nodes it is already 32,768 models. Of course, we would like to consider

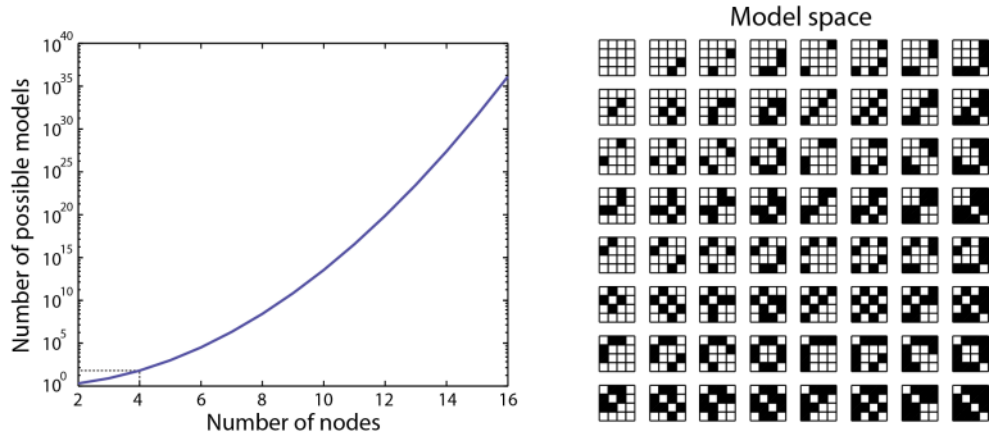


Figure 3.4 Model space and adjacency matrices. The plot on the left side shows the number of different models that one can assume as a function of the number of nodes (with the constraint on the models that each connection has to be bidirectional). On the right side we show all alternative models that can be considered, given four nodes.

even larger networks. From Figure 3.4, we can see that the number of models grows really rapidly.

3.3.2 *Post hoc* Bayesian model selection

To address a problem of selecting the best model among large number of model candidates, in [146] they introduced a Bayesian model selection procedure for *post hoc* inferences about reduced (nested) versions of a full model. This method enables to calculate the marginal likelihood for any reduced model that is nested within a larger (full) model as a function of the posterior density of the full model. Critically, this procedure requires only a single inversion of the full model, where all connections are allowed.

Consider a reduced model m_i , where a subset of parameters $\boldsymbol{\theta}_i$, $\boldsymbol{\theta}_i \subset \boldsymbol{\theta}_F$, is constrained to some special value, i.e. $\boldsymbol{\theta}_i = \boldsymbol{\theta}_0$, which is known as a sharp hypothesis, or "point null". In our case we consider $\boldsymbol{\theta}_0 = 0$. In contrast, a full model m_F assumes that $\boldsymbol{\theta}_i$ is free to vary; i.e. $\boldsymbol{\theta}_i \neq 0$. Now, if we consider the full model m_F and let $\boldsymbol{\theta}_i \rightarrow 0$, it effectively means that the full model becomes a reduced model m_i. In other words, the reduced model is nested under the full model. Under this assumption we can reformulate the Bayes factor (3.13) by considering only full model m_F, and dividing the posterior density for parameters $\boldsymbol{\theta}_i$ by the prior density for $\boldsymbol{\theta}_i$, at the point of interest; i.e. at $\boldsymbol{\theta}_i = 0$:

$$B_{iF} = \frac{p(\mathbf{y}|m_i)}{p(\mathbf{y}|m_F)} = \frac{p(\boldsymbol{\theta}_i = 0|\mathbf{y}, m_F)}{p(\boldsymbol{\theta}_i = 0, m_F)}. \tag{3.19}$$

This result is generally known as Savage-Dickey density ratio [147, 148]. From this formulation we can immediately derive the marginal likelihood of any reduced model:

$$p(\mathbf{y}|m_i) = p(\mathbf{y}|\boldsymbol{\theta}_i = 0, m_F) = \frac{p(\boldsymbol{\theta}_i = 0|\mathbf{y}, m_F)p(\mathbf{y}|m_F)}{p(\boldsymbol{\theta}_i = 0, m_F)}, \tag{3.20}$$

where we can see that the marginal likelihood under the reduced model is just the marginal likelihood of the full model times the posterior expectation of the prior density ratio. This when further transformed to log space is equal to:

$$\ln p(\mathbf{y}|m_i) = \ln p(\boldsymbol{\theta}_i = 0|\mathbf{y}, m_F) - \ln p(\boldsymbol{\theta}_i = 0, m_F) + \ln p(\mathbf{y}|m_F). \tag{3.21}$$

Here the last term on the right side of (3.21) is the log marginal likelihood of the full model, which is a constant for all reduced models, and therefore we can treat it as zero:

$$\ln p(\mathbf{y}|m_i) \approx \ln p(\boldsymbol{\theta}_i = 0|\mathbf{y}, m_F) - \ln p(\boldsymbol{\theta}_i = 0, m_F). \tag{3.22}$$

This also means that the full model shares the same likelihood with any reduced model.

At this point it is necessary to define an approximation to marginal likelihood, which accounts for uncertainty on posterior parameter estimates. In this case, one can approximate the log marginal likelihood by the variational free-energy $\mathcal{F}_i$ [62, 75]. Considering the result of full model inversion, which provides the posterior density on model parameters $p(\boldsymbol{\theta}_F|\mathbf{y}, m_F) = \mathcal{N}(\boldsymbol{\theta}_F, \mathbf{P}_F^{\theta})$, and also the fact that we have initialized the model inversion by reflecting the (real) structured priors on coupling parameters $p(\boldsymbol{\theta}_F, m_F) = \mathcal{N}(\boldsymbol{\eta}_F, \boldsymbol{\Sigma}_F^{\theta})$ as described in Section 3.2.2, it is possible to define the posterior and prior densities for any reduced model as:

$$p(\boldsymbol{\theta}_i = 0|\mathbf{y}, m_F) = \mathcal{N}(\boldsymbol{\theta}_i, \mathbf{P}_i^{\theta}) \quad with\ \mathbf{P}_i^{\theta} = \left[\boldsymbol{\Upsilon}_i^{\theta}\right]^{-1} \tag{3.23}$$

$$p(\boldsymbol{\theta}_i = 0, m_F) = \mathcal{N}(\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_i^{\theta}) \quad with\ \mathbf{P}_i^{\theta} = \left[\boldsymbol{\Pi}_i^{\theta}\right]^{-1}. \tag{3.24}$$

In particular, any reduced model can be created from the full model by collapsing the prior density over one or more parameters; i.e. by setting the corresponding elements of the prior mean $\boldsymbol{\eta}_i$ and precision $\boldsymbol{\Pi}_i^{\theta}$ to zero. Then the free-energy of reduced model m_i

can be expressed as a simple analytic function of the means and precisions[6] of the prior and posterior of the full model [146]:

$$\mathcal{F}_i = \frac{1}{2}\ln\frac{\left|\boldsymbol{\Pi}_i^{\theta}\right|\left|\boldsymbol{\Upsilon}_F^{\theta}\right|}{\left|\boldsymbol{\Upsilon}_i^{\theta}\right|\left|\boldsymbol{\Pi}_F^{\theta}\right|} - \frac{1}{2}\left(\boldsymbol{\theta}_F^T\boldsymbol{\Upsilon}_F^{\theta}\boldsymbol{\theta}_F + \boldsymbol{\eta}_i^T\boldsymbol{\Pi}_i^{\theta}\boldsymbol{\eta}_i - \boldsymbol{\eta}_F^T\boldsymbol{\Pi}_F^{\theta}\boldsymbol{\eta}_F - \boldsymbol{\theta}_i^T\boldsymbol{\Upsilon}_i^{\theta}\boldsymbol{\theta}_i\right), \quad (3.25)$$

where the reduced posterior precision $\boldsymbol{\Upsilon}_i^{\theta}$ is the posterior precision of the full model $\boldsymbol{\Upsilon}_F^{\theta}$ plus the difference between the reduced and full prior precisions $\boldsymbol{\Pi}_i^{\theta}$ and $\boldsymbol{\Pi}_F^{\theta}$, respectively. Similarly, the reduced posterior mean $\boldsymbol{\theta}_i$ is a mixture of precision-weighted means:

$$\boldsymbol{\Upsilon}_i^{\theta} = \boldsymbol{\Upsilon}_F^{\theta} + \boldsymbol{\Pi}_i^{\theta} - \boldsymbol{\Pi}_F^{\theta} \quad (3.26)$$

$$\boldsymbol{\theta}_i = \mathbf{P}_i^{\theta}\left(\boldsymbol{\Upsilon}_F^{\theta}\boldsymbol{\theta}_F + \boldsymbol{\Pi}_i^{\theta}\boldsymbol{\eta}_i - \boldsymbol{\Pi}_F^{\theta}\boldsymbol{\eta}_F\right). \quad (3.27)$$

The last equation is especially interesting, because it represents a reconstructed (re-weighted) vector of parameter estimates without contribution of certain parameters.

Importantly, this post-hoc estimate of the free-energy based approximation of log marginal likelihood can now be used to compute the Bayes factor, comparing the reduced model m_i with reduced model m_j, as:

$$B_{ij} = \exp\left(\mathcal{F}_i - \mathcal{F}_j\right). \quad (3.28)$$

Although, the *post hoc* model selection based on reduced free-energy enables to compare relatively large number of models (millions), still for a large size networks having seven and more nodes, the number of possible combinations among nodes becomes really huge (see Figure 3.4). In this case, one has to apply a greedy search through the space of reduced free-energy. This entails identifying a subset of parameters, with the least free-energy and searching over all reduced models within that subset. Redundant parameters are then removed and the procedure repeated until all model parameters have been considered or no further parameters can be removed [146, 149].

[6] Precision is the inverse of variance.

3.3.3 Network pruning

In the previous section we have shown how effectively we can compare a large number of different reduced models based only on a single inversion of one full model; i.e. the model where all possible connections between network nodes are allowed. Although this model selection does a great deal of work for us, it is still the first level of inference (model inversion) which must provide a confident estimates of coupling parameters, given the data. However, since there are always some random correlations between time courses (which correspond to particular nodes in the network), the inversion scheme does not set automatically the irrelevant coupling parameters to zero. Thus these spurious couplings (with non-zero variance) spoil the performance of model inversion. Therefore, we seek a procedure, which can automatically infer the relevant connections and suppress the irrelevant ones.

In principle, there are two ways how to achieve this. The first one supplements a penalty term to the objective function, which causes that the irrelevant couplings tend to zero value. This is effectively accomplished by introduction of shrinkage priors on coupling parameters, which are e.g. of Gaussian distribution $p(a_{ij}) = \mathcal{N}(0, 1/\lambda_{ij})$. Here λ_{ij} represents a regularizing constant (precision) associated with the coupling parameter a_{ij}. During the learning process of coupling parameters and regularizing constants, if the coupling parameter in question is not relevant then the precision λ_{ij} will be large, thus forcing the parameter to be close to zero. This approach is known as automatic relevance determination (ARD) [63, 150]. The second way involves an estimation of sensitivity of the error function to removal of a coupling parameter (when set to zero), where the connections with the least effect on the error function are subsequently removed. This clearly requires a threshold that has to be specified *a priori*. This approach is mostly known as network pruning, when probably the most popular pruning methods are the *optimal brain damage* [151] and the *optimal brain surgeon* [152], which perform the pruning of irrelevant couplings off-line. However, there are other pruning methods that can be applied also online [153-156]. Although the ARD and the network pruning approaches are often cast into two different groups, they have obviously a common aim and partly overlap in the use of the relevance measure. As we have already mentioned earlier, the inverse of the covariance matrix (precision), $\boldsymbol{\Upsilon}^{\theta} = \left(\mathbf{P}^{\theta}\right)^{-1}$, is equivalent (for Gaussian random variables) to the expected Fisher

information matrix (FIM), or also corresponds to the expected Hessian matrix, which can be used to study system identifiability [96]. Critically, all methods mentioned above use FIM (exact or its approximations) to construct an error function, or they include FIM as a term into the objective function. Therefore, both of these approaches can be considered as ARD methods, where maybe the most important difference is, that the first (Bayesian) approach uses optimization of regularizing constant (considering priors) to perform "soft" ARD, whereas pruning involves selection of some threshold, which we can be seen as a "hard" ARD.

The estimation framework that is considered in this work does not allow to optimize the priors and include them directly into Kalman filter. Therefore, we will proceed with the variant of hard ARD based on network pruning. In this sense, we will partly follow a pruning algorithms developed for online application [153-156].

We might start with the assumption that the connections between nodes are distributed sparsely with lots of small or absent couplings and relatively small number of strong couplings, which is our motivation for application of network pruning. We will also consider only off-diagonal elements of the connectivity matrix **A**, because we cannot attempt to prune an entire node that is associated with the measured time course. Next, we do not want to remove the irrelevant coupling completely from the network (so the number of parameters would change), but rather set its mean to zero and minimize its variance. Finally, we suggest to perform pruning after each iteration of SCKS algorithm, and calculate the relevance measure based on the time average estimates of the expected parameter means $\boldsymbol{\theta}$ and error covariance matrix $\mathbf{P}^{\theta}$. By saying this, we should not start pruning immediately after the first iteration of SCKS, since at this point we might not have reached a sufficient convergence. For example, we can consider a starting condition based on the change of log-likelihood (2.112), if this change is sufficiently small (as defined below), we enable the pruning procedure.

The pruning method considered in this work is based on computation of importance function of individual coupling parameters (or subset of them) by determining sensitivity on their removal; i.e. by setting $\boldsymbol{\theta}_i = 0$. Here both the importance function and the sensitivity are derived from a scaled FIM:

$$E = \boldsymbol{\theta}^T \boldsymbol{\Upsilon}^{\theta} \boldsymbol{\theta}, \tag{3.29}$$

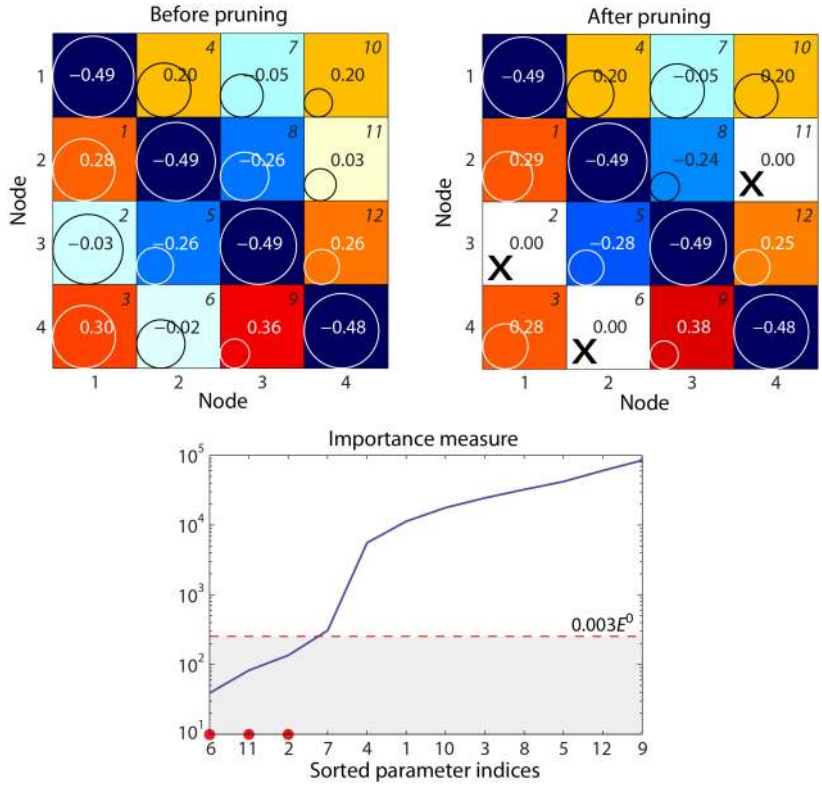


Figure 3.5 Illustration of the pruning procedure. The upper part shows the estimate of adjacency matrix obtained after i-th iteration (e.g. $i = 5$) before pruning procedure was applied (left) and after pruning (right). Here the coupling parameters scored with importance measure that is lower than *a priori* selected threshold (lower plot, *y*-axis is displayed in log scale to enable visualization of the threshold) are set to zero (black cross).

where precision $\boldsymbol{\Upsilon}^{\theta}$ is in this case a diagonal matrix (off-diagonal elements are ignored, i.e. effectively set to zero). The pruning procedure is then very simple. All the steps are summarized in Algorithm 9. Note that once we determine the coupling parameters that are not important, we perform their removing by using similar strategy as was described in Section 3.3.2, equations (3.26)-(3.27), where the corresponding elements of the prior mean $\boldsymbol{\eta}_i$ and precision $\boldsymbol{\Pi}_i^{\theta}$ are set to zero. This enables us to obtain a new vector of parameters that is reweighted with respect to the uncertainty of the remaining (important) coupling parameters.

In general, this pruning algorithm is very fast. Since it enables us to remove at least some of irrelevant parameters already during the optimization process, it contributes to the reduction of model space that we have to search after the model inversion. Critically, with each elimination of an irrelevant parameter by pruning, we improve the identifiability of the system, and with the next iteration of SCKS algorithm we thus

make an optimization step towards the new optimally selected reduced model. Finally, the effect of pruning becomes especially significant when larger networks are estimated.

Algorithm 9. Network pruning

- *Iterate SCKS algorithm and start pruning if the condition* $(\mathcal{L}_d - \mathcal{L}_{d-1}) < (\mathcal{L}_2 - \mathcal{L}_1)/4$ *is satisfied:*
    1. Calculate the average parameter error covariance matrix:
    $$\mathbf{P}^{\theta} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{P}_t^{\theta} \tag{3.30}$$
    2. Calculate the inverse of parameter error covariance matrix:
    $$\mathbf{\Upsilon}^{\theta} = \left(\mathbf{P}^{\theta}\right)^{-1} \tag{3.31}$$
    3. Evaluate the error function for all (off-diagonal) coupling parameters:
    $$E_k^0 = \left[\mathbf{\Upsilon}^{\theta}\right]_{kk}\boldsymbol{\theta}_k^2 \tag{3.32}$$
    4. Rearrange the list of indices i according to ascending order of E_k^0 values.
    5. Start removing the contribution of the coupling parameters (based on the order of i), by successively setting $\boldsymbol{\theta}_i = 0$, and evaluating the error function:
    $$E_k^i = \boldsymbol{\theta}_i^T\mathbf{\Upsilon}^{\theta}\boldsymbol{\theta}_i \tag{3.33}$$
    6. Once we obtained $\mathbf{E}^i = \left[E_1^i, \dots, E_K^i\right]^T$ up to the removal of the last parameter (i.e. all parameters were removed), we calculate the difference:
    $$\Delta\mathbf{E}^i = E_k^0 - \mathbf{E}^i \tag{3.34}$$
    with $E^0 = \sum_{k=1}^{K} E_k^0$.
    7. Subsequently, decide which parameters are not relevant by considering the threshold:
    $$\Delta\mathbf{E}^i \leq \delta E^0, \tag{3.35}$$
    with $\delta = 0.003$.
    8. Eliminate contribution of irrelevant parameters (based on the indices i that passed the threshold) by setting their prior precision to zero,

$\boldsymbol{\Pi}_i^{\theta} = 0$, ($\boldsymbol{\Pi}_i^{\theta} = \left(\boldsymbol{\Sigma}_i^{\theta}\right)^{-1}$), and update the error covariance and parameter vector:

$$\mathbf{P}^{\theta} = \left(\boldsymbol{\Upsilon}_F^{\theta} + \boldsymbol{\Pi}_i^{\theta} - \boldsymbol{\Pi}_F^{\theta}\right)^{-1} \tag{3.36}$$

$$\boldsymbol{\theta} = \mathbf{P}^{\theta}\left(\boldsymbol{\Upsilon}_F^{\theta}\boldsymbol{\theta}_F + \boldsymbol{\Pi}_i^{\theta}\boldsymbol{\eta}_i - \boldsymbol{\Pi}_F^{\theta}\boldsymbol{\eta}_F\right) \tag{3.37}$$

- *Proceed with the next iteration of SCKS algorithm, and consequently repeat the pruning procedure.*

3.4 Chapter summary

This chapter introduced a novel data-driven approach to evaluate effective connectivity. It represents a fully stochastic variant of dynamic causal modeling, where the coupling parameters and region-specific parameters of BOLD response are subsumed into the state vector and estimated jointly with the endogenous neuronal signals and hemodynamic states. This sequential estimation is performed by using the iterative scheme of square-root cubature RTS smoother that was developed in Chapter 2.

In order to approximate a macroscopic model of neuronal dynamics and interactions, we have considered a linear model in the form of stochastic differential equation (3.3). This model could be later extended to also accommodate nonlinear interactions or to model excitatory and inhibitory states of neuronal coupling.

We have also described two different methods to perform model selection. Since, there is clear need to score a large number of models, the *post hoc* Bayesian model selection [146] seems to be the best choice. In our case, we can apply this type of model selection under the assumption that the transient priors on coupling parameters, which are part of Kalman filter, are derived from the real structured priors (in Section 3.2.2). However, we should note that it would be more appropriate and also beneficial to consider real priors directly as a part of the optimization scheme. Optimization of regularizing constants with respect to the priors can have in this sense significant influence on the performance of estimation procedure, where it effectively performs automatic relevance determination (ARD). Nevertheless, in order to finesse the limitation of our method, we have achieved improved optimization of coupling

parameters by applying a pruning algorithm, which also allows ARD of important couplings (though it requires *a priori* selection of some small threshold).

In future the future it could be interesting to pursue a fully Bayesian sequential optimization procedure, e.g. through generalized filtering based on generalized coordinates of motion [52], but still keep the approximation of the probability densities using the cubature integration rules.

Chapter 4

Validation and application of the method

4.1 Introduction

The first three chapters described the theoretical background and introduced the new algorithm for the estimation of neuronal signal and evaluation of effective connectivity from fMRI data. Until this point, we have not shown or discussed any results that can be obtained by applying these algorithms. Therefore, it will be the focus of this chapter to provide a sufficient demonstration and validation of the proposed method. As in Chapters 2 and 3, we split the current chapter into two sections. The first section will focus on a simple demonstration of neuronal signal estimation from single fMRI time course, where we also focus on the identification problem of hemodynamic model parameters. The second section, which is of our main interest, will focus on the estimation of coupling parameters in neuronal interaction model. We demonstrate the estimation of effective connectivity (i.e. model inversion) by using the stochastic DCM. In particular, we will try to address the main concerns that are very often associated with the methods designed for evaluation of effective connectivity or connectivity analysis in general.

Finally, the validation and the performance evaluation will be performed mainly by using simulated data that have a character of endogenous activity present in resting-state fMRI data. Additionally, we will present an application to empirical fMRI data.

4.2 Single time course model inversion

In this section we take a closer look at the hemodynamic model inversion by the SCKS algorithm for a single time course. In this sense it will be very important to first discuss the model and algorithm initialization conditions, including some constrains that we consider on hidden states and model parameters. These will be summarized in the following section together with the description of data simulation and with the obtained results. Moreover, in the same section we will already consider a restriction on hemodynamic model parameter space, which is the result of model identifiability analysis provided in the Section 4.2.2.

4.2.1 Hemodynamic model inversion and estimation of neuronal signal

As a first step to test the model inversion by SCKS algorithm, we have to generate a synthetic data as a reasonable approximation to real fMRI time courses. Since we want to formulate the inversion problem as a nonlinear blind deconvolution, where we do not have any knowledge about the input causing the hemodynamics, which in real situation corresponds mainly to estimation of endogenous neuronal activity from resting-state data (but the same assumptions can be made also in the case of task data), we will consider simulated data having the character of resting-state fMRI time courses. After obtaining the data, we describe a general initialization of the SCKS algorithm and highlight the constraints we make about the hemodynamic states and their parameters to achieve the improved stability of the model inversion. Finally, we perform the model inversion by SCKS and discuss the results.

Simulations

The generation of synthetic data representing the resting-state time courses, starts with the selection of neuronal model. In this work, we have mentioned two models as an approximation to neuronal signal $\mathbf{z}_{1:T}$.The first one is based on simple discrete random-walk model (2.109) that was considered in Chapter 2. Note that this model is suitable only for modeling and estimation of single time course, which is now the case. However, because we want to use a unified framework for both single time course and multiple time courses model inversion, we choose here the second model based on stochastic differential equation (3.3). In this case, the neuronal model is restricted to a

single node with $a_{11} = -1/2$. The example using the first variant based on the discrete random-walk model can be found in our published papers [20, 120, 122].

In particular, to generate the neuronal signal $\mathbf{z}_{1:T}$ we employed the version of neuronal model described in footnote 4 on page 74, where the input $\mathbf{u}_{1:T}$ is a smoothed sequence of random Gaussian variables, so that it has a variance 0.02 and a Gaussian autocorrelation function 6 s. This neuronal signal caused changes in the states of hemodynamic model (1.1)-(1.4), i.e. inducing signal, blood flow, blood volume and deoxyhemoglobin content, all integrated with the time step $\delta = 100$ ms. Then the fMRI signal was generated by using BOLD observation equation (1.7). The hemodynamic model parameters were set to their usual prior means (see Table 1.1) and scaled with a deviate sampled from a log-normal distribution, $\boldsymbol{\theta}^h = \boldsymbol{\theta}_0^h \exp(\mathcal{N}(0,0.01))$, where the function of this scaling is explained below.

We also considered a small amount of additive Gaussian noise with the variance $3 \cdot 10^{-4}$ to all hidden hemodynamic states. To make the noise contribution more realistic, we further considered the noise sequences to be slightly correlated with the width of Gaussian autocorrelation function equal to 1 s. At the level of the observation signal we added a Gaussian innovations to produce the signal to noise ratio (SNR) equal to 2, where the SNR is defined as [142]:

$$\mathrm{SNR} = \frac{\sigma_{BOLD}}{\sigma_{noise}}. \tag{4.1}$$

This says that for SNR = 2, the standard deviation of the added observation noise σ_{noise} equals to half the standard deviation of the noise-free BOLD signal σ_{BOLD}. Based on the variance of input signal 0.02, we can expect the variance of noise-free BOLD signal (model output) to be about 0.1, then the variance of observation noise is chosen to be 0.025.

