Hemodynamic modelling of BOLD fMRI - A machine learning approach

Daniel J. Jacobsen

Kongens Lyngby 2006
IMM-PHD-2006-173

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This Ph.D. thesis concerns the application of machine learning methods to hemodynamic models for BOLD fMRI data.

Several such models have been proposed by different researchers, and they have in common a basis in physiological knowledge of the hemodynamic processes involved in the generation of the BOLD signal. The BOLD signal is modelled as a non-linear function of underlying, hidden (non-measurable) hemodynamic state variables.

The focus of this thesis work has been to develop methods for learning the parameters of such models, both in their traditional formulation, and in a state space formulation. In the latter, noise enters at the level of the hidden states, as well as in the BOLD measurements themselves.

A framework has been developed to allow approximate posterior distributions of model parameters to be learned from real fMRI data. This is accomplished with Markov chain Monte Carlo (MCMC) sampling techniques, including 'parallel tempering', an improvement of basic MCMC sampling.

On top of this, a method has been developed that allows comparisons to be made of the quality of these models. This is based on prediction of test data, and comparisons of learnt parameters for different training data. This gives estimates of the generalization ability of the models, as well as of their reproducibility. The latter is a measure of the robustness of the learnt parameters to variations in training data. Together, these measures allow informed model comparison, or model choice.

Using resampling techniques, a measure of the uncertainty about the generalization ability and reproducibility of the models is also obtained.

The results show that for some of the data, the standard so-called 'balloon' model is sufficient. More complex data have also been designed, however, and for these, the stochastic state space version of the standard balloon model is shown to be superior, although an augmented version of the standard balloon model is not found to be an improvement for either data set.

Resumé

Denne Ph.D. afhandling omhandler anvendelsen af machine learning metoder til hæmodynamisk modellering af BOLD fMRI data.

Flere sådanne modeller er blevet foreslået af forskellige forskere, og de har en fælles basis i fysiologisk viden om de for hæmodynamiske processer, der har betydning for BOLD signalets dannelse. BOLD signalet modelleres som en ikke-lineær funktion af underliggende, skjulte (ikke-målelige) hæmodynamiske tilstandsvariable.

Fokus for dette arbejde har været udviklingen af metoder til at lære parametrene for sådanne modeller, både i deres traditionelle formulering, og i en tilstandsmodel formulering. I sidsnævnte indtræder støj i de skjulte variable, såvel som i selve BOLD målingerne.

Et sæt metoder er blevet udviklede, som tillader læring af tilnærmede a posteriori fordelinger af modelparametre fra fMRI data. Dette er gjort ved Markov chain Monte Carlo (MCMC) sampling teknikker, heriblandt 'parallel temperering', en forbedring af standard MCMC sampling.

Ovenpå dette er en metode udviklet, som gør det muligt at sammenligne kvaliteten af disse modeller. Dette gøres gennem prædiktion af test data, og sammenligniger af lærte parametre for forskellige træningsdata. Hermed estimeres modellernes generaliseringsevne, såvel som deres reproducerbarhed. Reproducerbarhed er et mål for hvor robuste, de lærte parametre er overfor variationer i træningsdata. Sammen giver disse mål mulighed for informerede model sammenligninger, eller modelvalg.

Ved hjælp af resampling-teknikker gøres det yderligere muligt at vurdere usikkerheden af estimaterne af generaliseringsevne og reproducerbarhed.

Resultaterne viser, at den såkaldte 'ballon model' er tilstrækkelig for nogle data. Men mere komplekse data er også blevet designet, og for disses vedkommende er tilstands-model udgaven af ballon modellen bedst. En udvidet udgave af ballon modellen har ikke vist sig at være en forbedring for de anvendte data.

Preface

This thesis was prepared at Informatics Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The work is funded by DTU. The project commenced in September 2003 and was completed in October 2006. Throughout the period, the project was supervised by professor Lars Kai Hansen. The thesis reflects the work done during the Ph.D. project and concerns machine learning approaches for BOLD fMRI hemodynamic modelling. This work concerns different aspects of knowledge-based, mathematical modelling of hemodynamic processes in the human brain, and methods for evaluation of the quality of such models.

The thesis consists of a summary report and a collection of 3 research papers written during the period 2003–2006, and published or submitted for publication elsewhere.

Lyngby, October 2006

Daniel J. Jacobsen

Papers included in the thesis

Parts of the work presented in this thesis have been published in the form of a conference article and one journal article. A second article has been submitted for publication in Neural Computation. These publications are listed here in the order they appear in the appendices to this thesis.

Conference paper (with referee)

Daniel J. Jacobsen, Kristoffer H. Madsen, Lars Kai Hansen, "Identification of non-linear models of neural activity in BOLD fMRI", *In 3rd IEEE International Symposium on biomedical Imaging: Macro to Nano*, pp. 952-955, 2006 (appendix A).

Journal papers

Daniel J. Jacobsen, Kristoffer H. Madsen, Lars Kai Hansen, "Bayesian model comparison in non-linear BOLD fMRI hemodynamics", *accepted for publication in Neural Computation (submitted July 2006)* (appendix B).

Daniel J. Jacobsen, "Deterministic versus stochastic dynamics in non-linear hemodynamic BOLD fMRI - a Bayesian comparison using unscented Kalman filtering", *submitted to Neural Computation*, October 2006 (appendix C).

Acknowledgements