Finally, the simulated BOLD signal was downsampled with a sampling period (TR) which equals 2 s, resulting in signal length of $T = 256$ time points.

Model inversion

Before we approach the model inversion, we should mention the initialization conditions and eventually constrains that we apply to neuronal state, hemodynamic states, model parameters, and hyperparameters, respectively:

- *Neuronal state:* The initial condition for neuronal state is $z_0 = 0$, with the error variance $P_0^z = 0.01$. A very important is the prior variance of neuronal state noise, $Q_0^n = 5 \cdot 10^{-5}$, which is considered fixed during the estimation procedure. This value is suitable for BOLD signal change approximately about 1% (as it was simulated[7]) [26, 157]. By changing the variance of neuronal state we can control the smoothness of neuronal signal estimate (also partially control possible overfitting). As we will see in the consequent paragraph, this noise variance is assumed much larger than the noise variance for hemodynamic states. This effectively treats the neuronal fluctuations as the predominant source of hemodynamics and assumes that the hemodynamic fluctuations are largely neuronal in origin [15].

- *Hemodynamic states:* For hemodynamic states we consider a transformation to the log space, which guarantees that these states will always have positive values (the negative values of hemodynamic states would be physically meaningless), and ensures the numerical stability during the parameter optimization. Specifically, we convert the hemodynamic states equations, i.e. $\mathbf{x}(t) = [s(t), f(t), v(t), q(t)]^T$, to log space by considering a change of variables, $\mathbf{x}'(t) = \ln(\mathbf{x}(t))$ [142]. That is, for any given hemodynamic state variable $\mathbf{x}(t)$ with the state equation $d\mathbf{x}(t)/dt = \mathbf{f}(\mathbf{x}(t))$:

$$\mathbf{x}'(t) = \ln(\mathbf{x}(t)) \quad \Leftrightarrow \quad \mathbf{x}(t) = \exp(\mathbf{x}'(t)), \tag{4.2}$$

  then by applying the chain rule we get:

$$\frac{d\mathbf{x}'(t)}{dt} = \frac{d\ln(\mathbf{x}(t))}{dt} = \frac{\mathbf{f}(\mathbf{x}(t))}{\mathbf{x}(t)}. \tag{4.3}$$

  It is worth mentioning that this log-transformation does not affect the model parameters, only the hemodynamic states. In contrast, when evaluating the

[7] The signal change 1 % is in our case equivalent to the signal with a peak amplitude about 1 with respect to the zero baseline. This amount of percent signal change is common for fMRI data.

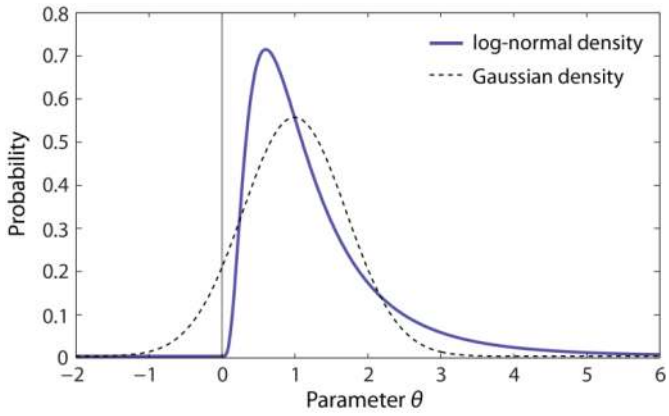


Figure 4.1 Log-normal probability density function. This log-normal function was created with mean of 1 and variance 0.5. For comparison, a Gaussian probability density function with identical mean and variance is shown (dashed line). Note that in contrast to the Gaussian, the support of log-normal density is restricted to positive numbers.

BOLD output equation, the hemodynamic states are exponentiated [142]. In other words, during each measurement update step of SCKS we use $\mathbf{x}_{t|t-1}$ to computed predicted BOLD signal at time t, not $\mathbf{x}'_{t|t-1}$. As a consequence of this log-transformation, we initialize the state vector of four hemodynamic states by $\mathbf{x}_0 = [0,0,0,0]^T$, with state error covariance matrix $\mathbf{P}_0^x = 0.01 \cdot \mathbf{I}_{n_x}$. The state noise covariance is then initialized as $\mathbf{Q}_0^h = 2 \cdot 10^{-9} \cdot \mathbf{I}_{n_x}$.

- *Model parameters:* For hemodynamic model parameters we use the empirical priors as summarized in Table 1.1. As a result of model identifiability analysis (described in consequent section) we choose to estimate only three model parameters $\boldsymbol{\theta}^h = [\theta^\kappa, \theta^\varphi, \theta^\varepsilon]^T$, representing the rate of signal decay (κ), the resting oxygen extraction fraction (φ), and ratio of intra- and extravascular signal (ε), respectively. Further, because these parameters cannot be negative, we apply a positivity constraint by considering the scaling of the prior parameter value by a log-normal density. For instance, the actual value of the rate of signal decay at time t is defined as $\kappa_t = \kappa_0 \exp(\theta_t^\kappa)$, where we are estimating the scaling parameter θ_t^κ. Similarly we define also the other two parameters. The example of this constrain by using a log-normal density function is depicted in - Figure 4.1. These scaling parameters are initialized as $\boldsymbol{\theta}_0^h = [0,0,0]^T$, with the error covarinace matrix $\mathbf{P}_0^{\theta^h} = \mathrm{diag}([5 \cdot 10^{-3}, 5 \cdot 10^{-3}, 1 \cdot 10^{-4}])$ and the noise covariance $\mathbf{O}_0^{\theta^h} = 1 \cdot 10^{-4} \cdot \mathbf{P}_0^{\theta^h}$.

- *Hyperparameters:* During the model inversion we adaptively estimate the noise statistics. The hemodynamic state noise covariance matrix and parameter noise covariance matrix are estimated through the Robbins-Monro stochastic approximation scheme (2.141). Besides the initialization of these matrices as mentioned above, we need to choose a forgetting factors. In this case we use $\lambda_x = 0.997$ for states and $\lambda_\theta = 0.99$ for parameters. Next, we need to initialize parameters of inverse-Gamma distribution for estimation of measurement noise. Assuming rather slow dynamics of the noise, we use $\rho = 0.997$. The initial variance is given by non-informative priors $\alpha_0 = 1$ and $\beta_0 = 1$. The VB algorithm is then iterated 5 times during each time step.

After proper initialization, we performed model inversion by SCKS according to the Algorithm 8, where we allowed maximum of 20 iterations, and considered the discretization of continuous model by using local linearization scheme with the time step 1 s. It means we linearly interpolated the observation sequence by factor 2. Note that, we specified the initializations by using variances, but we propagate their square-

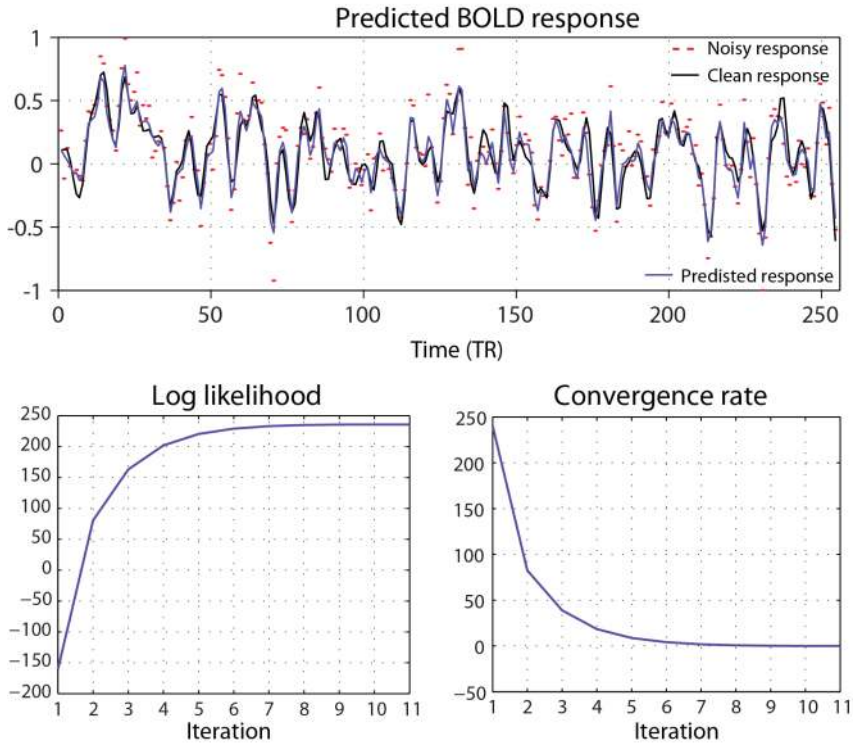


Figure 4.2 Results of single fMRI time course model inversion (part 1.). The upper plot shows the predicted BOLD responses by SCKS algorithm and provides the comparison with the noisy observed responses and the original noiseless signal. The lower plots show the increase of log-likelihood and the decrease of convergence rate, which indicate that algorithm converged after 11 iteration.

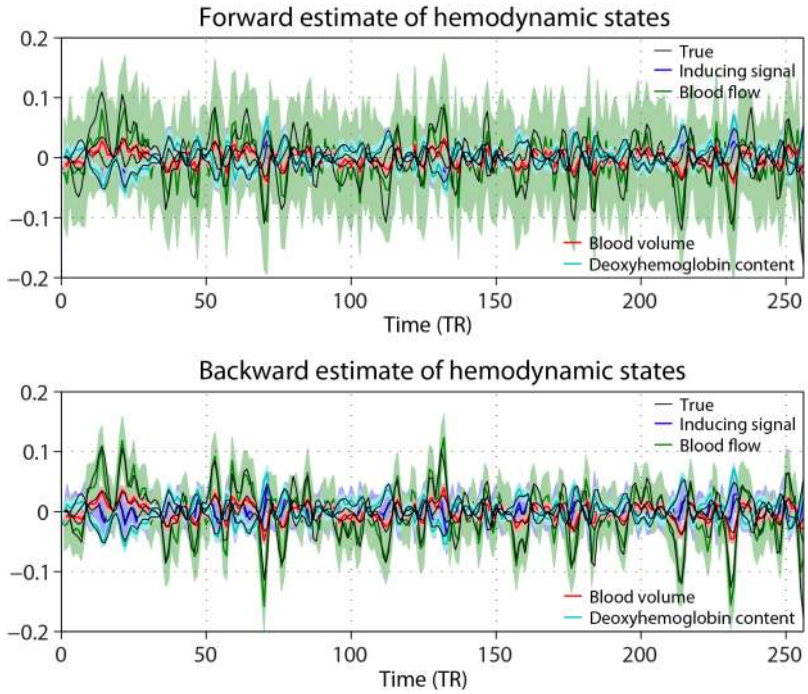


Figure 4.3 Results of single fMRI time course model inversion (part 2.). The upper plot shows the estimates of hidden hemodynamic states as they are provided by forward run of SCKS algorithm. The shaded area represents the 95 % posterior confidence intervals. The lower plot displays the estimated provided by the backward run (smoother). Note that in the later case the confidence intervals are already much smaller.

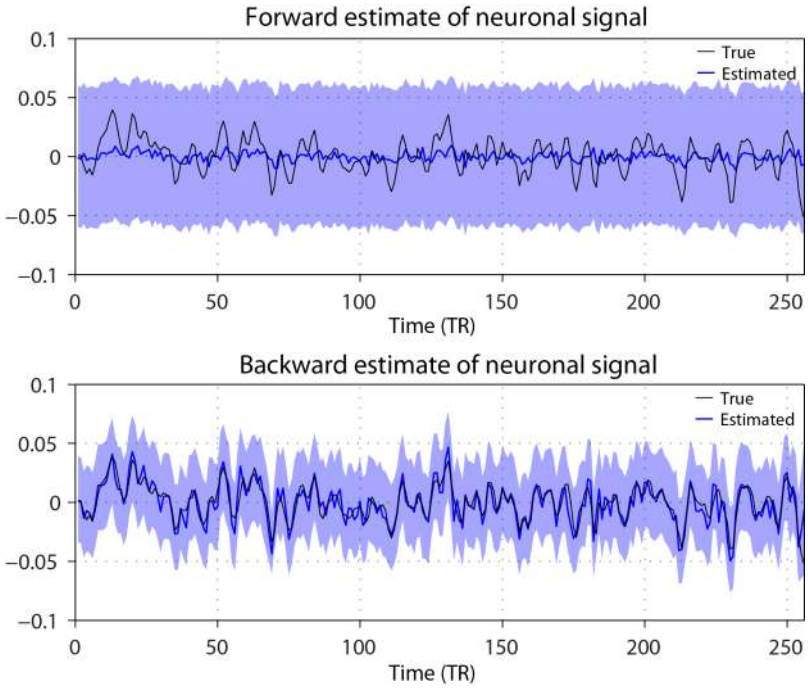


Figure 4.4 Results of single fMRI time course model inversion (part 3.). The upper plot shows the estimates of neuronal signal provided by forward run of SCKS algorithm. In this case there is significant difference between the true signal and the estimated signal. The estimated signal has much lower amplitude is also delayed. The lower plot shows the estimate of neuronal signal provided by the backward run of SCKS, which well much the true neuronal signal.

roots (i.e. standard deviations) through the SCKS algorithm.

Results

In this section we show only the results of the single model inversion. More simulations will be performed for a multivariate estimation, which is the main focus of this work. In Figure 4.2 we can see a prediction of BOLD responses compared to the noisy and to the clean BOLD signal. Due to employed estimation of measurement noisy we are not overfitting and the prediction corresponds well to the clean signal. In this case we have reached the convergence with 11 iterations, as it can be seen from the plots of log-likelihood and convergence rate. In Figure 4.3 we display the estimates of hemodynamic states as they are delivered by forward run of the filter and backward run of the smoother, respectively. Clearly, by performing the smoothing, the estimates are more correct and confident (narrow confidence interval around the posterior means). More importantly, when looking at the results of estimated of neuronal signal, we can see that only forward run is unable to recover the true neuronal signal correctly. Therefore, it is now obvious why one has to employ also backward run. By applying it we receive a correct estimate with much narrower confidence intervals.

4.2.2 Identifiability of hemodynamic model parameters

Model identifiability is a property which has to be satisfied in order to make the parameter inference possible. This is equivalent to saying that changes in parameter values must generate sufficiently different joint probability distributions of the observed variables [96]. In some cases, the model is unidentifiable, but it is still possible to learn the true values of a certain subset of the model parameters. In this case we say that the model is partially identifiable. In other cases it may be possible to learn the location of the true parameter up to a certain finite region of the parameter space, in which case the model is set identifiable.

Considering the nonlinear hemodynamic model, we know based on the previous results that the parameters are only set identifiable [158] or partially identifiable in very artificial environment; i.e. almost noiseless data and very sparse time distribution of the responses having higher temporal resolution [20]. In addition, a model identifibility also depends on the optimization procedure that is used, and whether a deterministic or

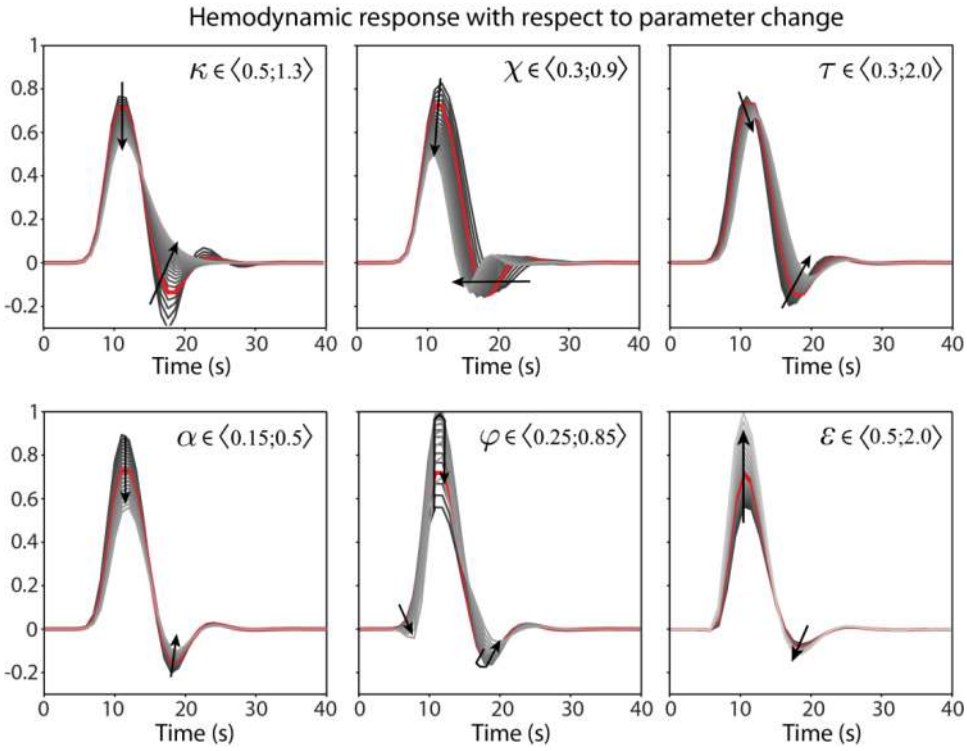


Figure 4.5 The effect of parameter change on output hemodynamic response. For each parameter, the range of values considered is reported, comprising 21 values.

stochastic representation of the model is considered. In our case the situation can be further complicated by the fact that we assume the model input to be unknown; i.e. that all hemodynamic changes are generated by endogenous activity. One can then also expect that the effect of some hemodynamic model parameters can be partly compensated by the estimated neuronal signal [20]. Therefore in our analysis, where we have a stochastic model and the aim is to estimate the neuronal signal, we want to ask a question: which part of the parameter space is best identifiable, and which set of parameters when estimated provides the best estimate of neuronal signal.

The hemodynamic model is mainly described by a set of six parameters $\boldsymbol{\theta} = \{\tau, \kappa, \chi, \varphi, \alpha, \varepsilon\}$, where the effect of the parameter change on the resulting hemodynamic response is shown in Figure 4.5. There we can see that some parameters influence the shape of hemodynamic response in quite similar manner; e.g. parameters α and ε. To find a subset of model parameters that provides the best model estimate, we search the model space consisted of all possible parameter combination. In order to make this space smaller, we focus only on parameters $\{\tau, \kappa, \chi, \varphi, \alpha\}$. The parameter ε, which is the ratio of intra- and extravascular signal and which takes the place in BOLD observation equation (1.7) is selected beforehand as an important parameter that should be included

in our final set [30]. Then the complete model space we search includes 32 different variants (see the bottom of Figure 4.6). For model inversion and parameter estimation by SCKS, we use the same principle of scaling the empirical parameter prior values through log-normal density function as described in the previous section. Each model inversion is then initialized with scaling parameters having zero mean and error variance $5 \cdot 10^{-3}$ ($1 \cdot 10^{-4}$ for ε). This entire process is repeated 20 times, when the time courses are generated according to the description provided in the previous section.

After inverting all models (i.e. together 640), we need to choose a criteria that we apply to find the most proper set of parameters. A very basic but helpful criterion is an assurance of increasing log-likelihood with each subsequent iteration. For this condition we considered maximum of 20 iterations during each model inversion and no threshold for convergence rate under which the estimation process is stopped (except divergence). From Figure 4.6 we can see that actually many parameter combinations led to the divergence during the optimization procedure (grey filled circles). Therefore, only those models (parameter sets) that were able to pass all 20 iterations were further considered (6 of total 32). As a second criterion we chose BIC representing model fit in general. From this point of view, the best fit is provided by model number 16 and 5. However, these still might not be the right choice, because if we look at the third measure represented by root-mean-square error (RMSE) between estimated and the true neuronal signal averaged over all 20 simulations, we observe that model 16 and 5 actually have the largest errors. As the last criterion we use RMSE between hemodynamic response reconstructed by the set of estimated parameters and hemodynamic response generated by the true parameters. These three measures map the trade-off between an accuracy of predicted responses and estimated neuronal signal. Since, our main goal is to estimate the neuronal signal and later infer coupling parameter describing the neuronal interaction model, at the end we choose the model 11 (red filled circles), where the parameters rate of signal decay (κ) and the resting oxygen extraction fraction (φ) are estimated.

Although, one could always perform more careful analysis of parameter space, e.g. by using profiles of log-likelihood [20] or based on scaled Fisher information matrix [96], our analysis provides sufficient indication that fits to our approach using SCKS algorithm. These two parameters plus ε will be used in this work to describe a variability of hemodynamic response. The rest of parameters is considered fixed to their

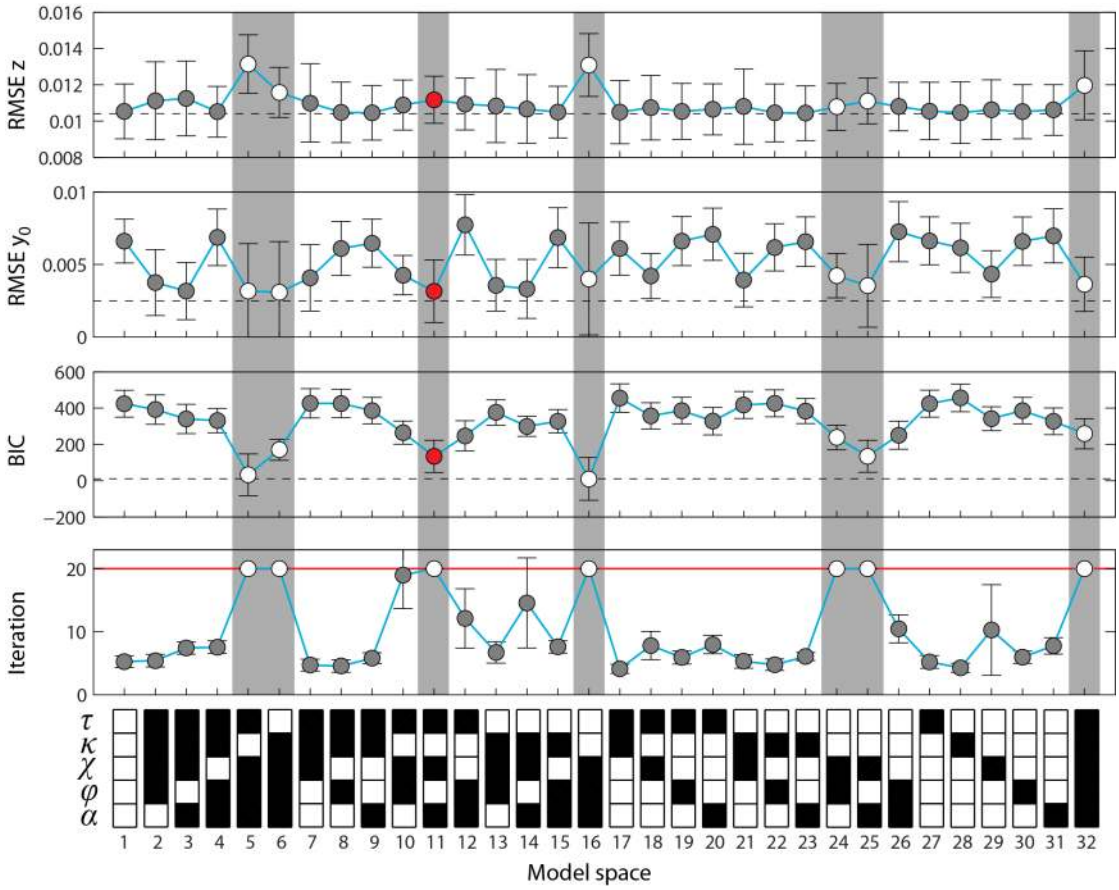


Figure 4.6 Results of hemodynamic model identifiability analysis. At the bottom we display the model space comprising of 32 different combinations of model parameters (white color means the parameter is estimated). From the bottom, in the second row we display average number of iteration after which was the estimation terminated. Here, only parameter combinations, which always allowed to reach the maximum number of iterations ($i = 20$) are considered as suitable (white circles). The third row represents the average BIC, with 95 confidence interval. Similarly, the fourth and fifth rows represent the average RMSE errors between theoretical response and the one reconstructed using estimated model parameters, and the RMSE errors between the true and estimated neuronal signals, respectively. The optimal parameter combination (as a trade-off between model fit and accuracy of the neuronal signal estimate) is highlighted with the red circles.

empirical prior values. Finally, we should note that the hemodynamic model becomes more identifiable with higher temporal resolution and less noisy data.

4.3 Stochastic DCM

The previous section provided a demonstration of model inversion using SCKS algorithm for a single fMRI time course based on which we have made an idea of how to setup the optimization, mainly how to initialize the transient priors on model

parameters and noise statistics in order to obtain an adequate estimate of neuronal signal. The current section will extend this demonstration to multivariate case with the aim to estimate the coupling parameters of effective connectivity. This effectively means that we will pursue a validation of SCKS based stochastic DCM.

Besides the general validation of model inversion and model selection, this section will discuss important topics such as: an effect of noisy data; effect of sampling period; variability of hemodynamic response function among different brain regions; possible confusion of causality by an influence of the third (missing) region; and application to the larger networks.

Similarly as in the previous section, we will assess the performance of the proposed algorithm through the Monte Carlo simulations.

4.3.1 Inversion of sDCM

To evaluate a standard inversion of sDCM by SCKS we will consider generation of resting-state data as described in [15]. We will be especially interested to identify how well we can estimate the coupling parameters among the network nodes and if we can correctly select the underlying network structure. In the later case, we will test the performance of *post hoc* Bayesian model selection after inverting the full model, where all connections are allowed; i.e. using the concept of reduced free-energy as an approximation to marginal likelihood. Additionally, we will also show how the correct model can be selected by the second approach, when the marginal likelihood is approximated by BIC.

Simulations

Let us consider a synthetic resting-state fMRI time courses that were generated through the network of four nodes. The resulting four time courses were acquired with a sampling period 3 s (i.e. TR = 3 s) and consist of 256 time points. To simulate these data we used the following generation process. First, the neuronal fluctuations were generated independently for each node of the network by smoothing a sequence of Gaussian random variables so that they had a variance of 0.02 and a Gaussian autocorrelation function of 6 s. Second, these time courses were used as an independent inputs **u** into the neuronal interaction model (as described in the footnote 4 on page 74),

where the effective connectivity between the nodes was assigned through the adjacency matrix **A** that had for example the following form:

$$\mathbf{A} = \begin{bmatrix} -0.5 & 0.24 & 0 & 0.2 \\ 0.34 & -0.47 & -0.2 & 0 \\ 0 & -0.32 & -0.48 & 0.29 \\ 0.25 & 0 & 0.33. & -0.49 \end{bmatrix}. \tag{4.4}$$

Here, the use of positive and negative coupling parameters between nodes produces the anti-correlated responses commonly seen in real resting-state data. Next, the outputs from the neuronal model, i.e. the neuronal states, entered the hemodymamic models involving the hemodynamic states (i.e. inducing signal, normalized blood flow, volume, and deoxyhemoglobin content), which were described by region (node) specific set of parameters. In particular, these parameters were set to their usual prior (Table 1**.**1) and scaled with a deviate sampled from normal distribution, $\boldsymbol{\theta}^h = \boldsymbol{\theta}_0^h \exp(\mathcal{N}(0,0.01))$. This produces a small interregional variability of hemodynamic responses as depicted in - Figure 4.15. Further, we considered an integration step equals to 1 s for discretization of neuronal and hemodynamic model equations. Finally, from the hemodynamic states we generated the BOLD signals corresponding to each node in the network. These were then downsampled to obtain the required TR = 3 s temporal resolution. Since we consider our generative model as a stochastic system, we also added a small Gaussian random fluctuations of variance $3 \cdot 10^{-4}$ to all hemodynamic states (producing a maximum signal change of about 1 % in the fMRI signals), and also added Gaussian noise with variance 0.025 to our observations. This was chosen to obtain signals with SNR ~ 2 (clean signals had variance about 0.1). Additionally, all additive fluctuations for both states and observations were considered slightly correlated with the width of Gaussian autocorrelation function equals to 1.5 s [15]. Remember that if not stated otherwise we will consider the same generation process also in the later experiments.

These simulated data were then used for model inversion by using SCKS, where we assumed the following initialization. The initialization of neuronal, hemodynamic states and hemodynamic model parameters was basically the same as for a single time course, only extended for multiple time courses. In multivariate case, we had to additionally specify structured priors and corresponding transient priors on coupling parameters in the adjacency matrix. These were chosen according to the description provided in the Section 3.2.2, to allow for full model inversion; i.e. with no restriction on connections.

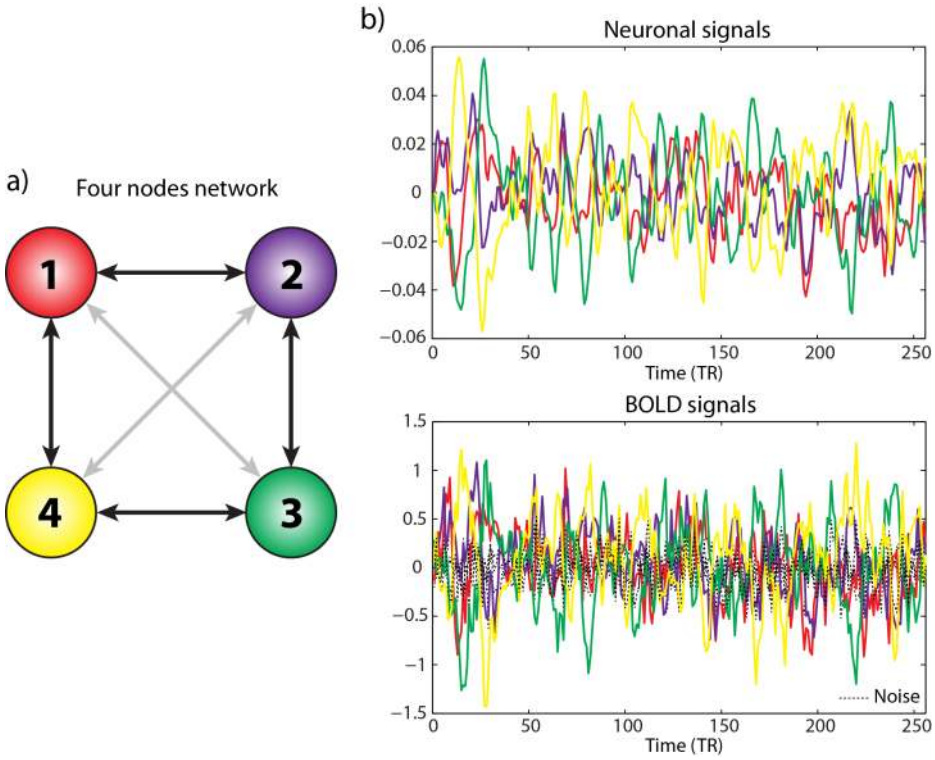


Figure 4.7 Example of simulated connectivity network. (a) Four nodes network with the connectivity structure given by adjacency matrix (4.4). (b) Neuronal signals generated through this dependency network structure, where the color of the time course corresponds to the color of the network node. The same color code is used to display the output BOLD signals.

Also the hyperparameters for the estimation of measurement noise covariance were the same as in the previous case.

Finally, we repeated the simulations and model inversion described above 100 times, where we allowed for maximum of 16 iterations. Data were generated by sampling the coupling parameters from a uniform distribution $a_{ij} \sim \mathcal{U}(0.15, 0.35)$ and switching the sign of reciprocal connections randomly (but still making sure that the adjacency matrix is stable). Connections were then eliminated using an adjacency matrix selected from the model space as shown in Figure 3.4. Here we considered only sparse structures of adjacency matrix having at least one pair of connections switched off and not less than three pairs of connections switched on. Self-connections were sampled from $a_{ij} \sim \mathcal{N}(-0.5, 1 \cdot 10^{-3})$. Candidate simulations were discarded if the simulated data exceeded 2% BOLD signal change.

Besides the MC simulations we also tested the model selection based on BIC and AIC. In this case, we generated data only using a single model of adjacency matrix as

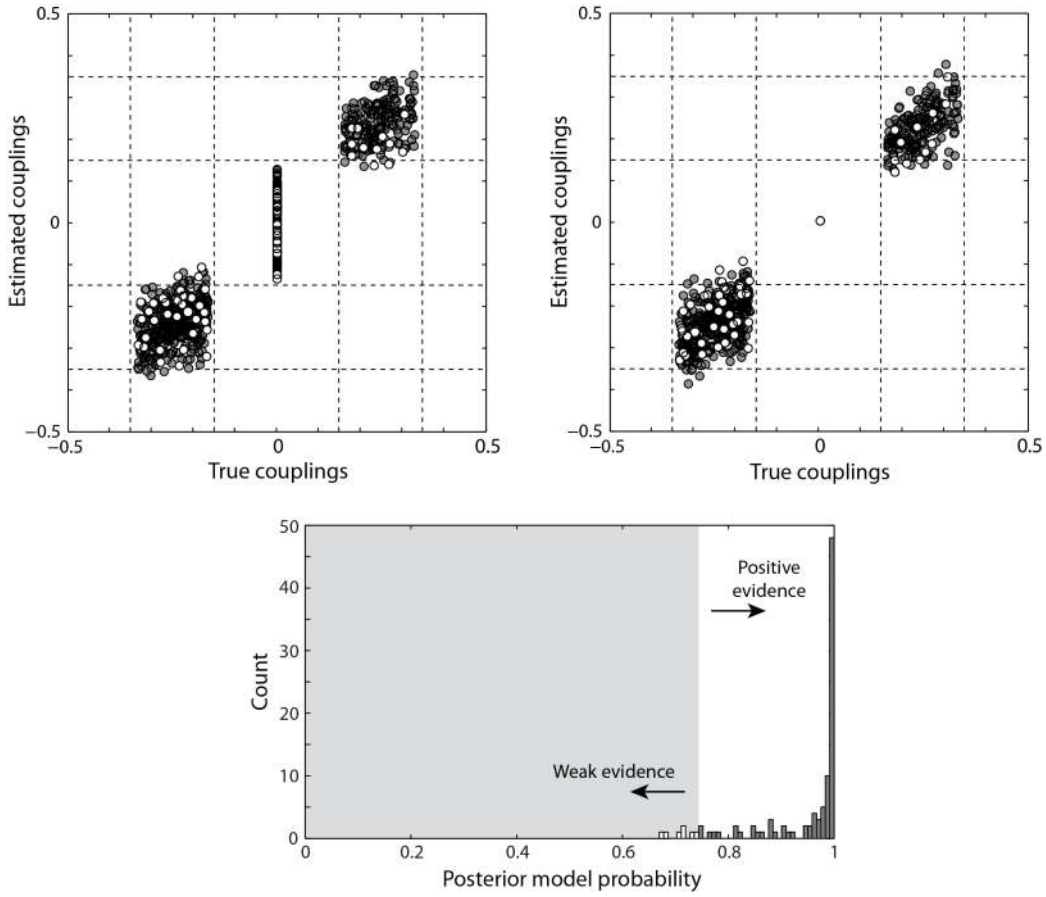


Figure 4.8 Results of sDCM model inversion based on Monte Carlo simulations. The upper plots display the estimates of coupling parameters (i.e. elements of adjacency matrix) for all 100 full model inversions. On the left side we show the results before application of *post hoc* BMS, and on the right side are the results obtained after BMS. The lower plot shows the histogram of posterior model probabilities (corresponding to the best model) for all 100 simulations. Here the models (and corresponding coupling parameters) with the weak evidence are displayed using white color.

shown in (4.4). Then we considered a model space with bidirectional connections as depicted in Figure 3.4, i.e. together 64 possible models. Each of these models is uniquely defined through the initialization of the priors, where the prohibited connections are set to have prior means and variances equal to zero. Finally, each of these models was inverted individually and scored by BIC and AIC measures. Note that in this particular case, we did not apply the pruning step during the model inversions.

Results

The results of the MC simulations are summarized in Figure 4.10. There the estimated coupling parameters are plotted against their true values for the full model (top-left) as they were identified by SCKS and for the reduced model as chosen by *post*

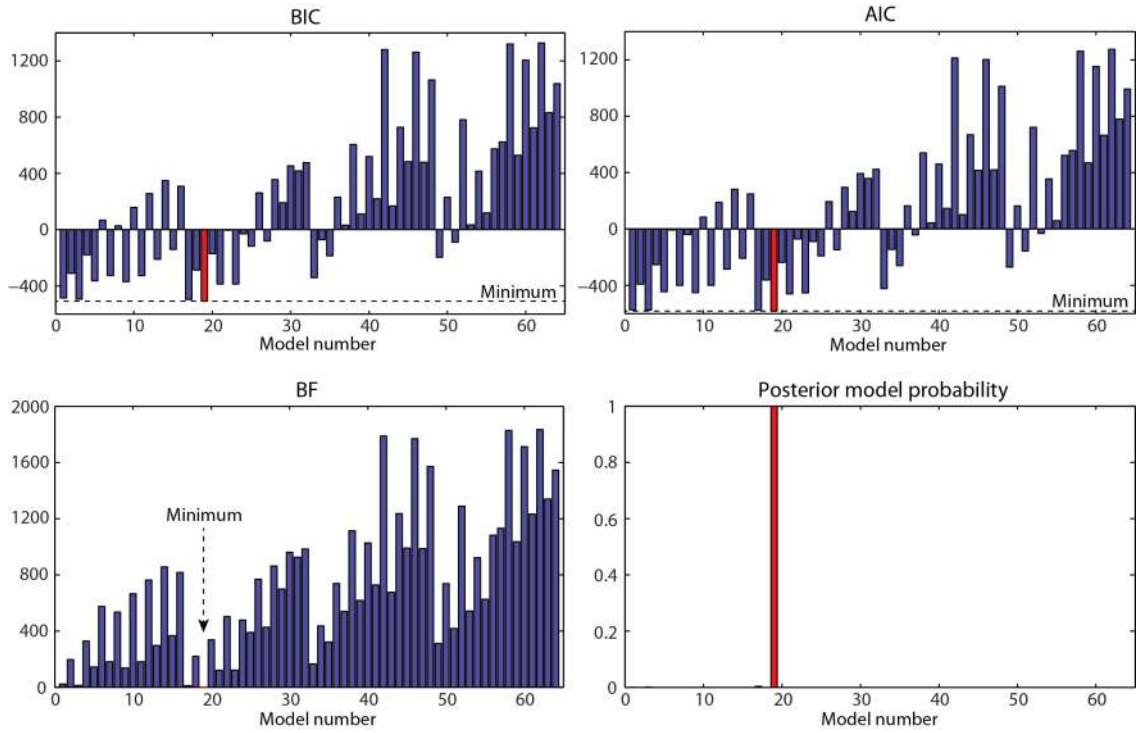


Figure 4.9 Results of sDCM model inversion using BIC for model selection. These results were obtained by individual inversion of all possible restricted models (i.e. 64 models for 4 nodes). The upper plots show difference between scoring models by BIC and AIC. In this case they are very similar, selecting the same model 19. The lower plots display Bayes factors (left) based on approximation of marginal likelihoods using BIC, and posterior model probability (right), which shows that the model 19 was selected with 100 % evidence.

hoc BMS. By comparing these two results, one can see characteristic (vertical) shrinkage of subset of parameters that were simulated as zeros to their prior mean and also some weight changes in "active" coupling parameters, which make the parameter more aligned around the diagonal. Note that *post hoc* BMS always selected the true model as it was assigned to simulated data. Moreover, it can be seen that the estimated couplings are mostly placed within a square space, which corresponds to the range of uniform distribution from which the true couplings were generated for a data simulation. The lower plot then shows the histogram of posterior model probabilities by which the true models were scored. One can see that a majority of models (about 93 %) were selected with positive, strong and very strong evidence, whereas few models (about 7 %) had rather weak evidence. The coupling parameters of these less certain models are distinguished by using white color.

Next, the results of the model selection based on BIC and AIC, where each version of reduced model was individually inverted by SCKS, are depicted in Figure 4.9. In this

case, both BIC and AIC criteria pointed to the same model, which was our simulated true model. By using Bayes factor (where the marginal likelihood is approximated by BIC) to compare all 64 models, we were able to score the true model with a very strong evidence, as it can be seen from the plot of posterior model probability. Although this approach to model selection was also able to successfully select the correct model, we should again emphasize that this approach is limited only to small size networks (up to 4 nodes), otherwise it is unbearably computationally expensive.

Finally, in Figure 4.9 we demonstrate a characteristic process of parameter estimation by SCKS as evolved in time and across 16 iterations (only network coupling parameters are depicted). In order to make a more complete picture about this process, we also show a time evolution for associated parameter error variances, parameter noise variances, and error variances of hidden hemodynamic states (all depicted as standard deviations, since a square-root covariance matrices are propagated through the SCKS algorithm). In these plots, it is especially important to point out the effect of pruning

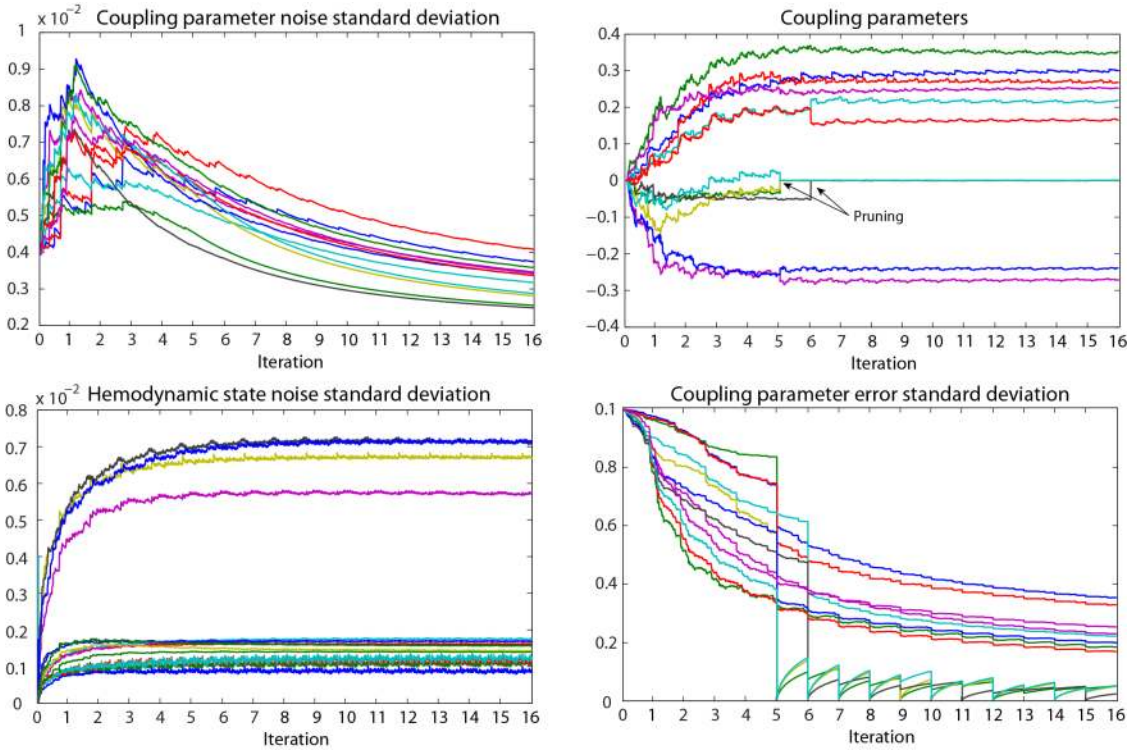


Figure 4.10 Visualization of the estimation process during sDCM model inversion. The upper plots show the time evolution (over consequent iterations) of coupling parameter noise standard deviations (left) and of coupling parameters estimates (given by forward run of SCKS) (right). The lower plots show the same time evolution but for the hemodynamic state noise standard deviations and for the error standard deviation of coupling parameters.

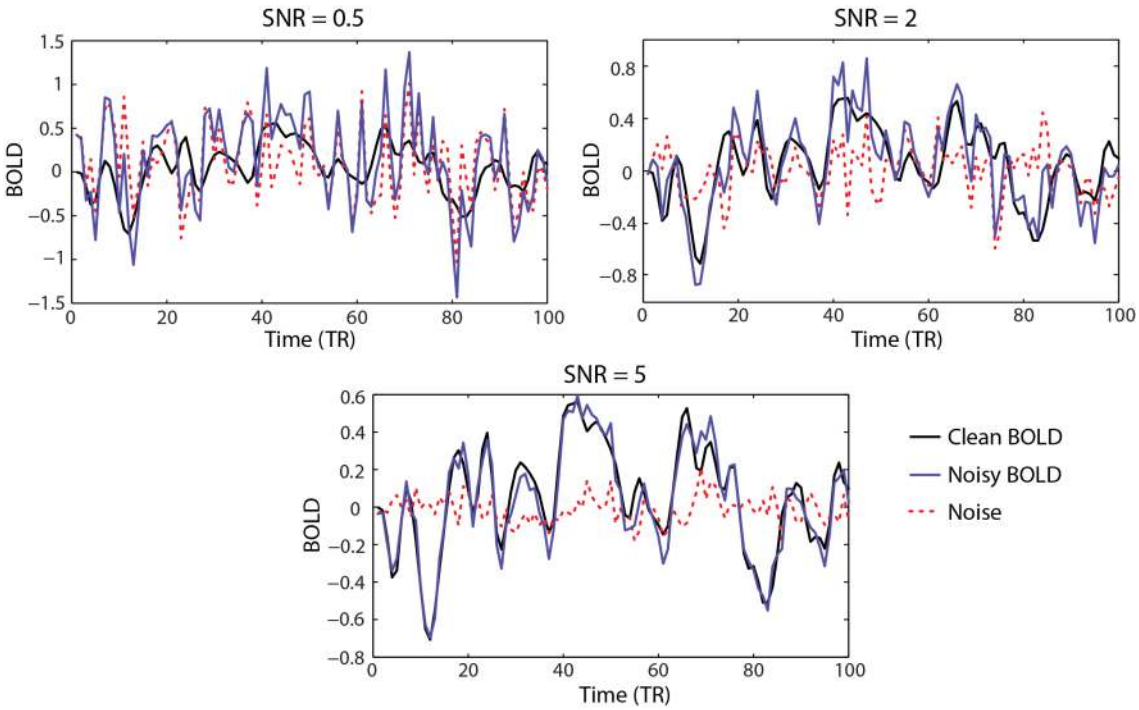


Figure 4.11 Demonstration of BOLD signals with different SNR levels. The examples show a time window (100 time point) of simulated BOLD signals for SNR 0.5, 2 and 5, respectively.

procedure, which started with the 5th iteration, on the parameter estimates and related error variances. We do not display a time evolution for the measurement noise variance, which was also estimated, but one can expect a very similar behavior as depicted in the consequent Section 4.3.2 in Figure 4.13 (right).

4.3.2 Effect of noisy data

It is always interesting to ask a question how robust is the optimization algorithm against the noise in the data. Therefore, in this section we try to investigate how sensitive is the performance of the proposed algorithm for the estimation of effective connectivity to different levels of observation noise. In this case, we use the same definition of SNR as defined earlier, where we specifically consider a simulated data that are contaminated with a very high, high, middle and low level of noise. Since in our algorithm we are estimating the amount of observation noise using VB scheme, it is a nice opportunity to show how this adaptive estimation method contributes to the final result of estimated network couplings and the uncertainty on these couplings.

Simulations

Let us consider a simple network of four nodes with the adjacency matrix as defined in (4.4) and the same generation process of synthetic data as described in the previous section. The only difference is in the amount of additive observation noise, where we used SNR = 0.5 for a very high level of noise, SNR = 1 for a high level of noise, SNR = 2 and 3 for middle levels of noise, and SNR = 5 for low level of noise, respectively. For better imagination, some of these noise levels are shown in Figure 4.11, displaying shorter time window within generated BOLD signal. Next we performed 50 Monte Carlo simulations, 10 for each noise level. Then for each of simulated data-set we considered two scenarios of model inversion by SCKS. The first one performed an uninformed model inversion of the full model with unknown variance of observation noise and the second one performed an informed model inversion with known model structure (only effective couplings are estimated) and with known variance of observation noise. Obviously, for the second scenario we did not have to employ the pruning procedure. Consequently, for the first scenario the best model was selected by *post hoc* BMS. Note that data for each repetition were generated with the same adjacency matrix, but with different sequences of neuronal signals and different random sequences of additive noise (slightly correlated as described earlier).

Results

Since we used the same adjacency matrix for all simulations it suggests to have a closer look at the estimates of particular coupling parameters. The results are summarized in - Figure 4.12, where we purposely skipped the results for self-inhibitory couplings because they did not change much from their prior means. Also, because in almost all cases the *post hoc* BMS selected the true model, we do not display the couplings that were correctly identified as irrelevant; i.e. equal to zero. By saying this, there were two models where during the model inversion of the full model by SCKS, the pruning procedure incorrectly switched off a couple of coupling parameters (blue dots). This means that we have a few false negative estimates, but no false positive estimates; i.e. none of zero couplings were identified as an effective connection between nodes.

Although, the model inversion followed by the model selection was quite successful for all noise levels and both scenarios, it is evident from Figure 4.12 that for

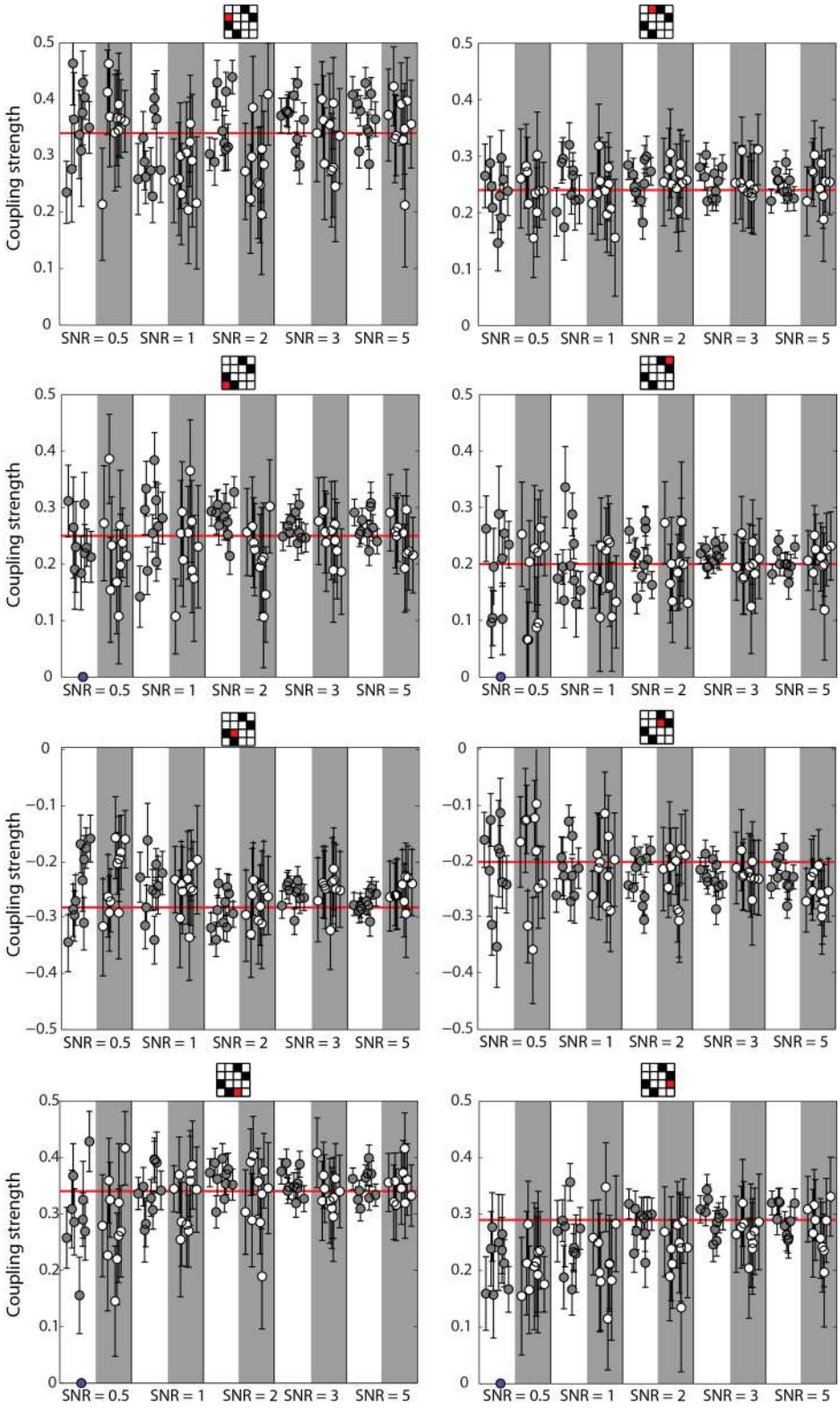


Figure 4.12 Results of SNR experiment (part 1.). Each plot represents estimates for one particular coupling parameter obtained for different levels of noise, where we further distinguish between unrestricted model inversion including estimation of measurement noise covariance (white strip) and restricted model inversion where both correct model structure and measurement noise covariance were known during model inversion (grey strip). The red line is the true value of coupling parameter and the blue dot represents parameter, which was incorrectly switched off during model inversion.

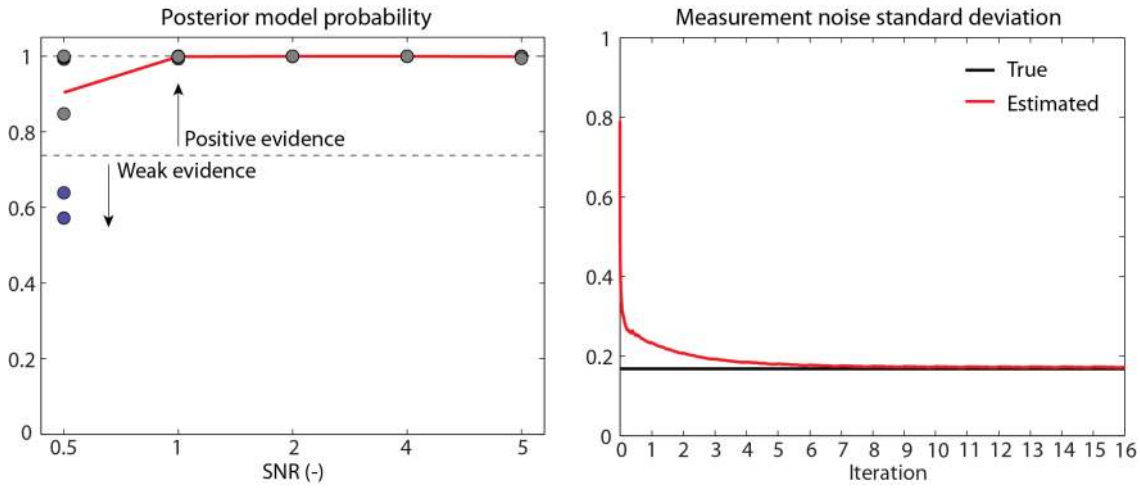


Figure 4.13 Results of SNR experiment (part 2.). The plot on the left side shows the posterior model probabilities that were assigned to the selected models for different levels of SNR (10 simulation for each SNR level). The blue dots represent models having lower evidence, which were incorrectly selected by BMS. The plot on the right side shows an example of estimated measurement noise standard deviation (square-root of variance) as it evolved over successive iterations of SCKS algorithm. We considered the average measurement noise variance for all time courses entering model inversion.

lower SNR, mainly SNR = 0.5 and 1 the accuracy and also the certainty (shown through the error bars) on parameter estimates degrades. Maybe surprisingly, one has to notice that the first uninformed scenario (gray dots on white strips), which was set to perform a full model inversion, including estimation of observation noise variance, provided a significantly better results than the informed one. Specifically, the estimated coupling parameters of the first scenario are much closer to the true values (red lines), with a smaller variability across the MC simulations, and mainly with a smaller error bars, which means that the estimates are more confident. This fact is consistent across all noise levels. Clearly, this demonstrates a great adaptive features of the SCKS algorithm and the ability to accurately estimate the observation noise variance. On the right side of Figure 4.13 we can see the estimation of observation noise variance as it evolved over time and successive iterations. The initialization of observation noise variance is purposely selected as we would expect a very noisy data. This effectively prevents the algorithm to become too confident in early phase of estimation procedure. Also, the dynamic of observation noise are assumed very slow (almost constant), which results in an exponential decay of observation noise variance over multiple iteration steps of SCKS algorithm to actually reach the true value.

On the left side of Figure 4.13 we show the posterior model probabilities for the selected models of the first scenario as they were calculated during the *post hoc* model selection (based on reduced free energy). We can see that except for two models under a very low SNR, which were scored with a weak evidence (these models contain a couple of incorrectly identified parameters as also depicted in Figure 4.12), the rest of the models were correctly selected with a strong and a very strong evidence. This demonstrates a very good sensitivity of *post hoc* BMS against both false negative and false positive estimates.

As a final remark to these simulations, in the real fMRI data one can usually expect SNR between 1 to 3 [26, 159]. However, the fMRI time series for DCM is rather lower in noise (SNR between 2 to 4) since it represents a summary signal of certain region, obtained by using a simple averaging of multiple voxels within the ROI, or more often by using the principal eigenvariate to ensure an optimum weighting of contribution for each voxel within the ROI. On this matter, one should be always very careful in selection of ROIs for analysis with sDCM. They should be smaller in size and should include only significantly activated voxels.

4.3.3 Effect of data sampling period

Another interesting question one can ask is how much is the accuracy of the inversion scheme dependent on the sampling period of the data; i.e. on time series resolution. This is also related to the performance of local linearization scheme that is used to discretize the continuous model. In this section, we perform this test by considering a range of different sampling periods (TR) ordered from small to very high.

Simulations

In order to perform this test we had to simulate data with finer temporal resolution. In this sense we considered integration step 100 ms. The rest of the simulation process was the same as in Section 4.3.1. It means, we also considered a small size network of four nodes with connection structure chosen as defined earlier, and resulting BOLD signals with SNR = 2. Once we obtained simulated data with a higher resolution, we considered five different sizes of sampling periods TR = 0.5 s, 1 s, 2 s, 3 s, and 5 s, respectively, for which we inverted the full model by SCKS. Additionally, we made an adjustment that

for TR > 1 s, the integration step applied in SCKS was selected to match the 1 s resolution, i.e. for TR = 2 s we apply $\delta = 1/2$, etc. Finally, we repeated the entire simulation and inversion procedure 10 times.

Results

The summary of MC simulations is depicted in Figure 4.14. We were especially interested in the effect of sampling period on the accuracy of estimated neuronal signal and coupling parameters. Since in our case we know the ground truth, we calculated the root-mean-square error (RMSE) between the simulated neuronal signals and our estimates obtained by SCKS. There we can clearly see (Figure 4.14, top-left) that with lower TR we receive more precise estimates of neuronal signal, even though the difference is not that high. Similar trend can be seen also if we compare the true and estimated coupling parameter (Figure 4.14, top-right). Here, the difference is even less significant. Although, one would like to see more significant improvement, the results demonstrate a good performance and robustness of the local linearization scheme, which allows an accurate estimation even with high sampling period (by considering relatively large integration step 1 s).

Further, we could have a look at the posterior model probability of the selected models. However, because in all cases (including high TR) the models were always correctly selected with a very strong evidence, we rather examined the effect of sampling period on marginal likelihood approximated by BIC (Figure 4.14, bottom). In this plot we can see considerably increased confidence (reflected in lower value of the BIC) about the estimates for TR = 0.5 s. This is simply because in this case we considered twice the number of time points, but the same number of iterations. If we decrease the number of observation to match the three other cases one can expect almost a linear trend of the improvement.

Overall, we did not notice a very significant improvements in the estimates of coupling parameters by increasing the temporal resolution of the data. This fact can be considered as good news, because it means that the estimation of effective connectivity by DCM formulated in continuous time is not sensitive to the temporal resolution of the data, which is an important advantage over the lag based method for identification of directionality such as GCM, where the temporal resolution plays the crucial role.

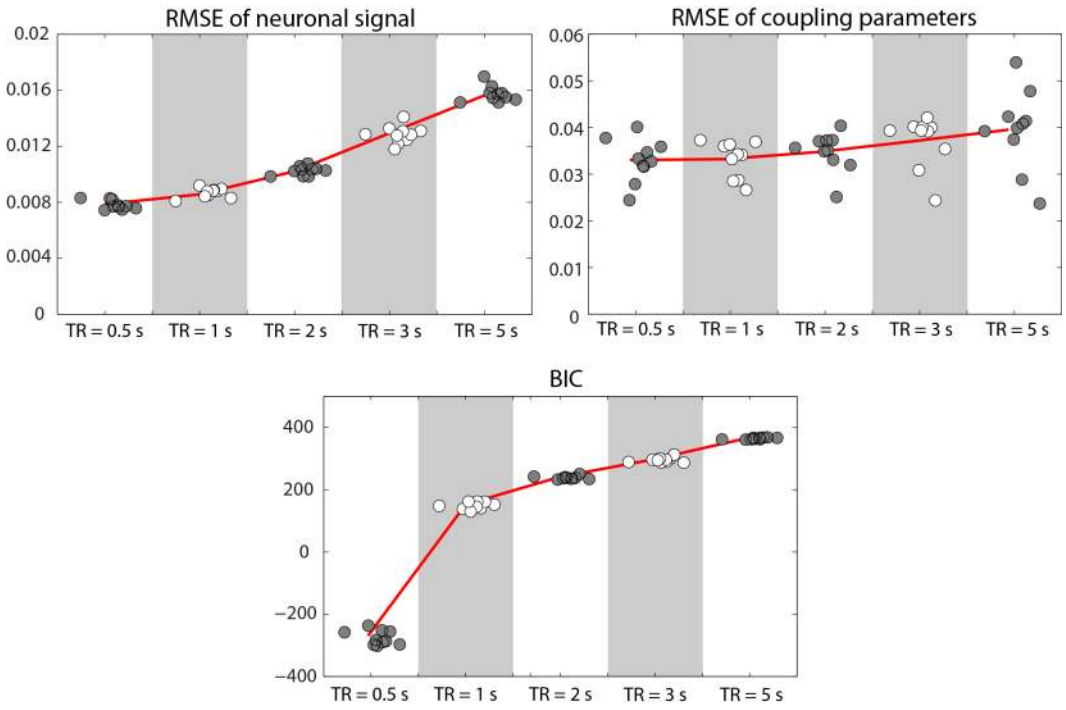


Figure 4.14 Results of sampling period experiment. The upper plots display the RMS errors between the true and estimated neuronal signals (left) and the RMS errors between the true and estimated coupling values (right) for different data sampling periods. The lower plot show dependence of BIC on sampling period.

Nevertheless, we witnessed that the higher temporal resolution of the data (with reasonable SNR) can always bring some improvement.

4.3.4 Effect of interregional variability of HRF

In the introduction of this work, we have emphasized that the main motivation for the nonlinear deconvolution approach (or generally approach based on generative model) is to eliminate a possible confusion in identifying directional connectivity due to variability of hemodynamic response across the brain regions. So far, in Section 4.2.2 we have discussed an identifiability of the hemodynamic model, and concluded that even though the hemodynamic model is not uniquely identifiable by its complete set of parameters, we are still able to recover the underlying neuronal signal with a satisfactory accuracy. In other words, the identifiability issue does not seriously hurt the model inversion (at least in the case of fully stochastic treatment of the model).

What remains to be answered is whether we can perform a model inversion for multiple time courses and simultaneously identify the true effective connectivity at the neuronal level if the shape of hemodynamic response significantly varies among the

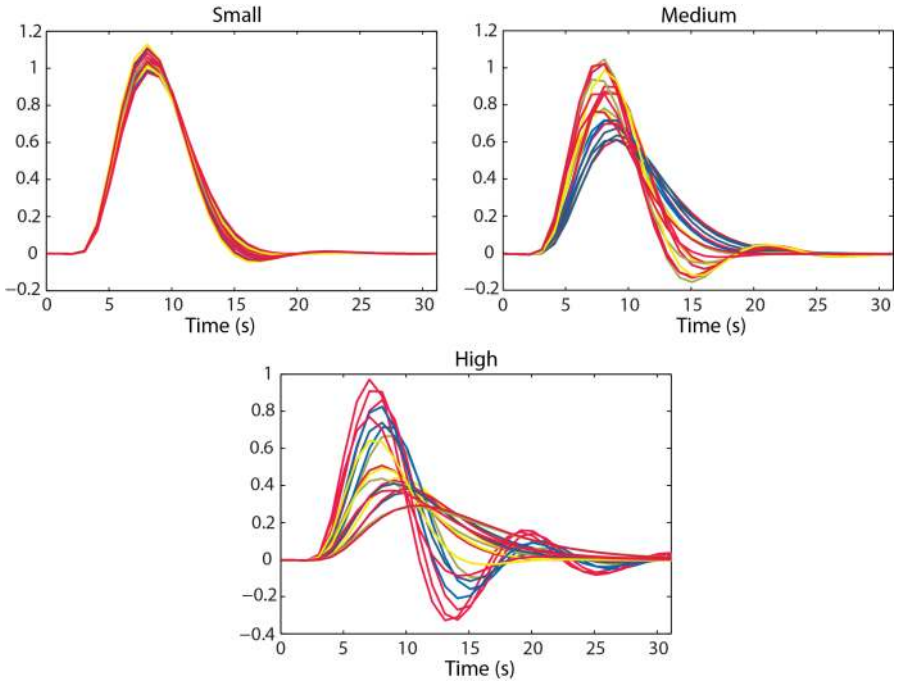


Figure 4.15 Demonstration of different levels of hemodynamic response variability.

brain regions. Therefore, in this section we test the performance of the proposed algorithm in these clearly difficult conditions.

Simulations

For these simulations we considered 5 nodes network with adjacency matrix similar to the one described by (4.4), but with an additional node connected to the fourth node through bidirectional coupling. Further we used the same generation process as in Section 4.3.1., but we considered four different levels of hemodynamic response variability assigned to different nodes (regions). This variability was obtained by scaling the empirical parameter priors (Table 1.1) using log-normal density function $\boldsymbol{\theta}^h = \boldsymbol{\theta}_0^h \exp(\mathcal{N}(0, \vartheta))$, where the variance ϑ was equal to 0 for the case of no variability; 0.01 for a small variability; 0.1 for a medium variability; and 0.5 for a large variability (see Figure 4.15). The resulting BOLD signal had the temporal resolution TR = 3 s and the SNR = 2. The simulation and model inversion was repeated 20 times for each level of hemodynamic response variability (considering the same structure of adjacency matrix).

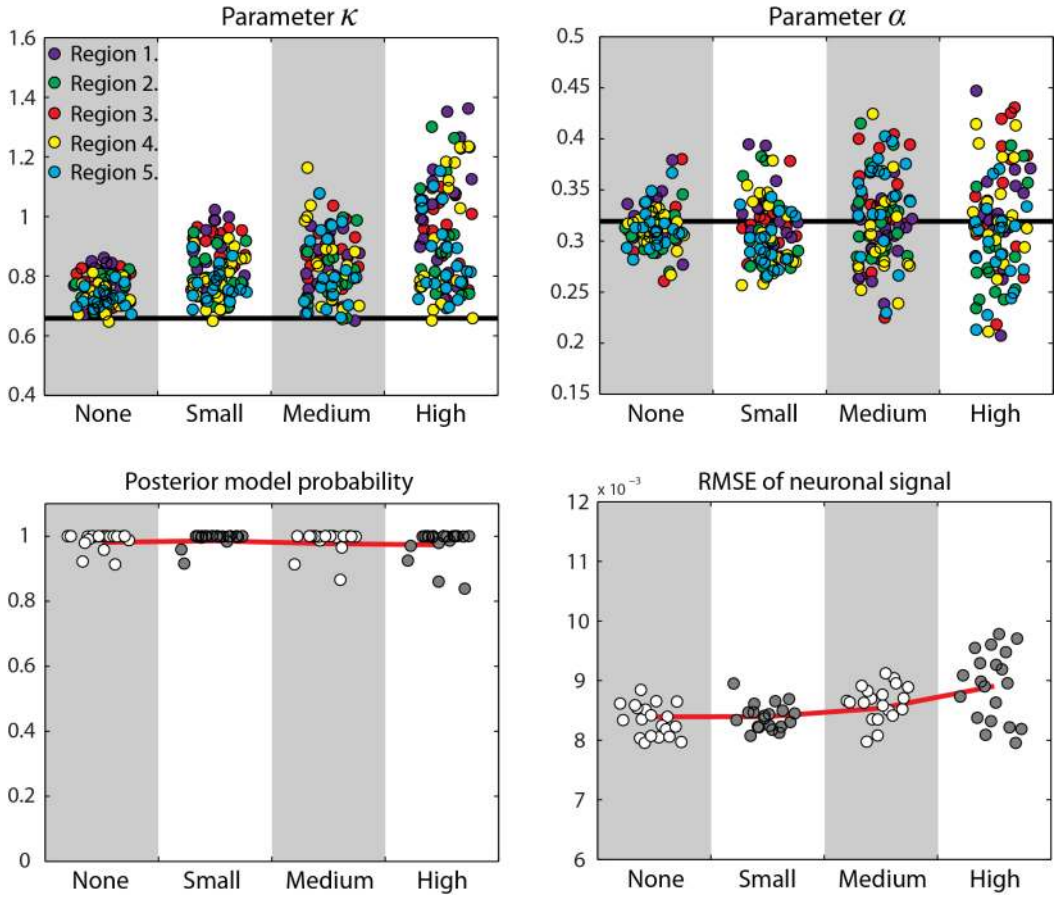


Figure 4.16 Results of interregional variability experiment. The upper plots show the effect of increasing variability of hemodynamic response among brain regions that are considered for model inversion, on hemodynamic model parameter estimates. Here we can see that with increased variability, the variability of estimates increases as well (note that this parameter estimates does not provide unique solution). The lower plots show the dependence of posterior model probability and RMS errors between the true and estimated neuronal signals on the level of hemodynamic response variability.

Results

The results from all inverted models by SCKS are plotted in Figure 4.16. Since we know that hemodynamic model parameters are not uniquely identifiable, i.e. their effect on the output hemodynamic signal can be interchangeable and also partly compensated by the neuronal signal, we cannot simply test for the error between estimated and the true hemodynamic parameters. However, we considered four levels of hemodynamic response shape variability, therefore we can at least check how this variability is reflected in hemodynamic parameter estimates. This is depicted at the top of Figure 4.16. for parameters rate of signal decay (κ) and resting oxygen extraction fraction (φ), which were selected to fit the shape of hemodynamic response (see Section 4.2.2). In

these plots we can see that by increasing variability in simulated data the variability of the estimated parameter values increases as well. Now, taking a look at the results of posterior model probabilities of selected model (connectivity structure) by *post hoc* BMS (at the bottom of Figure 4.16), we can see that there is almost no difference between no, small, medium and high variability of hemodynamic response among different network nodes. In all cases, we selected the correct model with high posterior model probability. Next, we calculated the average RMSE between the estimated and the true neuronal signals. Here we already experience some decrease in accuracy that comes with the high variability of hemodynamic response, but the increase in the error is negligibly small.

These results prove that the stochastic DCM based on SCKS is not sensitive to variability of hemodynamic response function across different brain regions. This statement is valid even in case of longer sampling period, as it was used in these simulations (TR = 3), and one can expect an improvement by increasing the temporal resolution and SNR of the data.

4.3.5 Third region influence

Probably the highest risk of a false positive identification of causal relationship is caused by spurious correlations between time courses. This spurious correlation represent a mathematical relationship in which two events have no direct causal connection, but they are likely to be wrongly inferred that they do, due to the coincidence or the presence of a third unseen event (region). Let us imagine two simple scenarios of causal influence as shown in Figure 4.17. In the first case, we have a causal chain among three regions where region 1 has a causal influence on region 3, mediated through region 2; i.e. $1 \rightarrow 2 \rightarrow 3$. In the second case, we have a common source represented by region 2, which causes dynamics in regions 1 and 3, and there is no influence between region 1 and 3; i.e. $2 \rightarrow 1$ and $2 \rightarrow 3$. Assuming a bidirectional connectivity, then these two cases are always present. From the previous simulations we know that if all regions are included in the analyzed network than SCKS is able to effectively infer the correct causal relationships. However in the case we neglect some of essential regions such as region 2 (the common cause), then we are facing a "missing region" problem that might lead to causal fallacy. Therefore, in this section we will try

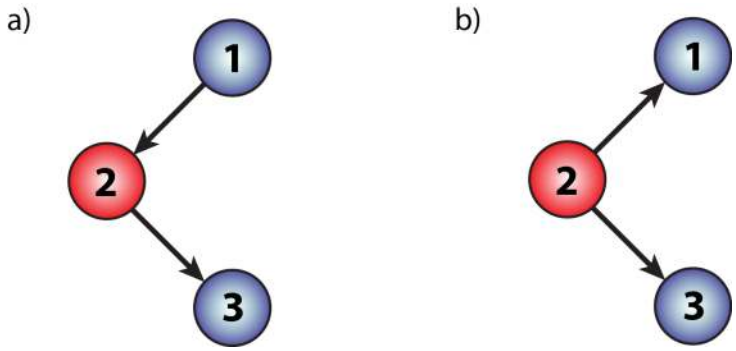


Figure 4.17 Two common connections involving the influence of the third region. (a) The influence from the region one to the region three is mediated by region two. (b) The region two is influencing both regions one and three.

to address a question: what are our chances to identify a correct causal influence in the case of missing (essential) region. Additionally, we would like to know if there are any properties of fMRI time courses, such as sampling period and amount of observation noise, which help to overcome this problem.

Simulations

We generated a network composed of 5 nodes as shown in Figure 4.18, where the first node has relatively strong causal influence on the second and the third node, where we also consider reciprocal connections. There is no influence between the second and the third node. The third node is further connected to the fourth node, which is connected to the fifth node, all through bidirectional connections. Having this particular network, we simulated higher resolution data with integration step 100 ms. Now, in order to test the robustness of the algorithm in case that the common (essential) node is missing, we considered three following scenarios. In the first scenario we invert the complete 5 nodes network. In the second scenario we exclude the time course associated with the first node and invert model having only time courses 2-5. In the third scenario, we consider 3 nodes network, where we further exclude the 5th time course (see Figure 4.18). These scenarios were tested for three different levels of observation noise (SNR = 1, 2, and 3) and for two different sampling periods (TR = 1 s and 3 s). Finally, each particular case was repeated 10 times considering the same adjacency matrix but different innovations.

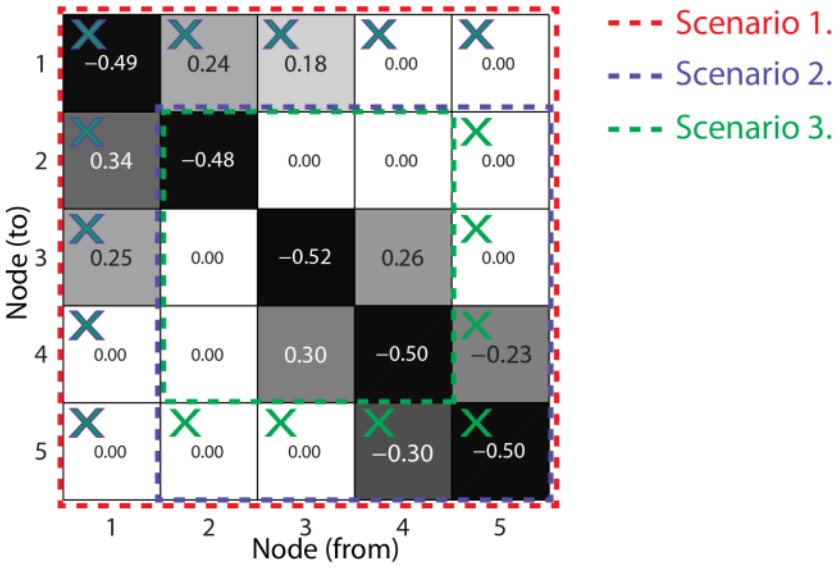


Figure 4.18 Demonstration of missing region problem. The first scenario considers all nodes for model inversion (red dashed line). The second scenario leaves out the first node for model inversion (blue dashed line) and the third scenario leaves out also the fifth node (green dashed line).

Results

The results from the above MC simulations are reported in Figure 4.19, for all three scenarios, where we scored the performance of the SCKS algorithm by comparing the posterior model probabilities of selected models, and by comparing the RMSE between the true and estimated network couplings.

The results obtained during the first scenario, where all five nodes (time courses) are included, do not show any new information that we would not know already from the previous simulations. Clearly, we were able to estimate and select model that is equivalent to the true model, where we can see an improvement in both posterior model probabilities and RMSE for higher SNR. Moreover, with a smaller sampling period, one can estimate the coupling parameters slightly more precisely.

The situation is already more interesting when looking at the results of the second scenario. In this case, by neglecting a common source of influence on 2nd and 3rd regions, we expect a presence of spurious correlations between the 2nd and 3rd region that might possibly hurt the performance of estimation procedure. And indeed, we can see an overall decrease of posterior model probabilities (approximately by 10 %), where among selected models we have several models (blue dots) that were incorrectly identified (with false positive couplings). Nevertheless, a majority of the models were

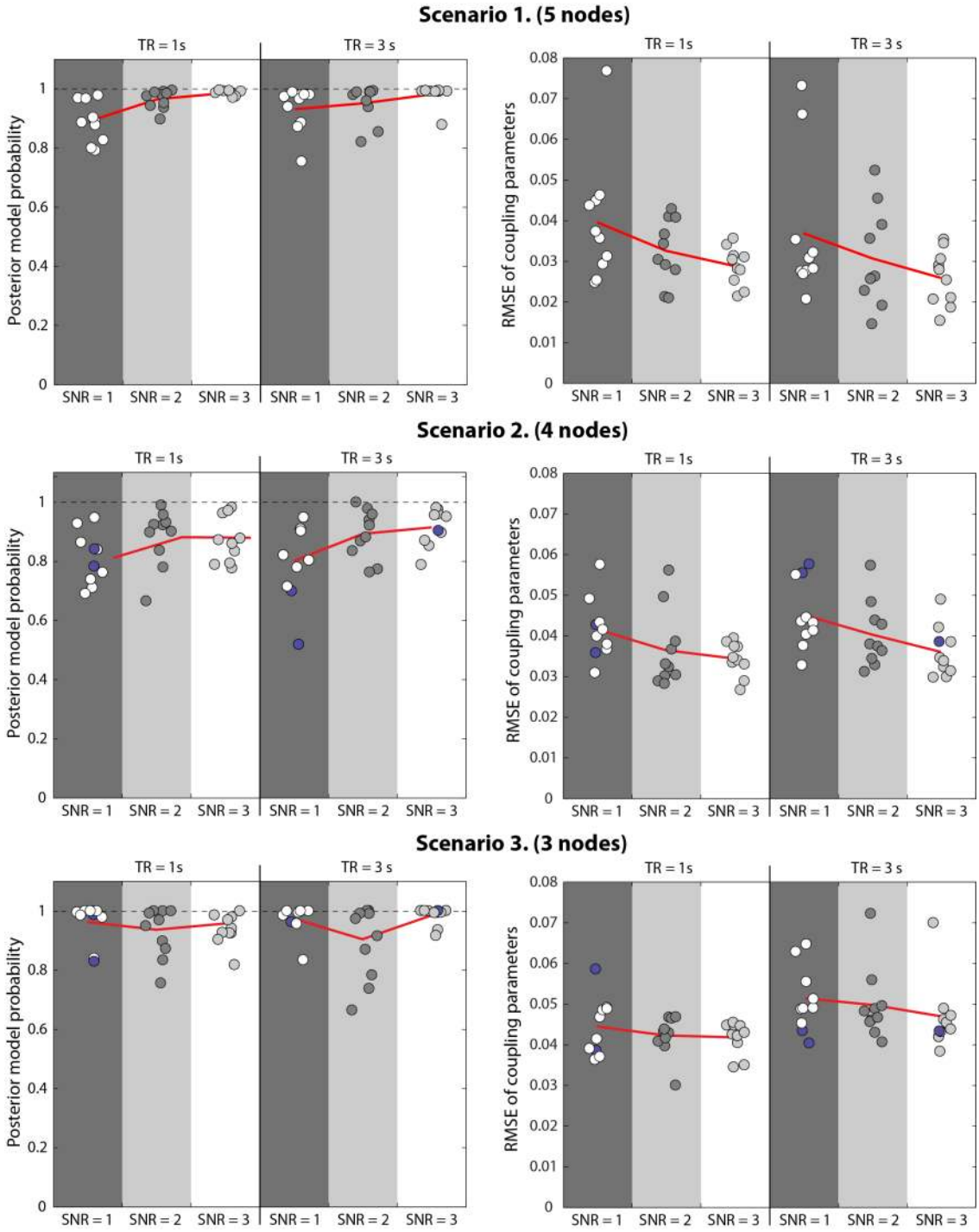


Figure 4.19 Results of missing region experiment. Each row in the figure describes one particular scenario regarding the missing region problem. The plots on the left side show the effect of different sampling period (TR = 1s and 3 s) and the effect of different amount of noise added to data on posterior model probability of selected model. The same effects are then displayed in the plots on the right side, but with respect to the average RMS error between the estimated and the true coupling parameters.

still correctly identified. There is a very small difference between estimates obtained with lower and higher resolution, but increasing the SNR brings an improvement.

Finally, in the third scenario we had similar situation but considered only 3 nodes network. Although in this case we scored models with relatively high posterior probabilities, the number of incorrectly selected models is the same as in the previous case. It means that if we try to identify coupling parameter within a very small network, there will be a higher danger of obtaining a very certain (nonzero) estimates even of connections that are spurious.

Obviously, this test represents only a very limited insight into the entire problem. It can be more dramatic in real data situation, for example when multiple sources are missing, etc. Therefore, we expect further developments to improve the regularization process of parameter estimation, which could more effectively eliminate the spurious coupling estimates.

4.3.6 Larger networks

The problem of missing region described in the previous section could be potentially overcome by covering a higher number of brain regions in analysis of effective connectivity, where we hope that all important regions are included. Considering larger networks is also interesting from the perspective of analyzing resting-state data. In this case one usually first applies methods such as independent component analysis [160] to the whole brain fMRI data, which summarizes a spatially distributed brain activations into a smaller set of components (modes). The relevant components and their associated time courses can be then considered as a nodes of analyzed network by stochastic DCM. Therefore in this section we want to ask a question: is there a bright future of stochastic DCM based on SCKS to by applied also into the larger networks?

Simulations

Because it is not that simple to generate a larger network based on random permutations of connections in adjacency matrix that would result in a stable matrix, we considered an adjacency matrix, which is sometimes called a tandem network. The adjacency matrix of the tandem network has non-zero couplings only on the first diagonal above the main diagonal and on the first diagonal below the main diagonal; i.e. all nodes are connected through bidirectional couplings. This type of network was first created for 8 nodes and later for 16 nodes. The strengths of connections were in both cases generated

randomly from the uniform distribution and from the normal distribution in case of self-connections as described earlier. The sets of fMRI time courses generated in a standard manner (TR = 3 s; SNR = 2) were then used for model inversion by SCKS.

Results

The results for the 8 nodes network are depicted at the top of Figure 4.20. There we can see the estimated adjacency matrix obtained right after the model inversion by SCKS

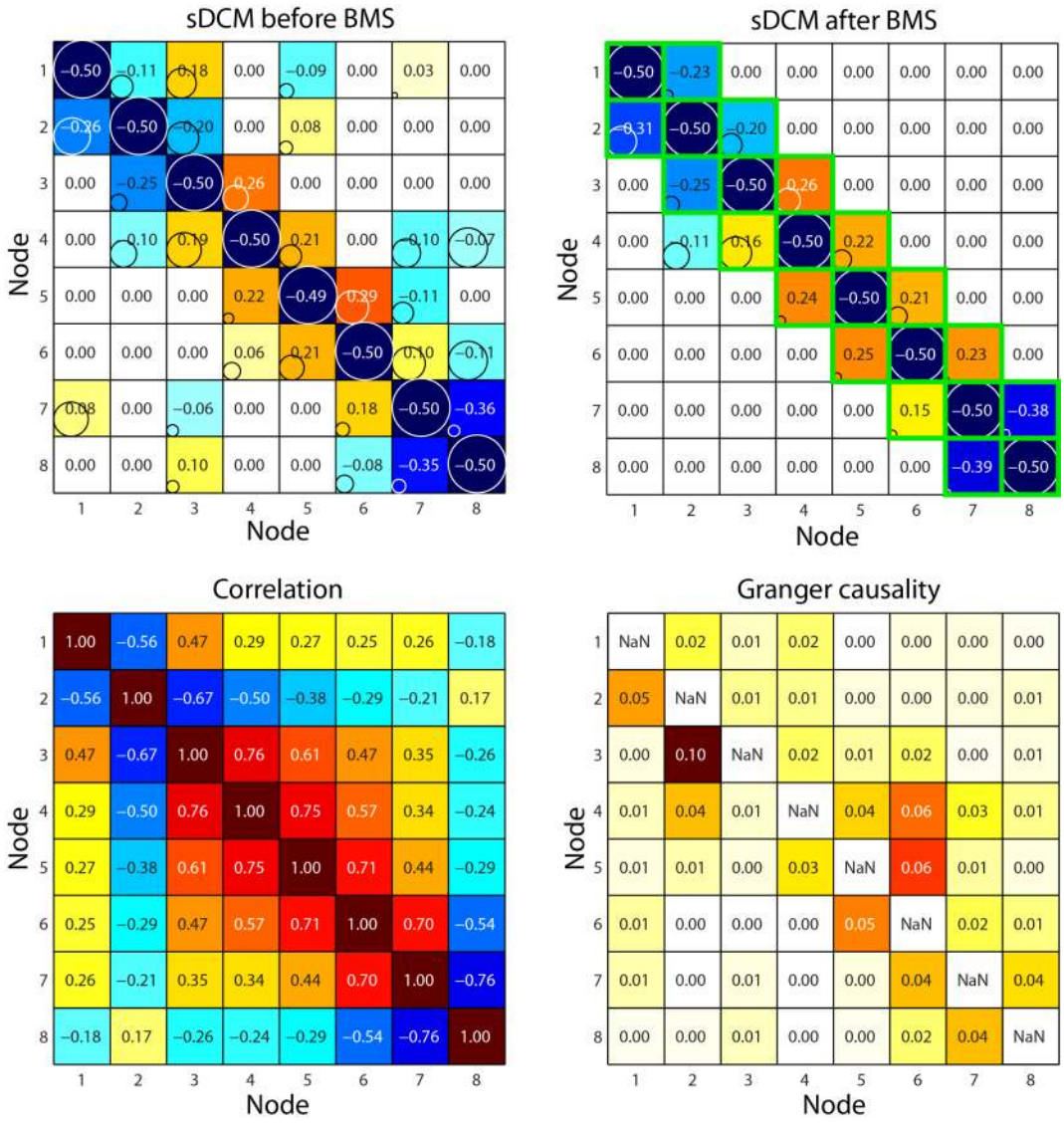


Figure 4.20 Results of lager network experiment (part 1.). The upper row displays connectivity matrices (for 8 nodes) obtained by model inversion using the SCKS algorithm before (left) and after (right) applying model selection based on *post hoc* BMS. The green mesh represents the true connectivity structure used for data simulation. The lower row shows the results that one obtains by applying simple pair-wise correlation between observed BOLD time courses (left) and by applying Granger causality modeling based on standard multivariate autoregressive model of the second order.

algorithm, where many of irrelevant couplings were already switched to zero by the pruning procedure during model inversion. After applying *post hoc* BMS we already received a result that is very close to the true connectivity structure (green mesh), where only one connection ($2 \rightarrow 4$) was wrongly identified. Because we have network consisting of more than six nodes, we already employed a greedy search procedure to perform *post hoc* BMS.

At the bottom of the same figure, we also show the result obtained by pair-wise correlation between BOLD signals. Correlation is very often used as the simplest method for evaluation of functional connectivity (also in resting-state data). The correlation will give us a basic picture about interdependences between time courses, but from the perspective of effective connectivity, many of these interdependences (correlations) are spurious. Critically, just using a correlation with any threshold can never recover the real connectivity structure. As the last comparison we also show the result obtained by Granger causality modeling (GCM) based on simple multivariate autoregressive model, where the order of the model (the lag) was estimated using BIC and AIC (both suggested the model order 2). In this case, GCM fails to estimate the correct connectivity structure mainly because the temporal resolution of BOLD signals is too low (TR = 3). We expect that GCM could do better if higher resolution signals are used and if GCM is formulated as a state-space model [127, 128].

The result of the second network with 16 nodes is depicted in Figure 4.21. There we show already the final result as obtained after *post hoc* BMS. Although there are several irrelevant couplings that were not switched off neither by BMS or by pruning, their strength is smaller than any other correctly identified connections. Additionally, we have one coupling parameter ($12 \rightarrow 11$) that was incorrectly switched off by pruning during the model inversion. Just for comparison we again display correlation matrix capturing the dependences between BOLD signals.

In general, these results show that it will be possible to effectively apply stochastic DCM also to the larger networks. However, the computational demands increases markedly. For example for 5 nodes network one iteration of SCKS takes about 60 s, while for 16 nodes one iteration takes already about 30 minutes[8].

[8] Evaluated on personal laptop HP Pavilon with CPU i7-2630QM at 2.0 GHz and 8GB RAM.

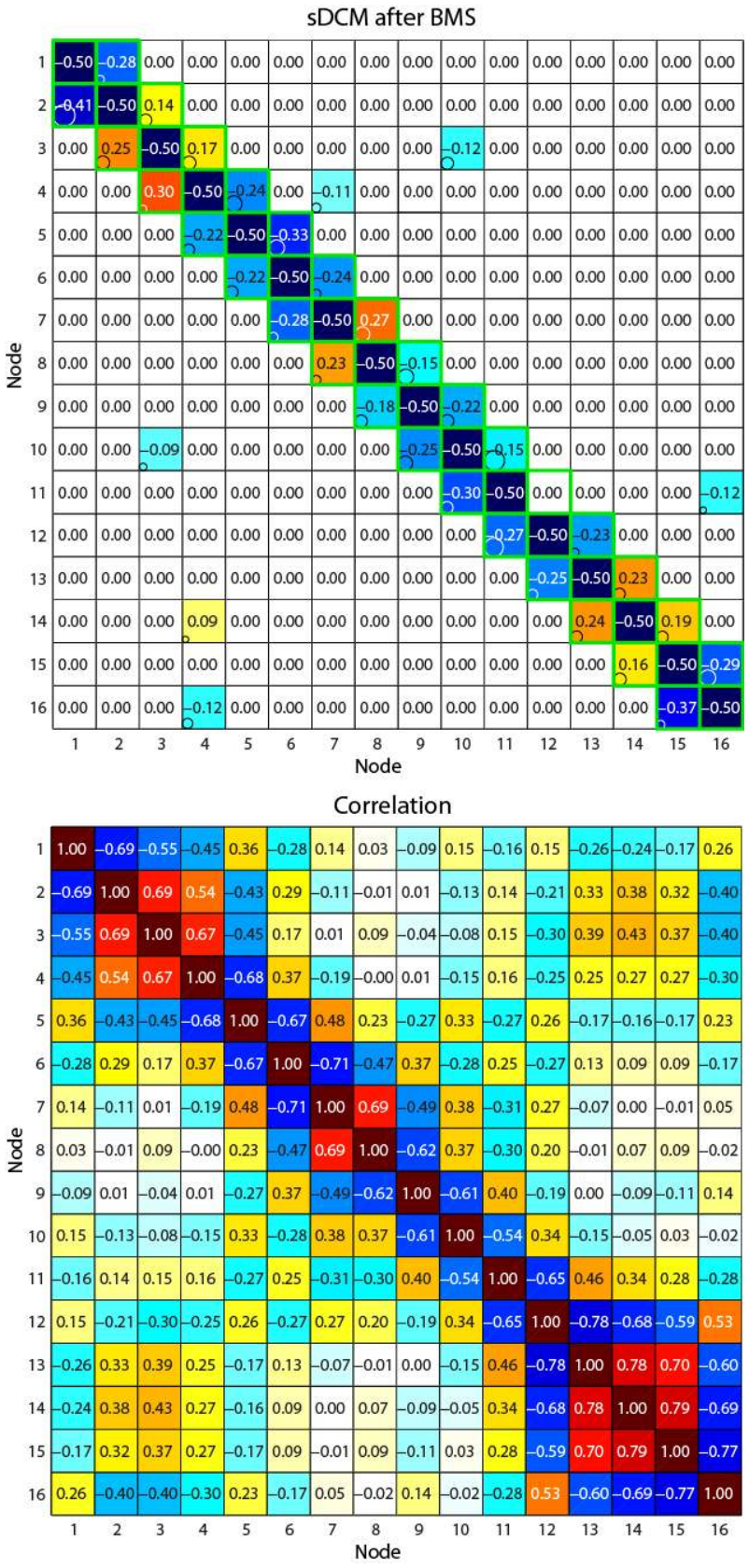


Figure 4.21 Results of lager network experiment (part 2.). The upper panel shows the estimated connectivity matrix (16 nodes) obtained by model inversion using the SCKS algorithm followed by the *post hoc* BMS. The green mash corresponds to the true connectivity structure. The lower panel shows the result obtained by pair-wise correlation.

4.3.7 Analysis of empirical fMRI data

As the last test, we attempt to apply the SCKS algorithm to empirical data. Although, we mostly emphasized on application of stochastic DCM to resting-state data, there is no reason why this approach should not work also with a task data. Further, since there is very little known about effective connectivity in resting state data, we choose to demonstrate the algorithm performance on task data, but under the assumption that we do not know the exogenous (experimental) input. In particular, we apply SCKS to empirical data-set that has been used previously to describe developments in causal modeling and related analysis [11, 15, 17, 123, 142, 161].

Performing model inversion of task data by SCKS without considering the known input is potentially interesting because it allows one to quantify how much neuronal activity can be attributed to the evoked responses (i.e., the experimental design or exogenous inputs) relative to endogenous activity [15]. In what follows, we will briefly describe the data used for our analysis and then report the results.

Table 4.1 Selected regions for sDCM analysis.range.

Name	Description	Location (mm)	Number of ($3\ mm^3$) voxels
VIS	Striate and extrastriate cortex	-12 -81 -6	300
AG	Angular gyrus	-66 -48 -21	51
STS	Superior temporal sulcus	-54 -30 -3	269
PPC	Posterior parietal cortex	-21 -57 66	168
FEF	Frontal eye fields	-33 -6 63	81
PFC	Prefrontal cortex	-57 21 30	48

Data description

Data were acquired from a healthy subject at 2 Tesla using a Magnetom VISION (Siemens, Erlangen) whole body MRI system, during a visual attention study. Contiguous multi-slice images were obtained with a gradient echo-planar sequence (TE=40 ms; TR=3.22 s; matrix size=64×64×32, voxel size 3×3×3mm). Four consecutive 100 scan sessions were acquired, comprising a sequence of ten scan blocks of five conditions. The first was a dummy condition to allow for magnetic saturation effects. In the second, *Fixation*, subjects viewed a fixation point at the centre of a

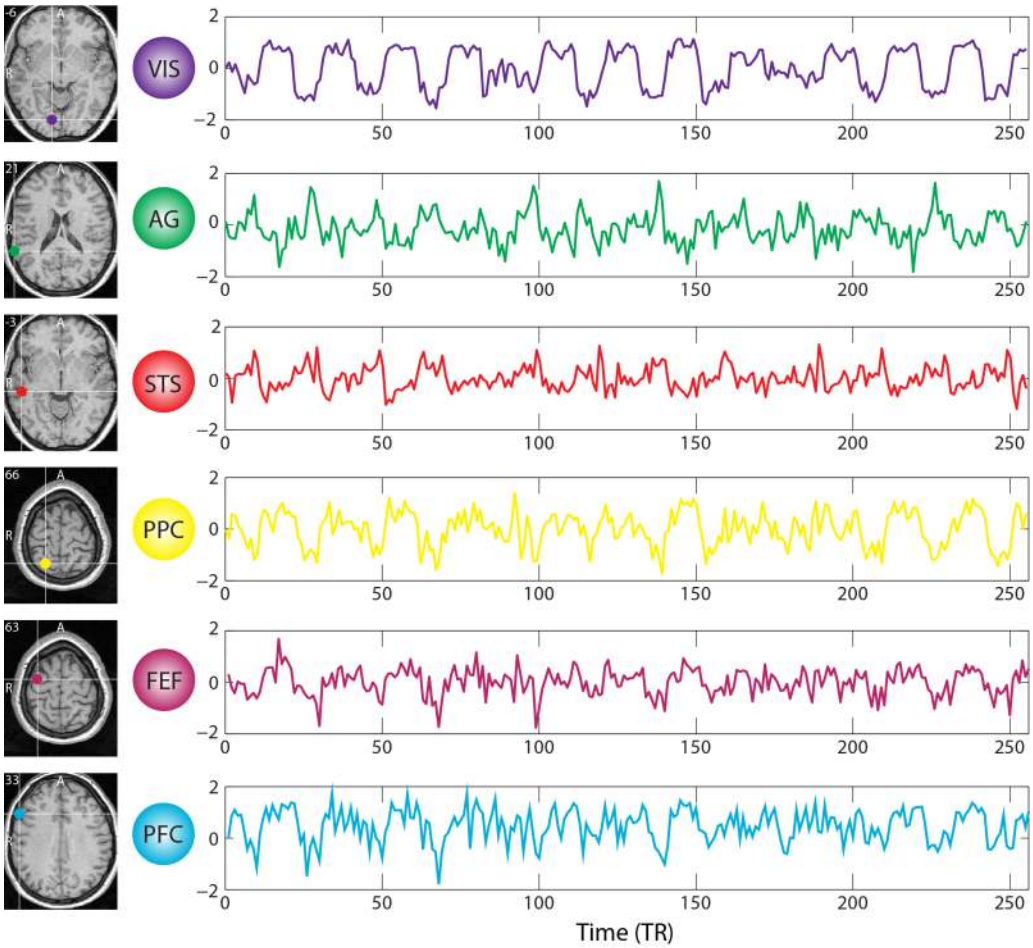


Figure 4.22 Selected regions for sDCM analysis and their associated time courses.

screen. In an *Attention* condition, subjects viewed 250 dots moving radially from the centre at 4.7 degrees per second and were asked to detect changes in radial velocity. In *No attention*, the subjects were asked simply to view the moving dots. In a *Static* condition, subjects viewed stationary dots. The order of the conditions alternated between *Fixation* and visual stimulation (*Static*, *No Attention*, or *Attention*). In all conditions subjects fixated the centre of the screen. No overt response was required in any condition and there were no actual changes in the speed of the dots. The data were analyzed using a conventional SPM[9] analysis. The regions chosen for network analysis were selected in a rather *ad hoc* fashion to ensure that the regional summaries were defined functionally by selecting regions showing evoked responses. Six representative regions were defined as clusters of contiguous voxels surviving an F-test for all effects of interest at $p<0.001$ (uncorrected) in the conventional SPM analysis. These regions were chosen to cover a distributed network (of largely association cortex) in the right

[9] Software available at http://www.fil.ion.ucl.ac.uk/spm.

hemisphere, from visual cortex to frontal eye fields (see Table 4.1 for details). The activity of each region (node) was summarized with its principal eigenvariate to ensure an optimum weighting of contributions for each voxel within the ROI (see Figure 4.22). In this example, one can see evoked responses in visual areas (every 60 s) with a progressive loss of stimulus-bound activity and a hint of attentional modulation and other fluctuations in higher regions [15].

Results

Results were obtained after 24 iterations of SCKS algorithm (using standard initialization), when it reached the convergence criteria. Then the best model was selected among 32,768 possible models using *post hoc* BMS with posterior model probability about 80 %. The result before and after model selection are depicted in - Figure 4.23. We can see that the final result has relatively sparse structure of

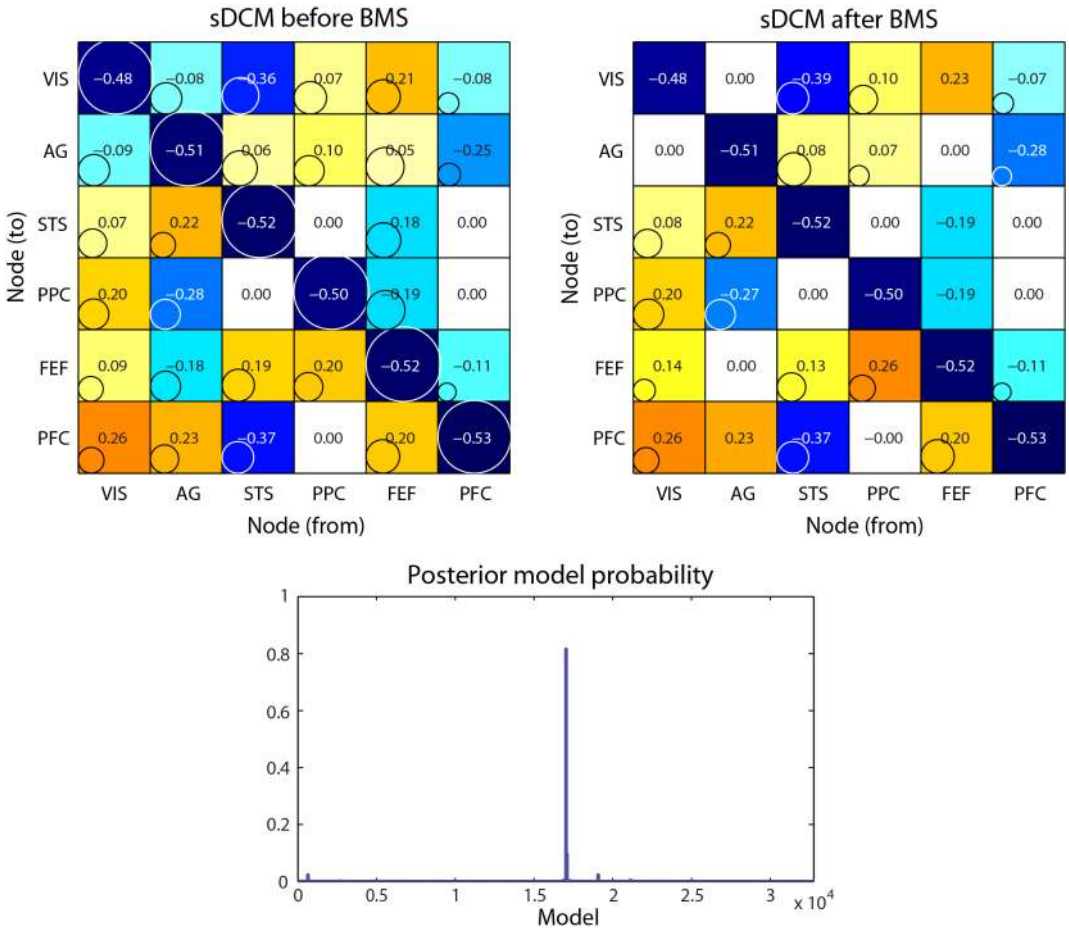


Figure 4.23 Results of empirical data analysis (part 1.). The upper row shows the results of connectivity structures obtained by model inversion using SCKS algorithm before (left) and after *post hoc* model inversion. In this case the best model was selected with more than 80 % evidence as depicted in the lower plot.

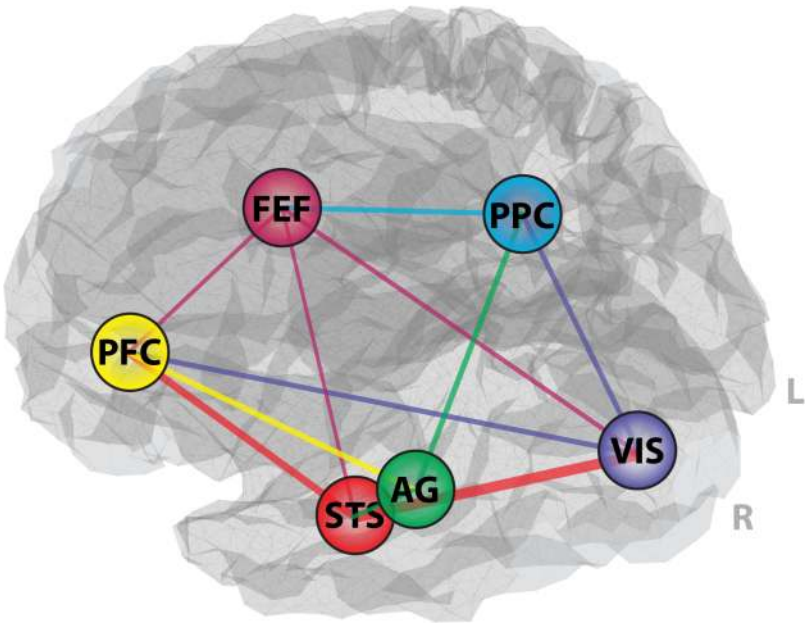


Figure 4.24 Results of empirical data analysis (part 2.). Visualization of identified connectivity structure in anatomical space. The color of the lines reports the source of the strongest bidirectional connection and the width represents its absolute (positive or negative) strength.

connectivity matrix with four bidirectional connections switched off. The architecture of this identified network is then shown also in an anatomical space (Figure 4.24), where the color of the arrow reports the source of the strongest bidirectional connection and the width represents its absolute (positive or negative) strength. This visualization refers to undirected graphs, although our scheme provides separate estimates for both directions of reciprocal connections. As maybe expected, there are stronger forward connections coming from the visual cortex, which can be considered as a bottom of functional hierarchy, to posterior parietal cortex and prefrontal cortex. Interestingly, there are also many backward connections that are stronger than the forward ones. For example from frontal eye fields, which could be considered as the top of the functional hierarchy, to the visual cortex, prefrontal cortex and to superior temporal sulcus. This is quite sensible given the greater amount of backward connections (neuronal pathways) anatomically, both within the cortical hierarchy and from cortex to subcortical structures [15, 162]. Finally, most of identified connections (but not quite all) are in agreement with a previous results obtained by analyzing effective connectivity using the same data-set [15, 123].

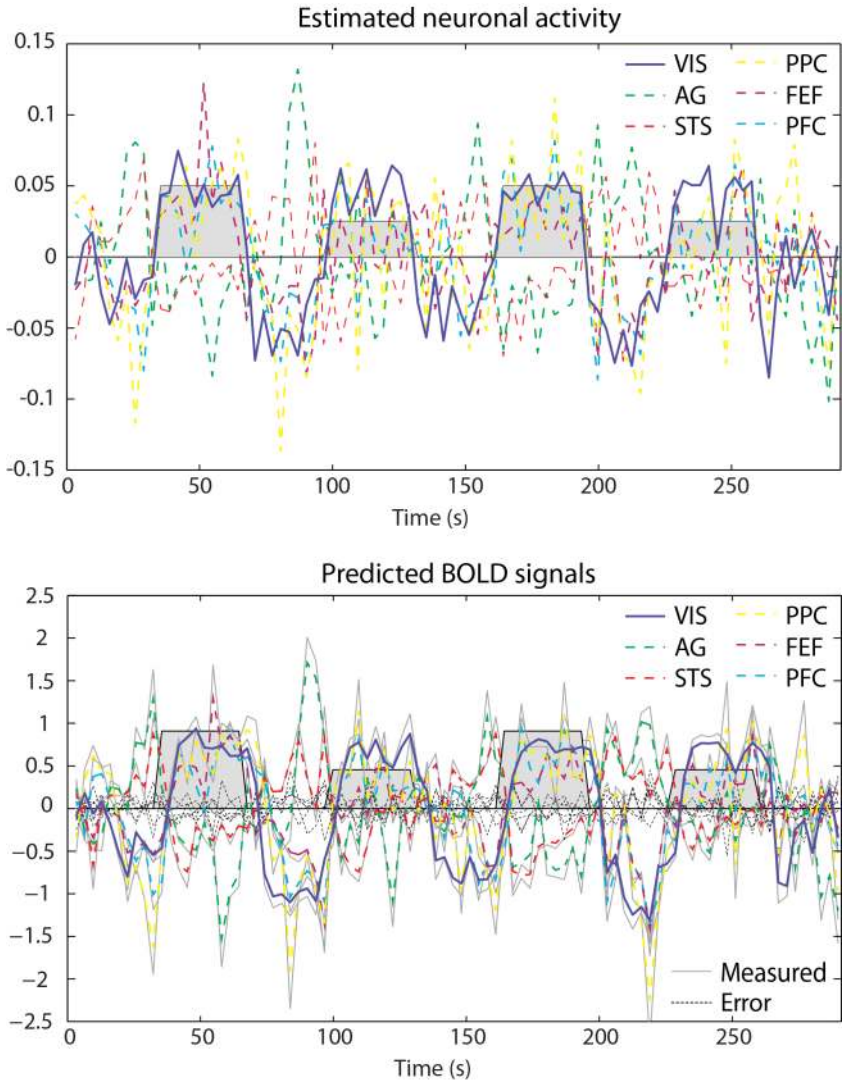


Figure 4.25 Results of empirical data analysis (part 3.). The upper plot displays the estimated neuronal signals, where we highlight the neuronal responds of the visual cortex. The shaded represents the paradigm of the visual stimulation. Similarly, the lower plot displays predicted BOLD responses.

Besides identification of effective connectivity we can be also interested in estimated time courses of neuronal signals (bottom of Figure 4.25). Specifically, because we know the experimental paradigm of visual stimulation (grey filled areas - high for attention and low for no attention), we can compare it to our estimates. In particular, looking at the estimate of neuronal signal associated with the visual cortex (highlighted in blue) we can clearly see that the activation and deactivation phase of this region match exactly to the start and end phase of the stimulus. This confirms that model inversion has effectively estimated neuronal activity from observed BOLD signal and that this estimate is veridical in relation to the experimental manipulation. For comparison to neuronal estimates we also show the original observed BOLD signals

and our predictions (bottom of Figure 4.25), where there is a delay of several seconds (about 6 s) between the start of stimulation and actual increase in BOLD signal.

Before closing we should again emphasize that the model inversion was not informed by a known stimulation function but can still recover the evoked responses.

4.4 Chapter summary

In this chapter we attempted to validate the proposed estimation framework based on the SCKS algorithm for estimation of neuronal signal and effective connectivity from simulated and empirical data. The first part of this chapter was focused on the estimation of the neuronal signal from single fMRI time course, where it mainly served as a tutorial on how to initialize and perform successful model inversion with SCKS. We also discussed identification issues of hemodynamic model parameters, and suggested a suitable set of parameters, which estimation can lead to the improved estimate of the neuronal signal. Note that in future, we might elaborate on this problem with more thorough analysis.

Primary method validation was then carried out for the multivariate case, where we tested the performance of SCKS accompanied by a *post hoc* Bayesian model selection as an approach to stochastic DCM. In this case the main results can be summarized as follows: the approach seems to be robust enough in situations of very noisy data with relatively large sampling period. We expect an improved accuracy of coupling estimates by increasing data SNR and temporal resolution. Next, the method is able to account for hemodynamic response variability across different brain regions. In other words, our method is not sensitive to this sort of variability. This feature makes our approach superior to other approaches which do not consider the generative model. The above mentioned conclusions generalize also to the inversion of a single time course.

The results also suggest that it should be possible to apply the method to the networks consisting of larger number of nodes. Although it is still not possible to think about the application to the whole data consisting of thousands of voxels, one might imagine its application to the brain activity summarized by spatially distributed modes (often around 20), such as those obtained from the independent component analysis [5, 128].

The results obtained by testing for the robustness against the missing region problem are slightly less promising. In this case we observed a decrease in posterior model probability estimates and an increase of false positive identifications. Although generally these results are still satisfactory, in future work we expect to enhance regularization techniques for parameter estimation that could improve the performance in this particular case.

We have also demonstrated an application of proposed algorithm to empirical (task) data and obtained reasonable results, although these results are not deeply discussed, as it is difficult to make any conclusion based on only single subject data-set.

The fact that the simulated data are based on the same model of neuronal interaction as later used for model inversion can be considered as a limitation of our method validation. Though there are features such as higher integration step and modeling of correlated noise for observations and hidden states that are included during the data generation, but are not applied or are ignored during the model inversion, it would be interesting to see if the approach can maintain the same performance in cases of more realistic interaction models (including transmission delays with respect to spatial location [138, 163]) are used for data generation (while still keeping the simple neuronal interaction model for data inversion).

Although we did not provide a comparison to other methods, some comparison can be found in our published paper [20] that focused on single time course model inversion. We should note that at present there is one more approach that can achieve the same aim [52], i.e. estimation stochastic DCM without knowing the exogenous input. However, this method, which definitely deserves distinction, is also very recent and still under development. Clearly, it will be subject of our future research to make comparisons and learn more about its properties. Importantly, the existence of these two competing methods suggests an interesting possibility for cross-validation when applied to real fMRI data.

Chapter 5

Conclusions and future work

In order to evaluate effective connectivity among different brain regions we need to model interactions at the neuronal level. In the case of fMRI data this is complicated by the fact that the measured BOLD signal is only an indirect representation of neuronal activations. The chain of physiological processes that connect the neuronal activation to the BOLD signal can be described by a continuous nonlinear hemodynamic model. Clearly, no model is perfect, which means that it is very important to allow for random fluctuations in unobserved (hidden) neuronal and physiological states by assuming a stochastic representation. Moreover, if we are not restricted to a deterministic model, we are able to account also for (endogenous) autonomous dynamics that cannot be explained by known (exogenous) experimental inputs. We can even throw away the prior knowledge about experimental causes of observed responses and make the evaluation of effective connectivity completely data-driven. Crucially, this enables us to assess causal influence at the neuronal level even from the resting-state fMRI data.

To allow evaluation of this stochastic model we consider the brain as a learning machine that infers information about states and parameters from the observed data. This inference requires representation of uncertainty. Probability theory provides a language for representing the uncertainty beliefs and a framework for maintaining these beliefs in consistent manner. Utilizing probability theory and the general descriptive power of dynamic state-space models, recursive Bayesian estimation provides a theoretically well founded and mathematically robust framework to facilitate sequential probabilistic inference in systems where reasoning under uncertainty is essential. However, because the hemodynamic model we employ is nonlinear, the optimal Bayesian solution to the probabilistic inference problem under consideration is intractable. Therefore, we have to consider only an approximate solution.

In this thesis, we have focused on an approximate solution provided by the Gaussian integration method based on cubature integration rules which was recently

introduced to nonlinear Kalman filtering [55]. Specifically, we have proposed a new approach based on cubature Kalman filtering and Rauch-Tung-Striebel smoothing to a joint estimation problem, where both model states and parameters are estimated sequentially, which also models the interaction (conditional dependences) between them. This framework was further extended to meet all requirements given by the model and the fMRI data we work with. First, we have introduced an extension of this approach to the continuous-discrete time systems, where the accurate and stable discretization of the process model was achieved by a local-linearization scheme [56]. Second, to allow the model inversion in a situation, where we *a priori* do not know the noise statistics of the observed BOLD signal, we have adopted an iterative variational Bayesian approach [64] to sequential estimation of measurement noise variance. Third, to improve the convergence of joint state and parameter estimation, we have proposed an adaptive scheme for the estimation of the parameters and state process noise covariance by efficient Robbins-Monro stochastic approximation scheme. Fourth, since we deal with observed data of a limited length, the forward cubature Kalman filter pass and the backward cubature RTS smoother pass, were wrapped into a simple iterative scheme that maximizes the log-likelihood with each iteration and provides fast convergence. Finally, to further improve the numerical stability of the filter the entire scheme was considered in its square-root form.

All these developments and extensions, which were in detail described in Chapter 2, had one common aim: to enable the estimation of the neuronal signal from a noisy BOLD signal, while considering a realistic nonlinear generation model of the observed signal which also includes stochastic fluctuations contributing to the hidden hemodynamic and neuronal states. In addition, the proposed approach to inversion of the model has a character of (generalized nonlinear) blind deconvolution, because the unknown endogenous neuronal signal (input) to hemodynamic model, which contains unknown parameters is estimated (only) from observed BOLD signal. While sufficient theoretical description of the methods and particular algorithms was provided in Chapter 2, the global validation was shown in the first part of Chapter 4. Additionally, several performance tests using toy examples are summarized in Appendix A and comparison to a related approach (dynamic expectation maximization) [14] can be found in a separate publication [20].

Although a very advanced and efficient method was proposed in Chapter 2, it fulfils only a first part of the goal that was described in this thesis. The second (main) goal was to enable evaluation of effective connectivity at the level of neuronal signals

given the observed hemodynamic responses. Clearly, the methodological framework described in Chapter 2 solves the more difficult part, i.e. provides estimates of (endogenous) neuronal activity. In Chapter 3 we have extended this framework to multivariate case, where the main interest is estimation of (effective) coupling parameters that inform the neuronal interaction model. In this case we have considered a neuronal interaction model in a form of linear stochastic differential equations, which define interactions as the communication of slow dynamics among macroscopic variables; i.e. brain regions (nodes). By connecting this neuronal model to region-specific hemodynamic models that link the neuronal activation to observations, we enabled full model inversion, which provides conditional estimates of coupling parameters, region-specific neuronal signals, hemodynamic states, and associated hemodynamic model parameters. All that is possible by applying the approach (iterated square-root cubature RTS smoother) developed in Chapter 2. Importantly, this neuronal interaction model allows one to estimate bidirectional connectivity (causal influence) between different nodes. Further, as an extension to the estimation framework, we have introduced an automatic detection of irrelevant coupling parameters using a network pruning algorithm based on calculation of scaled Fisher information matrix. This addition was necessary to improve the performance of estimating coupling (connectivity) parameters, especially in cases when the spurious correlations between observed signals are present. A complete form of this model inversion represents a stochastic treatment of dynamic causal modeling that makes it possible to estimate effective connectivity even in case of unknown model input; i.e. in case of resting-state data, where the neuronal signals causing the hemodynamic responses have purely endogenous character. This is an important departure from the original dynamic causal modeling [11], which was limited to a discrete model of hemodynamic states and assumed *a priori* knowledge of the model input.

This novel approach represents the first level of inference where we are especially interested in conditional estimates of coupling parameters and in the associated error covariance matrix. However, it is still necessary to perform the second level of inference, where we identify (select) the most likely model candidate, which in the case of effective connectivity corresponds to the most likely connectivity matrix. In Chapter 3 we considered model comparison based on calculation of Bayes factor (defined as a ratio of marginal likelihoods of restricted models) and discussed two different approximations to marginal likelihood through common Bayesian information criteria and through recently introduced concept of reduced free energy [146]. We have emphasized the convenience of the later approximation, because it requires only a single

inversion of the full model, where all connections all between nodes are allowed. Moreover, this approach is also well suited for model selection in larger networks.

In Chapter 4, we demonstrated the performance of the proposed approach, first focusing on estimation of the neuronal signal from a single fMRI time course. We also addressed the principal questions one might have regarding the performance of the introduced method to correctly infer the coupling parameters of neuronal interaction model. In this case we were able to show that the method is robust even when applied to data with lower SNR and larger sampling period. Also it is not sensitive to variability of hemodynamic response function across different brain regions. These are important properties, which make the approach superior to other approaches for the evaluation of effective connectivity that are not based on generative models and are not formulated in continuous time. We also showed that there is a good perspective for this approach to be applied to larger networks, where possibly all relevant brain regions are included. As a relative weakness we found that the method is only partly immune to the strong spurious correlations caused by exclusion of relevant region (node) from the analysis.

In conclusion, we have made a significant progress in the development of appropriate methodology for evaluation of effective connectivity in fMRI data, however it is clear that much work still lies ahead. To point out a few future aims:

- In this work we have considered a local approximation of Gaussian probability density, which might provide only locally optimal estimates. We are certainly aware that we could improve identifiability of the inversion problem by employing global estimators. However, global approximation methods are not practical in our case due to the curse of dimensionality. Nevertheless, we hope that the utilization of local approximation based on cubature integration rules, which also considers locally distributed sampling points brings significant improvement. This direction obviously leaves much space for further tests and developments.

- We expect to investigate other approaches to improve the regularization of sequentially optimized parameters. We suspect that significant improvement could be obtained by including real priors on parameters (or states) into the estimation scheme. Within a fully Bayesian framework we could use the priors to regularize or finesse the "ill-possedness" of the inverse problem. Even though this might not be possible with the current approach based on Kalman filter/smoother, we hope to find a solution in recently introduced generalized

filtering [52], which can also enable sequential optimization of both state and parameter conditional densities by using the concept of generalized motion (coordinates) of hidden states. However, also in this case we expect to apply cubature integration rules to approximate Gaussian probability density.

- There is an increasing evidence that more realistic neurobiological modeling could be improved by considering non-Markovian process [15, 164], which eschews (implausible) Markovian assumptions about the serial independence of random fluctuations. Solution to this problem can be provided again by applying the concept of generalized coordinates of motion [52], which assumes random differential equations [165] instead of standard stochastic differential equations. However, it should be tested whether this assumption can really improve the final estimates.

- Certainly other work can be focused on extending the generative models toward further biophysical realism. Each extension means an increase of model complexity, which introduces other identifiability problems. Therefore it will be very important to make a compromise between the biophysical realism and the model identifiability, since both are necessary to answer difficult questions about the brain function.

- We believe that the simulations carried out in this thesis are useful, but they should be extended to cover grater realism in neurodynamics [138, 163]. Nevertheless, it will be also very useful to have standardized experimental data from animals as a resource for model testing [130]. One can imagine a combined data-set that provides intracranial recordings of neural sources, higher resolution BOLD-fMRI, surface EEG, diffusion tensor imaging tractography, etc., that would be hugely beneficial and useful for method developers.

As a final remark we should very briefly describe the journey that we undertook and which led us to the results presented in this thesis.

At the beginning of this project in 2008 it was not very clear which direction is the best to go because we had almost no experience with fMRI data and with the analysis of functional and effective connectivity in general. There were two main methodological concepts of analyzing effective connectivity introduced to the neuroscience community: Granger causality modeling (GCM) [166] and dynamic causal modeling (DCM) [11], which were rapidly gaining a wide interest. Since the mathematical background of

GCM based on multivariate autoregressive modeling is much simpler than the one employed by DCM, GCM seemed to be a good approach to start with. Therefore, we chose GCM and focused on its limitation with respect to fMRI data. In particular, we tried to improved the GCM performance by considering time-varying estimation, first using simple windowing technique [167, 168], and later introducing a state-space formulation based on two-pass Kalman filtering [128]. However, there was an increasing evidence [9, 126] that a presents of hemodynamic response variability between different brain regions can seriously confuse the identification of causal relationship. Having the best intentions, we went even further and introduced a new GCM approach that considers a generative nonlinear hemodynamic model for estimation of neuronal signals [129], from which the connectivity parameters are later inferred. Nevertheless, we realized that at this point we were already so close to the concept of DCM that it did not have any further sense to rescue a slowly (but surely) sinking concept of GCM in the application to fMRI, but rather switch the conceptual framework directly to DCM. By experiencing this lesson, we ended up on the trail that led us to the development of the nonlinear blind deconvolution technique for the estimation of neuronal signal from fMRI data and to the stochastic DCM, as they are presented in this thesis.

The process undergone while working on this thesis can be described the best by the following quote:

"I may not have gone where I intended to go, but I think I have ended up where I intended to be."

— Douglas Adams

Appendix A

Toy examples

A.1 Performance of local linearization scheme

In this section we test the performance of continuous-discrete cubature Kalman filter based on local linearization (LL) scheme with time update described in Algorithm 6 and compare it with the recently introduced continuous-discrete cubature Kalman filter based on Itô-Taylor expansion of order 1.5 (IT-1.5) [109]. For this comparison, we will repeat part of the air-traffic-control experiment described in [109]. It should be noted that in [109] the IT-1.5 approach demonstrated superior performance compared to the continuous-discrete versions of extended and unscented Kalman filters.

Radar tracker for coordinated turns

In this illustrative example we consider a typical air-traffic-control scenario, where the objective is to track the trajectory of an aircraft that executes a maneuver at (nearly) constant speed and turn rate in the horizontal plane. Specifically, the motion in the horizontal plane and the motion in the vertical plane are considered to be decoupled from each other. In the aviation language, this kind of motion is commonly referred to as (nearly) coordinated turn. Hence, we may write the coordinated turn in the three-dimensional space, subject to fairly "small" noise modeled by independent Brownian motions as shown by:

$$d\mathbf{x}(t) = \mathbf{f}\big(\mathbf{x}(t)\big)dt + \sqrt{\mathbf{Q}}d\mathbf{w}(t), \tag{A.1}$$

where the seven-dimensional state of the aircraft $\mathbf{x} = [\epsilon\ \dot{\epsilon}\ \eta\ \dot{\eta}\ \zeta\ \dot{\zeta}\ \omega]^T$ with ϵ, η and ζ denoting positions and $\dot{\epsilon}$, $\dot{\eta}$ and $\dot{\zeta}$ denoting velocities in the x, y and z Cartesian coordinates, respectively; ω denotes the turn rate; the drift function $\mathbf{f}(\mathbf{x}) = [\dot{\epsilon}, (-\omega\dot{\eta}), \dot{\eta}, \omega, \dot{\epsilon}, \dot{\zeta}, 0, 0]^T$; the noise term $\mathbf{w}(t) = [w_1(t), w_2(t), \dots, w_7(t)]^T$ with

$\{w_i(t)\}$, $i = 1, 2, \ldots, 7$, being all mutually independent standard Brownian motions, accounts for unpredictable modeling errors due to turbulence, wind force, etc.; and finally the diffusion matrix $\mathbf{Q} = \text{diag}([0, q_1^2, 0, q_1^2, 0, q_1^2, q_2^2])$. For the experiment at hand, the radar is located at the origin and equipped to measure the range, d, azimuth angle, θ, and elevation angle, ϕ, at measurement sampling time T. Hence, we write the measurement equation:

$$\begin{pmatrix} d_t \\ \theta_t \\ \phi_t \end{pmatrix} = \begin{pmatrix} \sqrt{\epsilon_t^2 + \eta_t^2 + \zeta_t^2} \\ \tan^{-1}\left(\frac{\eta_t}{\epsilon_t}\right) \\ \tan^{-1}\left(\frac{\zeta_t}{\sqrt{\epsilon_t^2 + \eta_t^2}}\right) \end{pmatrix} + \mathbf{r}_t, \tag{A.2}$$

where the measurement noise is $\mathbf{r}_t \sim \mathcal{N}(0, \mathbf{R})$.

The data generation followed exactly the procedure described in [109], where other details can be found. We considered process diffusion matrix $\mathbf{Q} = \text{diag}([0, 0.2, 0, 0.2, 0, 0.2, 7 \cdot 10^{-6}])$ and the variance of measurement noise $\mathbf{R} = \text{diag}([50^2, 0.01, 0.01])$. Further, the true initial state was $\mathbf{x}_0 = [1000\text{ m}, 0\text{ ms}^{-1}, 2560\text{ m}, 150\text{ ms}^{-1}, 200\text{ m}, 0\text{ ms}^{-1}, 4.5°s^{-1}]^T$ and the data sampling interval was $\Delta t = 6$ s. The independent aircraft trajectories were generated using IT-1.5 with $K = 1000$ integration steps inside sampling interval.

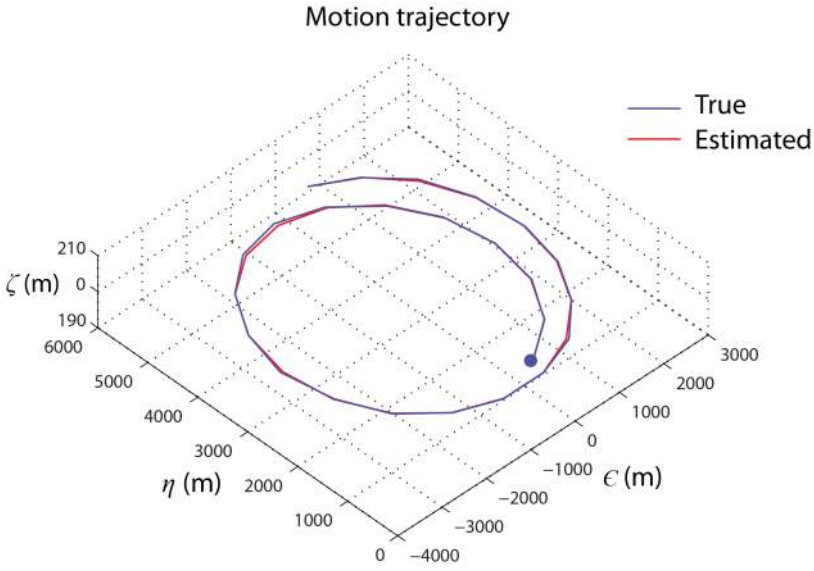


Figure A.1 Change of motion trajectory for the turn rate $\omega = 4.5°$.

Finally, we are ready to compare performance of CKF based on IT-1.5 and CKF based on LL. We considered 50 independent Monte Carlo runs, each collected for time interval of 210 s. Example of the motion trajectory is shown in Figure A.1. Both Bayesian filters were initialized with the same initial condition for each run. The initial state density was assumed to be Gaussian and the two-point differencing method, which uses the first two measurements to estimate the states' statistics [109]. We repeated the model inversion 7 times for different number of integration steps ($K = 2^0, 2^1, \dots, 2^6$); i.e. number of time update iterations, between fixed time interval $\Delta t = 6$ s.

To compare these two filters, we used the accumulative RMSE of the position, velocity and the turn rate. For example, we define the accumulative RMSE in position as:

$$\sqrt{\frac{1}{TN}\sum_{t=1}^{T}\sum_{n=1}^{N}\left((\epsilon_t^n - \hat{\epsilon}_t^n)^2 + (\eta_t^n - \hat{\eta}_t^n)^2 + \left(\zeta_t^n - \hat{\zeta}_t^n\right)^2\right)} \qquad \text{(A.3)}$$

where $(\epsilon_t^n, \eta_t^n, \zeta_t^n)$ and $\left(\hat{\epsilon}_t^n, \hat{\eta}_t^n, \hat{\zeta}_t^n\right)$ are the true and estimated positions at time t and in the n-th Monte Carlo run. Similarly we consider accumulative RMSE for velocity and turn rate. Note that accumulative RMSEs were computed only for a period of 60-210 s.

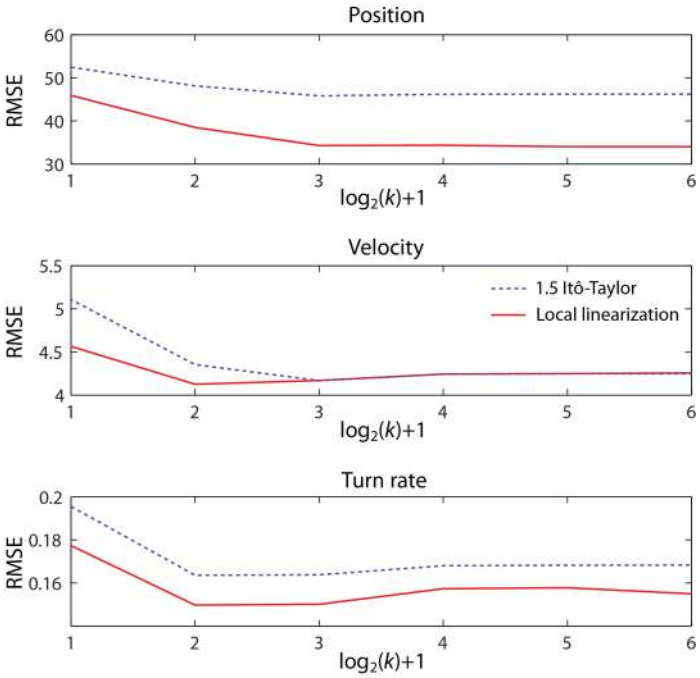


Figure A.2 Results of air-traffic-control experiment. Accumulative RMSE for position, velocity and turn rate with respect to the applied number of integration steps (k) between measurements.

The results are displayed in Figure A.2., where we can see that for both filters the accuracy of estimate (mostly) increases with a smaller integration step and gets stable with more than 8 integration steps between measurements. Overall, the estimates provided by CKF using local linearization for model discretization are of the same or better accuracy than estimates based on CKF using Itô-Taylor expansion of order 1.5. However, it should be noted that local linearization, since it requires calculation of matrix exponential, is much slower.

A.2 Performance of joint estimation scheme

In this section we demonstrate performance of iterated joint estimation scheme based on forward pass of square-root cubature Kalman filter and square-root cubature RTS smoother (square-root version of Algorithm 8). In particular, we show how this scheme is simultaneously able to estimate states and parameters. Additionally, the noise statistics are adaptively estimated as well, which means that we are trying to solve a triple estimation problem.

Lorenz attractor

The model of the Lorenz attractor exhibits deterministic chaos, where the path of the hidden states diverges exponentially on a butterfly-shaped strange attractor in a three dimensional state-space. There are no inputs in this system; the dynamics are autonomous, being generated by nonlinear interactions among the states and their motion. The path begins by spiraling onto one wing and then jumps to the other and back in chaotic way:

$$\begin{aligned} \frac{x_1(t)}{dt} &= \theta_1 x_2(t) - \theta_1 x_1(t) \\ \frac{x_2(t)}{dt} &= \theta_3 x_1(t) - 2x_1(t)x_3(t) - x_2(t) \\ \frac{x_3(t)}{dt} &= 2x_1(t)x_2(t) + \theta_2 x_3(t). \end{aligned} \tag{A.4}$$

We consider the output to be the simple sum of all three states at any time point, $y_t = x_{1,t} + x_{2,t} + x_{3,t}$. The output observation is further contaminated with additive Gaussian noise having zero mean and variance 1. We consider a small amount of the state noise with variance equals $1 \cdot 10^{-5}$. We generated 120 time samples using this

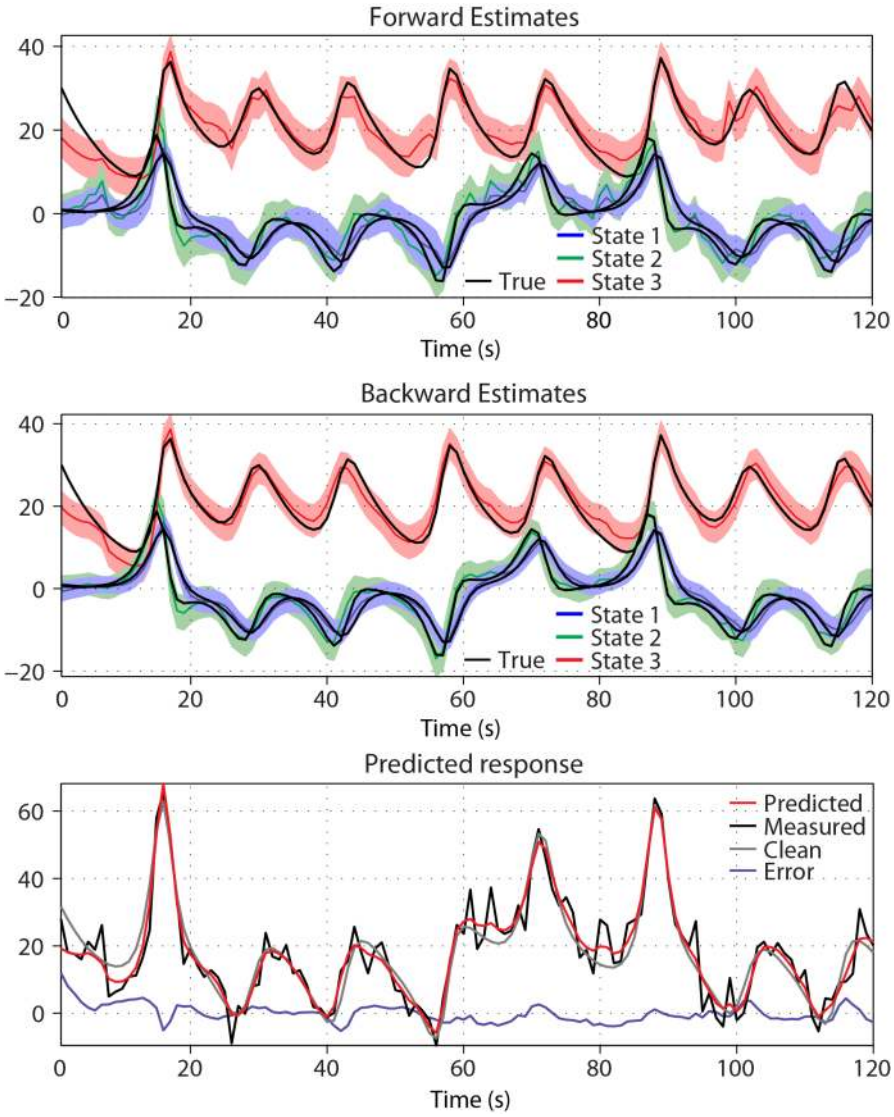


Figure A.3 Results of Lorenz attractor model inversion. The plots in the first and second rows display the results obtained by the forward pass and the backward pass of the SCKS algorithm, respectively. In the third row we can see the predicted responses compared to the clean and observed noisy signals.

model, with initial state conditions $\mathbf{x}_0 = [0.9, 0.8, 30]^T$, parameters $\boldsymbol{\theta} = [18, -4, 46.92]^T$ and an integration step $\Delta t = 1/32$ (using local lineariazation).

This sort of chaotic system shows sensitivity to initial conditions; which, in the case of unknown initial conditions, is a challenge for any inversion scheme. This is even more difficult if also the model parameters are unknown (or at least most of them). Therefore, we want to test the performance of SCKS algorithm in these difficult conditions. Moreover, we will go even further. Besides unknown initial conditions of the states and unknown parameters, we also consider the measurement noise variance to be unknown. It means we have to solve a triple estimation problem.

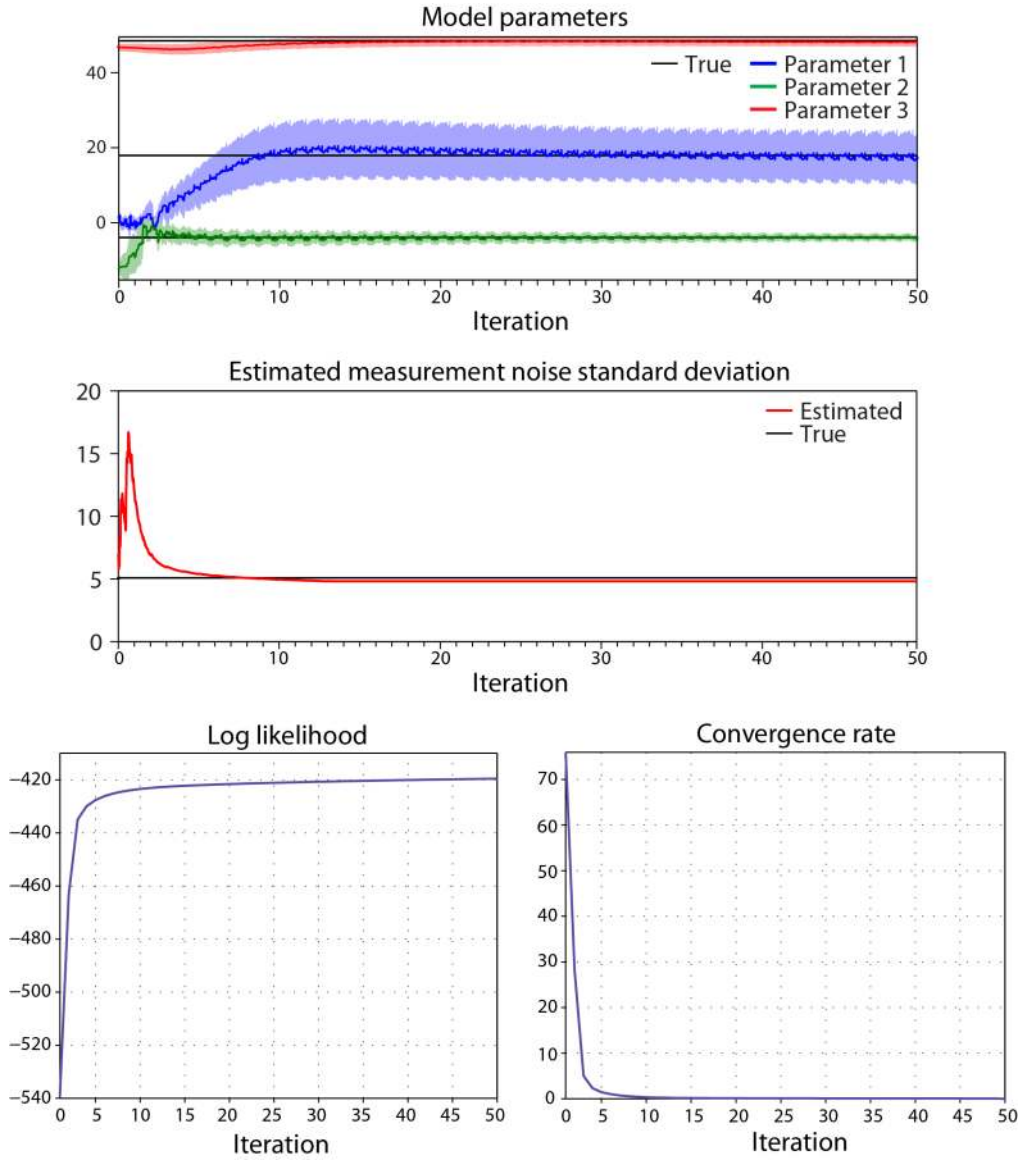


Figure A.4 Estimated parameter of Lorenz attractor model. In the first row are the estimated model parameters compared to their true values (black lines). In the second row we show the estimate of the measurement noise standard deviation that converged very close to the ground truth. Finally the third row displays changes in the log-likelihood and the convergence rate over successive iterations.

We used the initial state conditions $x_0 = [2, 8, 15]^T$ and the parameters $\theta_0 = [2, -12, 44]^T$, where their true values were the same as above. The error covariance matrices for both states and parameters were initialized as $\mathbf{P}_0^x = \mathbf{P}_0^\theta = \mathbf{I}_{n_x} \cdot 0.01$, and the state and parameter noise covariance matrices, $\mathbf{Q}_0 = \mathbf{I}_{n_x} \cdot 2 \cdot 10^{-3}$, $\mathbf{O}_0 = \mathrm{diag}([0.1, 0.1, 1 \cdot 10^{-3}])$, respectively. We allowed the SCKS algorithm to iterate until the convergence (with the tolerance $1 \cdot 10^{-4}$).

The algorithm converged after 50 iterations. This is because we have considered a very small threshold on convergence rate, under which the algorithm should be automatically terminated. The results of the state estimates are depicted in Figure A.3. There we can see that including also backward pass provides more accurate estimates. Additionally, one can see that due to iterations of SCKS algorithm we get closer to the initial state conditions of the simulated data.

The results of parameter estimates are depicted at the top of Figure A.4, where we were able to correctly estimate the true parameter values. In the same figure we also show the (time-varying) estimate of measurement noise variance (displayed as standard deviation), which also converged to the true value. In this particular model, the sequential estimation of the measurement noise variance is crucial for successful joint estimation of the states and parameter in cases where both initial parameters and initial state conditions are significantly different from their true values.

A.3 Performance of measurement noise estimation

In this section we demonstrate the performance of sequential estimation of measurement noise covariance by variational Bayesian approach [119] involved in cubature Kalman filter as it was described in Algorithm 7. We employ very similar example as was used in original paper that introduced this technique for general

Range-only tracking in a non-homogeneous noise field

In this simple example we illustrate the performance of the adaptive filter by tracking a moving target with sensors, which measure the distances to the target moving in 2-dimensional (u, v) space. The measurements are corrupted with noise having time-varying correlations between the sensors.

The state vector contains the position and velocity of the target $\mathbf{x}_t = [u_t \; v_t \; \dot{u}_t \; \dot{v}_t]^T$, where the dynamics of the target are modeled by the standard Wiener velocity model. The distance measurements from m sensors is given by:

$$y_t^i = \sqrt{(s_u^i + u_t)^2 + (s_v^i + v_t)^2} + r_t^i, \quad i = 1, \dots, m, \tag{A.5}$$

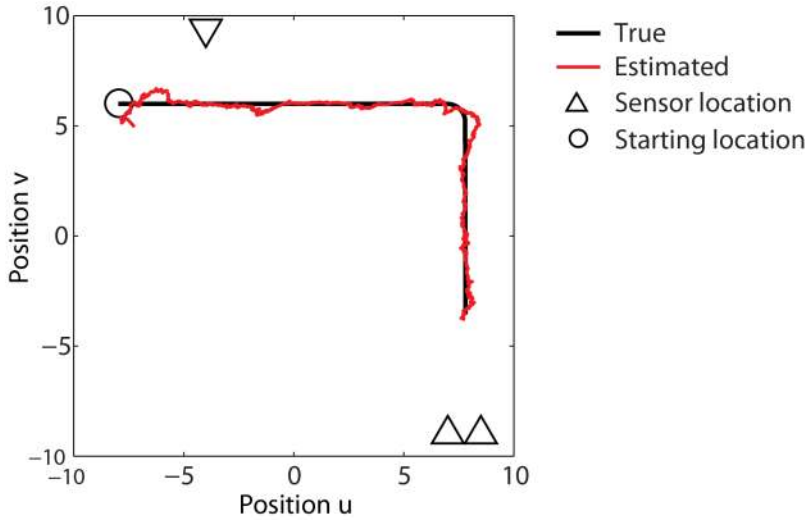


Figure A.5 Simulation scenario of the Range-Only Tracking. Circle denotes the starting position of the target and triangles represent the position of sensor The true trajectory of the target is the black line and the estimated trajectory is the red line.

where $\left(s_u^i, s_v^i\right)$ is the position of *i*-th sensor and r_t^i k is the *i*-th component of a Gaussian distributed noise vector $\mathbf{r}_t = \mathcal{N}(0, \mathbf{R}_t)$. In this experiment we consider a time varying covariance matrix of measurement noise that is changed randomly every 50th time step. This is generated by $\mathbf{R}_t = \boldsymbol{\sigma}_t \boldsymbol{\sigma}_t^T$, where the vector $\boldsymbol{\sigma}_t$ is generated from the Gaussian distribution having zero mean and variance equal to 1, $\boldsymbol{\sigma}_t = \mathcal{N}(0, \mathbf{I})$. The spectral density of the process noise was set to $q = 2$ and the time step to $\delta = 0.01$. The trajectory shown in Figure A.5 was discretized to 1000 time steps and then measurements were generated according to the procedure described above. Given the measurements, the target was tracked by adaptive cubature Kalman filter with estimation of measurement noise variance using variational Bayesian. The parameters of inverse-Gamma distribution were set to $\alpha = 1$, $\beta = 1$, and because we expect time-varying noise covariance the forgetting constant was set to $\rho = 0.9$, to allow faster dynamics. Finally, we considered 5 iterations of VB approach at each time step.

The obtained results for tracked target trajectory are displayed in Figure A.5 (red line). The results of estimated measurement noise covariance, which we in this case restricted to diagonal elements, are depicted in Figure A.6. We can see that the estimated variances (green lines) nicely tracks the true variances (black lines).

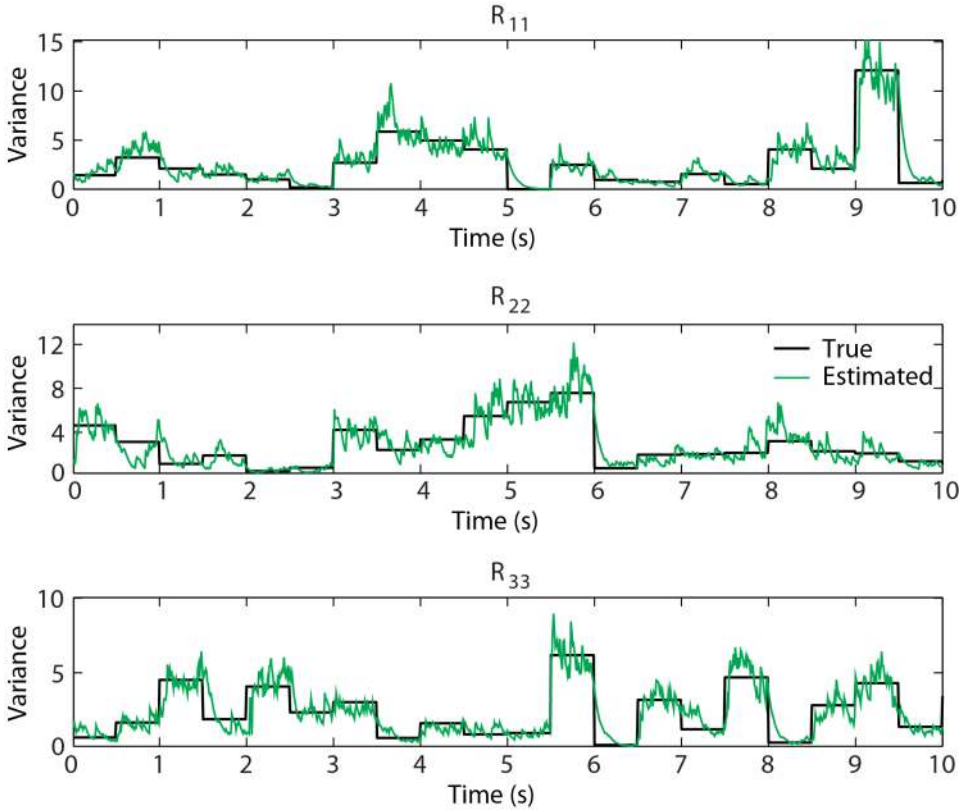


Figure A.6 Results of the measurement noise variance estimation. Only diagonal elements of measurement covariance matrix are displayed.

Appendix B

SCKS toolkit

B.1 SCKS estimation toolkit for Matlab®

The SCKS algorithm introduced in this thesis is represented by a set of functions that were implemented under Matlab®. This toolkit allows estimation of neuronal signal from fMRI data and also evaluation of stochastic dynamic causal modeling. The toolkit also contains several examples, which demonstrate how to use this algorithm. The algorithm uses several functions that are part of SPM8[10] toolbox (with update 4290 or higher) and requires installation of this freeware software. Although, the SCKS toolkit was developed purposely for application to fMRI data, it can be easily applied to any problem that is formulated as a continuous-discrete system. The toolkit package was tested using Matlab 2009 and 2010.

Finally, the toolkit can be download from my personal webpage: https://sites.google.com/site/havlicekmartin, where also the new updates will be later available.

[10] http://www.fil.ion.ucl.ac.uk/spm

Bibliography

[1] K. Friston, "Functional and effective connectivity: a review," *Brain Connectivity,* vol. 1, 2011.

[2] A. R. McIntosh, "Towards a network theory of cognition," *Neural Networks,* vol. 13, pp. 861-870, 2000.

[3] K. J. Friston, "Functional and effective connectivity in neuroimaging: a synthesis," *Human Brain Mapping,* vol. 2, pp. 56-78, 1994.

[4] M. McKeown and T. Sejnowski, "Independent component analysis of fMRI data: examining the assumptions," *Human Brain Mapping,* vol. 6, 1998.

[5] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "A method for making group inferences from functional MRI data using independent component analysis," *Human Brain Mapping,* vol. 14, pp. 140-151, 2001.

[6] K. Friston, C. Buechel, G. Fink, J. Morris, E. Rolls, and R. Dolan, "Psychophysiological and modulatory interactions in neuroimaging," *Neuroimage,* vol. 6, pp. 218-229, 1997.

[7] G. K. Aguirre, E. Zarahn, and M. D'esposito, "The variability of human, BOLD hemodynamic responses," *Neuroimage,* vol. 8, pp. 360-369, 1998.

[8] D. Handwerker, J. Ollinger, and M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses," *Neuroimage,* vol. 21, pp. 1639-1651, 2004.

[9] O. David, I. Guillemain, S. Saillet, S. Reyt, C. Deransart, C. Segebarth, and A. Depaulis, "Identifying neural drivers with functional MRI: an electrophysiological validation," *PLoS Biology,* vol. 6, pp. 2683-2697, 2008.

[10] J. Vartiainen, M. Liljeström, M. Koskinen, H. Renvall, and R. Salmelin, "Functional magnetic resonance imaging blood oxygenation level-dependent signal and magnetoencephalography evoked responses yield different neural functionality in reading," *The Journal of Neuroscience,* vol. 31, p. 1048, 2011.

[11] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling," *Neuroimage,* vol. 19, pp. 1273-1302, 2003.

[12] I. Shoji, "A comparative study of maximum likelihood estimators for nonlinear dynamical system models," *International Journal of Control,* vol. 71, pp. 391-404, 1998.

[13] J. Riera, J. Watanabe, I. Kazuki, M. Naoki, E. Aubert, T. Ozaki, and R. Kawashima, "A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signals," *NeuroImage,* vol. 21, pp. 547-567, 2004.

[14] K. J. Friston, N. Trujillo-Barreto, and J. Daunizeau, "DEM: a variational treatment of dynamic systems," *Neuroimage,* vol. 41, pp. 849-885, 2008.

[15] K. J. Friston, B. Li, J. Daunizeau, and K. E. Stephan, "Network discovery with DCM," *Neuroimage,* pp. 1202-1221, 2010.

[16] G. Deshpande, S. LaConte, G. James, S. Peltier, and X. Hu, "Multivariate Granger causality analysis of brain networks," *Human Brain Mapping,* vol. 30, pp. 1361-1373, 2009.

[17] B. Li, J. Daunizeau, K. E. Stephan, W. Penny, D. Hu, and K. Friston, "Generalised filtering and stochastic DCM for fMRI," *NeuroImage,* pp. 442-457, 2011.

[18] T. Obata, T. T. Liu, K. L. Miller, W. M. Luh, E. C. Wong, L. R. Frank, and R. B. Buxton, "Discrepancies between BOLD and flow dynamics in primary and supplementary motor areas: application of the balloon model to the interpretation of BOLD transients," *NeuroImage,* vol. 21, pp. 144-153, 2004.
[19] M. Havlicek, K. J. Friston, J. Jan, M. Brazdil, and V. D. Calhoun, "Effective connectivity in fMRI resting state data via blind deconvolution," in *17th Annual Meeting of the Organization for Human Brain Mapping*, Quebec, 2011, pp. 1-2.
[20] M. Havlicek, K. J. Friston, J. Jan, M. Brazdil, and V. D. Calhoun, "Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering," *NeuroImage,* vol. 56, pp. 2109-2128, 2011.
[21] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proceedings of the National Academy of Sciences,* vol. 87, p. 9868, 1990.
[22] S. Ogawa, R. S. Menon, D. W. Tank, S. G. Kim, H. Merkle, J. M. Ellermann, and K. Ugurbil, "Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. A comparison of signal characteristics with a biophysical model," *Biophysical Journal,* vol. 64, pp. 803-812, 1993.
[23] R. B. Buxton, K. Uludag, D. J. Dubowitz, and T. T. Liu, "Modeling the hemodynamic response to brain activation," *Neuroimage,* vol. 23, pp. S220-S233, 2004.
[24] K. Uludag "To dip or not to dip: Reconciling optical imaging and fMRI data," *Proceedings of the National Academy of Sciences,* vol. 107, p. 23, 2010.
[25] K. Uludag, B. Müller-Bierl, and K. Ugurbil, "An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging," *Neuroimage,* vol. 48, pp. 150-165, 2009.
[26] N. K. Logothetis, "The neural basis of the blood–oxygen–level–dependent functional magnetic resonance imaging signal," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* vol. 357, p. 1003, 2002.
[27] R. B. Buxton, E. C. Wong, and L. R. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: the balloon model," *Magnetic Resonance in Medicine,* vol. 39, pp. 855-864, 1998.
[28] J. B. Mandeville, J. J. A. Marota, C. Ayata, G. Zaharchuk, M. A. Moskowitz, B. R. Rosen, and R. M. Weisskoff, "Evidence of a cerebrovascular postarteriole windkessel with delayed compliance," *Journal of Cerebral Blood Flow & Metabolism,* vol. 19, pp. 679-689, 1999.
[29] K. J. Friston, A. Mechelli, R. Turner, and C. J. Price, "Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics," *NeuroImage,* vol. 12, pp. 466-477, 2000.
[30] K. E. Stephan, N. Weiskopf, P. M. Drysdale, P. A. Robinson, and K. J. Friston, "Comparing hemodynamic models with DCM," *Neuroimage,* vol. 38, pp. 387-401, 2007.
[31] Y. Zheng, J. Martindale, D. Johnston, M. Jones, J. Berwick, and J. Mayhew, "A model of the hemodynamic response and oxygen delivery to brain," *Neuroimage,* vol. 16, pp. 617-637, 2002.
[32] R. C. Sotero and N. J. Trujillo-Barreto, "Modelling the role of excitatory and inhibitory neuronal activity in the generation of the BOLD signal," *NeuroImage,* vol. 35, pp. 149-165, 2007.

[33] R. C. Sotero, N. J. Trujillo-Barreto, J. C. Jiménez, F. Carbonell, and R. Rodríguez-Rojas, "Identification and comparison of stochastic metabolic/hemodynamic models (sMHM) for the generation of the BOLD signal," *Journal of computational neuroscience,* vol. 26, pp. 251-269, 2009.
[34] C. Iadecola, "CC commentary: intrinsic signals and functional brain mapping: caution, blood vessels at work," *Cerebral Cortex,* vol. 12, pp. 223-224, 2002.
[35] P. Magistretti and L. Pellerin, "Cellular Mechanisms of Brain Energy Metabolism and their Relevance to Functional Brain Imaging," *Philos Trans R Soc Lond B Biol Sci,* vol. 354, pp. 1155-1163, 1999.
[36] D. Attwell, A. M. Buchan, S. Charpak, M. Lauritzen, B. A. MacVicar, and E. A. Newman, "Glial and neuronal control of brain blood flow," *Nature,* vol. 468, pp. 232-243, 2010.
[37] G. S. Berns, A. W. Song, and H. Mao, "Continuous functional magnetic resonance imaging reveals dynamic nonlinearities of "dose-response" curves for finger opposition," *Journal of Neuroscience,* vol. 19, pp. 1-6, 1999.
[38] A. Mechelli, C. Price, and K. Friston, "Nonlinear coupling between evoked rCBF and BOLD signals: a simulation study of hemodynamic responses," *NeuroImage,* vol. 14, pp. 862-872, 2001.
[39] K. L. Miller, W. M. Luh, T. T. Liu, A. Martinez, T. Obata, E. C. Wong, L. R. Frank, and R. B. Buxton, "Nonlinear temporal dynamics of the cerebral blood flow response," *Human Brain Mapping,* vol. 13, pp. 1-12, 2001.
[40] K. J. Friston, "Bayesian estimation of dynamical systems: an application to fMRI," *NeuroImage,* vol. 16, pp. 513-530, 2002.
[41] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magnetic Resonance in Medicine,* vol. 34, pp. 537-541, 1995.
[42] G. Krüger and G. H. Glover, "Physiological noise in oxygenation sensitive magnetic resonance imaging," *Magnetic Resonance in Medicine,* vol. 46, pp. 631-637, 2001.
[43] J. C. Jimenez and T. Ozaki, "Local linearization filters for non-linear continuous-discrete state space models with multiplicative noise," *International Journal of Control,* vol. 76, pp. 1159-1170, 2003.
[44] S. S. Haykin, *Kalman filtering and neural networks*: Wiley Online Library, 2001.
[45] J. Riera, J. Watanabe, I. Kazuki, M. Naoki, E. Aubert, T. Ozaki, and R. Kawahima, "A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signals," *NeuroImage,* vol. 21, pp. 547-567, 2004.
[46] R. M. Birn, Z. S. Saad, and P. A. Bandettini, "Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response," *Neuroimage,* vol. 14, pp. 817-826, 2001.
[47] L. A. Johnston, E. Duff, I. Mareels, and G. F. Egan, "Nonlinear estimation of the BOLD signal," *NeuroImage,* vol. 40, pp. 504-514, 2008.
[48] L. Murray and A. Storkey, "Continuous time particle filtering for fMRI," *Advances in Neural Information Processing Systems,* vol. 20, 2008.
[49] Z. Hu, X. Zhao, H. Liu, and P. Shi, "Nonlinear analysis of the BOLD signal," *EURASIP Journal on Advances in Signal Processing,* vol. 2009, pp. 1-13, 2009.
[50] D. Jacobsen, L. Hansen, and K. Madsen, "Bayesian model comparison in nonlinear BOLD fMRI hemodynamics," *Neural computation,* vol. 20, pp. 738-755, 2008.
[51] K. Friston, "Variational filtering," *Neuroimage,* vol. 41, pp. 747-766, 2008.

[52] K. J. Friston, K. E. Stephan, and J. Daunizeau, "Generalised Filtering," *Mathematical Problems in Engineering*, vol. 2010, 2010.
[53] G. E. Hinton and D. van Camp, "Keeping the neural networks simple by minimizing the description length of the weights," in *In proceedings of COLT-93*, 1993, pp. 5-13.
[54] D. J. C. MacKay, "Developments in probabilistic modelling with neural networks-ensemble learning," in *In proceedings of 3rd Annual Symposium on Neural Networks*, Nijmegen, Netherlands, 1995, pp. 191-198.
[55] I. Arasaratnam and S. Haykin, "Cubature Kalman Filters," *IEEE Transactions on Automatic Control*, vol. 54, pp. 1254-1269, 2009.
[56] T. Ozaki, "A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach," *Statistica Sinica*, vol. 2, pp. 113–-135, 1992.
[57] J. J. Riera, J. Watanabe, I. Kazuki, M. Naoki, E. Aubert, T. Ozaki, and R. Kawashima, "A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signals," *NeuroImage*, vol. 21, pp. 547-567, 2004.
[58] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, pp. 400-407, 1951.
[59] R. Van der Merwe, "Sigma-point Kalman filters for probabilistic inference in dynamic state-space models," Ph.D.thesis, Oregon Graduate Institute of Science and Technology, 2004.
[60] K. P. Murphy, "Dynamic bayesian networks: representation, inference and learning," PhD thesis, UC Berkeley, Computer Science Division, 2002., 2002.
[61] S. Haykin, *Neural networks: a comprehensive foundation*: Prentice hall, 1999.
[62] W. D. Penny and S. J. Roberts, "Bayesian multivariate autoregressive models with structured priors," in *IEE Proc.-Vis. Image Signal Process.*, 2002, pp. 33-41.
[63] D. J. C. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, pp. 415-447, 1992.
[64] S. Särkkä, "Recursive Bayesian inference on stochastic differential equations," Ph.D. Thesis: Helsinki University of Technology Laboratory of Computational Engineering, 2006.
[65] B. D. O. Anderson, J. B. Moore, and J. Barratt, *Optimal filtering*: Prentice-Hall, 1979.
[66] I. Arasaratnam and S. Haykin, "Cubature Kalman smoothers," *Automatica*, vol. 47, pp. 2245-2250, 2011.
[67] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, pp. 35-45, 1960.
[68] A. T. Nelson, "Nonlinear estimation and modeling of noisy time-series by dual Kalman filtering methods," Ph.D thesis, Oregon Graduate Institute of Science and Technology, 2000.
[69] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, pp. 1445-1450, 1965.
[70] S. Julier, J. Uhlmann, and H. F. Durrant-Whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *Automatic Control, IEEE Transactions on*, vol. 45, pp. 477-482, 2002.
[71] M. Norgaard, N. K. Poulsen, and O. Ravn, "New developments in state estimation for nonlinear systems," *Automatica*, vol. 36, pp. 1627-1638, 2000.
[72] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems," *Automatic Control, IEEE Transactions on*, vol. 45, pp. 910-927, 2002.

[73] R. S. Bucy and K. D. Senne, "Digital synthesis of non-linear filters," *Automatica,* vol. 7, pp. 287-298, 1971.
[74] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*: Springer Verlag, 2001.
[75] T. S. Jaakkola, "Tutorial on variational approximation methods," *Advanced mean field methods: theory and practice,* pp. 129-159, 2000.
[76] C. Fernandez-Prades and J. Vila-Valls, "Bayesian Nonlinear Filtering Using Quadrature and Cubature Rules Applied to Sensor Data Fusion for Positioning," in *In proceeding of IEEE International Conference on Communications,* 2010, pp. 1-5.
[77] P. Li, J. Yu, M. Wan, and J. Huang, "The augmented form of cubature Kalman filter and quadrature Kalman filter for additive noise," in *IEEE Youth Conference on Information, Computing and Telecommunication, YC-ICT '09.*, 2009, pp. 295-298.
[78] P. S. Maybeck, *Stochastic models, estimation and control* vol. 141: Academic press, 1979.
[79] G. Welch and G. Bishop, "An introduction to the Kalman filter," *University of North Carolina at Chapel Hill, Chapel Hill, NC,* vol. 7, 1995.
[80] I. Arasaratnam, S. Haykin, and R. J. Elliott, "Discrete-time nonlinear filtering algorithms using Gauss-Hermite quadrature," *Proceedings of the IEEE,* vol. 95, pp. 953-977, 2007.
[81] M. Simandl, J. Královec, and T. Söderström, "Advanced point-mass method for nonlinear state estimation," *Automatica,* vol. 42, pp. 1133-1145, 2006.
[82] T. Gerstner and M. Griebel, "Numerical integration using sparse grids," *Numerical algorithms,* vol. 18, pp. 209-232, 1998.
[83] I. H. Sloan, "Lattice methods for multiple integration," *Journal of Computational and Applied Mathematics,* vol. 12, pp. 131-143, 1985.
[84] R. Cools, "Advances in multidimensional integration," *Journal of computational and applied mathematics,* vol. 149, pp. 1-12, 2002.
[85] T. Lefebvre, H. Bruyninckx, and J. De Schuller, "Comment on A new method for the nonlinear transformation of means and covariances in filters and estimators [and authors' reply]," *Automatic Control, IEEE Transactions on,* vol. 47, pp. 1406-1409, 2002.
[86] P. Kaminski, A. Bryson Jr, and S. Schmidt, "Discrete square root filtering: A survey of current techniques," *Automatic Control, IEEE Transactions on,* vol. 16, pp. 727-736, 1971.
[87] P. Kaminski, A. Bryson Jr., and S. Schmidt, "Discrete square root filtering: A survey of current techniques," *Automatic Control, IEEE Transactions on,* vol. 16, pp. 727-736, 1971.
[88] S. Haykin, *Kalman filtering and neural networks*: Wiley Online Library, 2001.
[89] J. Dunik and M. Simandl, "Design of Square-Root Derivative-Free Smoothers," in *7th international Ph.D. workshop: Young generation viewpoint, Proceeding on,* Prague, 2006, pp. 33-41.
[90] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 1-38, 1977.
[91] T. K. Moon, "The expectation-maximization algorithm," *Signal Processing Magazine, IEEE,* vol. 13, pp. 47-60, 1996.

[92] B. M. Bell and F. W. Cathey, "The iterated Kalman filter update as a Gauss-Newton method," *Automatic Control, IEEE Transactions on,* vol. 38, pp. 294-297, 1993.
[93] L. Ljung, "Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems," *Automatic Control, IEEE Transactions on,* vol. 24, pp. 36-50, 1979.
[94] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory (v. 1)*: Prentice Hall PTR, 1993.
[95] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior Cramér-Rao bounds for discrete-time nonlinear filtering," *Signal Processing, IEEE Transactions on,* vol. 46, pp. 1386-1396, 1998.
[96] J. F. M. Van Doren, S. G. Douma, P. M. J. Van den Hof, J. D. Jansen, and O. H. Bosgra, "Identifiability: From qualitative analysis to model structure approximation," in *15th IFAC Symp. System Identification*, 2009, pp. 664–669.
[97] Z. Ji and M. Brown, "Joint State and Parameter Estimation For Biochemical Dynamic Pathways With Iterative Extended Kalman Filter: Comparison With Dual State and Parameter Estimation," *Open Automation and Control Systems Journal,* vol. 2, pp. 69-77, 2009.
[98] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations, Stochastic Modeling and Applied Probability, 3rd editon*: Springer, 1999.
[99] A. T. Fuller, "Analysis of nonlinear stochastic systems by means of the Fokker-Planck equation(Nonlinear systems disturbed by random white noise analyzed by Fokker-Planck equation for probability density in state space)," *International journal of control, first series,* vol. 9, pp. 603-655, 1969.
[100] A. H. Jazwinski, *Stochastic processes and filtering theory* vol. 63: Academic Press, 1970.
[101] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Transactions of the ASME. Series D, Journal of Basic Engineering,* vol. 83, pp. 95-107, 1961.
[102] W. Bangerth and R. Rannacher, *Adaptive finite element methods for differential equations*. Boston, MA: Birkhäuser, 2003.
[103] J. Svácha and M. Simandl, "Nonlinear state prediction by separation approach for continuous-discrete stochastic systems," *Kybernetika,* vol. 44, pp. 61-74, 2008.
[104] K. Kastella, "Finite difference methods for nonlinear filtering and automatic target recognition," *Multitarget-multisensor tracking: Applications and advances.*, vol. 3, pp. 233-258, 2000.
[105] A. Budhiraja, L. Chen, and C. Lee, "A survey of numerical methods for nonlinear filtering problems," *Physica D: Nonlinear Phenomena,* vol. 230, pp. 27-36, 2007.
[106] C. Berzuini, N. G. Best, W. R. Gilks, and C. Larizza, "Dynamic conditional independence models and Markov chain Monte Carlo methods," *Journal of the American Statistical Association,* vol. 92, pp. 1403-1412, 1997.
[107] J. D. Hoffman, *Numerical methods for engineers and scientists*. New York: McGraw-Hill, 1992.
[108] S. Särkkä, "Continuous-time and continuous-discrete-time unscented Rauch-Tung-Striebel smoothers," *Signal Processing,* vol. 90, pp. 225-235, 2010.

[109] I. Arasaratnam, S. Haykin, and T. R. Hurd, "Cubature Kalman Filtering for Continuous-Discrete Systems: Theory and Simulations," *Signal Processing, IEEE Transactions on*, vol. 58, pp. 4977-4993, 2010.
[110] T. Ozaki, "A local linearization approach to nonlinear filtering," *International Journal of Control*, vol. 57, pp. 75-75, 1993.
[111] L. Ljung and T. Söderström, *Theory and practice of recursive identification*. Cambridge, MA: MIT Press, 1983.
[112] J. C. Jimenez, "A simple algebraic expression to evaluate the local linearization schemes for stochastic differential equations* 1," *Applied Mathematics Letters*, vol. 15, pp. 775-780, 2002.
[113] C. Van Loan, "Computing integrals involving the matrix exponential," *Automatic Control, IEEE Transactions on*, vol. 23, pp. 395-404, 1978.
[114] R. Mehra, "Approaches to adaptive filtering," *Automatic Control, IEEE Transactions on*, vol. 17, pp. 693-698, 1972.
[115] K. Myers and B. Tapley, "Adaptive sequential estimation with unknown noise statistics," *Automatic Control, IEEE Transactions on*, vol. 21, pp. 520-523, 1976.
[116] A. P. Sage and G. W. Husa, "Algorithms for sequential adaptive estimation of prior statistics," in *Adaptive Processes (8th) Decision and Control, IEEE Symposium on*, 1969, pp. 61-69.
[117] J. Dunik, M. Simandl, and O. Straka, "Methods for estimating state and measurement noise covariance matrices: aspects and comparison," in *15th IFAC Symposium on System Identification*, 2009, pp. 372-377.
[118] I. Arasaratnam and S. Haykin, "Nonlinear Bayesian Filters for Training Recurrent Neural Networks," *MICAI 2008: Advances in Artificial Intelligence*, pp. 12-33, 2008.
[119] S. Särkkä and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *Automatic Control, IEEE Transactions on*, vol. 54, pp. 596-600, 2009.
[120] M. Havlicek, J. Jan, M. Brazdil, and V. D. Calhoun, "Estimation of neuronal signal from hemodynamic responses," in *33rd IEEE EMBS conference, Proceeding of*, Boston, 2011, pp. 8122-8125.
[121] S. Saha, E. Ozkan, F. Gustafsson, and V. Smidl, "Marginalized particle filters for Bayesian estimation of Gaussian noise parameters," in *Information Fusion (FUSION), 13th Conference on*, 2010, pp. 1-8.
[122] M. Havlicek, J. Jan, M. Brazdil, and V. D. Calhoun, "Deconvolution of neuronal signal from hemodynamic response," in *International Conference on Acoustics, Speech, and Signal Processing, IEEE ICASSP, Proceeding of*, 2011, pp. 617-620.
[123] C. Buchel, "Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI," *Cerebral Cortex*, vol. 7, pp. 768-778, 1997.
[124] A. R. Mclntosh and F. Gonzalez-Lima, "Structural equation modeling and its application to network analysis in functional brain imaging," *Human Brain Mapping*, vol. 2, pp. 2-22, 1994.
[125] A. Roebroeck, E. Formisano, and R. Goebel, "The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution," *Neuroimage*, vol. In Press, 2009.

[126] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for FMRI," *Neuroimage,* vol. 54, pp. 875-891, 2010.
[127] A. Roebroeck, E. Formisano, and R. Goebel, "Mapping directed influence over the brain using Granger causality and fMRI," *Neuroimage,* vol. 25, pp. 230-242, 2005.
[128] M. Havlicek, J. Jan, M. Brazdil, and V. D. Calhoun, "Dynamic Granger causality based on Kalman filter for evaluation of functional network connectivity in fMRI data," *NeuroImage,* vol. 53, pp. 65-77, 2010.
[129] M. Havlicek, J. Jan, S. PLIS, M. Brazdil, and V. D. Calhoun, "Dynamic Granger causality with embedded hemodynamic model," in *16th Annual Meeting of the Organization for Human Brain Mapping*, Barcelona, 2010, pp. 88-89.
[130] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston, "Effective connectivity: Influence, causality and biophysical modeling," *NeuroImage,* vol. 58, 2011.
[131] A. Roebroeck, E. Formisano, and R. Goebel, "Reply to Friston and David:: After comments on: The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution," *NeuroImage,* 2009.
[132] A. Roebroeck, E. Formisano, and R. Goebel, "The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution," *Neuroimage,* p. In Press, 2009.
[133] O. David, "fMRI connectivity, meaning and empiricism Comments on: Roebroeck et al. The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution," *NeuroImage,* p. In Press, 2009.
[134] K. Friston, "Dynamic casual modeling and Granger causality Comments on: The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution," *Neuroimage,* p. In Press, 2009.
[135] A. Roebroeck, A. K. Seth, and P. Valdes-Sosa, "Causal Time Series Analysis of functional Magnetic Resonance Imaging Data," in *JMLR: Workshop and Conference Proceedings*, 2011, pp. 65-94.
[136] W. Penny, K. Stephan, A. Mechelli, and K. Friston, "Comparing dynamic causal models," *NeuroImage,* vol. 22, pp. 1157-1172, 2004.
[137] P. Dayan and L. F. Abbott, *Theoretical neuroscience: Computational and mathematical modeling of neural systems*: MIT Press, 2001.
[138] G. Deco, V. K. Jirsa, P. A. Robinson, M. Breakspear, and K. Friston, "The dynamic brain: from spiking neurons to neural masses and cortical fields," *PLoS computational biology,* vol. 4, pp. 1-35, 2008.
[139] J. M. Kilner, J. Mattout, R. Henson, and K. J. Friston, "Hemodynamic correlates of EEG: a heuristic," *Neuroimage,* vol. 28, pp. 280-286, 2005.
[140] M. J. Rosa, J. Kilner, F. Blankenburg, O. Josephs, and W. Penny, "Estimating the transfer function from neuronal activity to BOLD using simultaneous EEG-fMRI," *Neuroimage,* vol. 49, pp. 1496-1509, 2010.
[141] A. C. Marreiros, S. J. Kiebel, and K. J. Friston, "Dynamic causal modelling for fMRI: a two-state model," *Neuroimage,* vol. 39, pp. 269-278, 2008.
[142] K. E. Stephan, L. Kasper, L. M. Harrison, J. Daunizeau, H. E. M. den Ouden, M. Breakspear, and K. J. Friston, "Nonlinear dynamic causal models for fMRI," *NeuroImage,* vol. 42, pp. 649-662, 2008.
[143] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association,* vol. 90, pp. 773-795, 1995.

[144] R. E. Kass and A. E. Raftery, "Bayes factors and model uncertainty," *Journal of the American Statistical Association,* vol. 90, 1995.
[145] C. C. Liu and M. Aitkin, "Bayes factors: Prior sensitivity and model generalizability," *Journal of Mathematical Psychology,* vol. 52, pp. 362-375, 2008.
[146] K. Friston and W. Penny, "Post hoc Bayesian model selection," *NeuroImage,* pp. 2089-2098, 2011.
[147] E. J. Wagenmakers, T. Lodewyckx, H. Kuriyal, and R. Grasman, "Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method," *Cognitive psychology,* vol. 60, pp. 158-189, 2010.
[148] I. Verdinelli and L. Wasserman, "Computing Bayes factors using a generalization of the Savage-Dickey density ratio," *Journal of the American Statistical Association,* vol. 90, pp. 614-618, 1995.
[149] J. D. Ramsey, S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, and C. Glymour, "Six problems for causal inference from fMRI," *Neuroimage,* vol. In press, 2009.
[150] D. J. C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural computation,* vol. 4, pp. 448-472, 1992.
[151] Y. Le Cun, J. S. Denker, S. A. Solla, R. E. Howard, and L. D. Jackel, "Optimal brain damage," *Advances in neural information processing systems,* vol. 2, p. 1990, 1990.
[152] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," *Advances in neural information processing systems,* pp. 164-164, 1993.
[153] S. J. Chang, C. S. Leung, K. W. Wong, and J. Sum, "A local training and pruning approach for neural networks," *International journal of neural systems,* vol. 10, pp. 425-438, 2000.
[154] J. Sum, C. Leung, G. H. Young, L. Chan, and W. Kan, "An adaptive Bayesian pruning for neural networks in a non-stationary environment," *Neural computation,* vol. 11, pp. 965-976, 1999.
[155] J. Sum, C. Leung, G. H. Young, and W. Kan, "On the Kalman filtering method in neural network training and pruning," *Neural Networks, IEEE Transactions on,* vol. 10, pp. 161-166, 1999.
[156] P. Hering and M. Simandl, "Structure Adaptation of Multi-Layer Perceptron Network for On-Line System Identification," in *17th World CongressThe International Federation of Automatic Control. Proceedings of,* Seoul, 2008, pp. 7022-7027.
[157] K. K. Kwong, J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, and R. Turner, "Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation," *Proceedings of the National Academy of Sciences,* vol. 89, pp. 56-75, 1992.
[158] T. Deneux and O. Faugeras, "Using nonlinear models in fMRI data analysis: model selection and activation detection," *NeuroImage,* vol. 32, pp. 1669-1689, 2006.
[159] W. Van der Zwaag, S. Francis, K. Head, A. Peters, P. Gowland, P. Morris, and R. Bowtell, "fMRI at 1.5, 3 and 7 T: characterising BOLD signal changes," *Neuroimage,* vol. 47, pp. 1425-1434, 2009.
[160] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI data by blind separation into

independent spatial components," *Human Brain Mapping,* vol. 6, pp. 160–188, 1998.
[161] L. Harrison, W. Penny, and K. Friston, "Multivariate autoregressive modeling of fMRI time series," *NeuroImage,* vol. 19, pp. 1477-1491, 2003.
[162] A. M. Sillito and H. E. Jones, "Corticothalamic interactions in the transfer of visual information," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* vol. 357, p. 17391752, 2002.
[163] J. Cabral, E. Hugues, O. Sporns, and G. Deco, "Role of local network oscillations in resting-state functional connectivity," *NeuroImage,* vol. 57, pp. 130-139, 2011.
[164] O. Faugeras, J. Touboul, and B. Cessac, "A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs," *Frontiers in computational neuroscience,* vol. 3, pp. 1-28, 2009.
[165] Y. K. Belyaev, "Analytic random processes," *Theory of Probability and its Applications,* vol. 4, pp. 402-409, 1959.
[166] R. Goebel, A. Roebroeck, D. S. Kim, and E. Formisano, "Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping," *Magnetic resonance imaging,* vol. 21, pp. 1251-1261, 2003.
[167] M. Havlicek, J. Jan, V. Calhoun, M. Brázdil, and M. Mikl, "Evaluation of Functional Network Connectivity in Event-related FMRI Data Based on ICA and Time-frequency Granger Causality," in *World Congress on Medical Physics and Biomedical Engineering*, Munich, 2009, pp. 716-719.
[168] M. Havlicek, J. Jan, V. D. Calhoun, and M. Brazdil, "Extended time-frequency granger causality for evaluation of functional network connectivity in event-related FMRI data," in *Engineering in Medicine and Biology Society, EMBC*, Minneapolis, 2009, pp. 4440-4443.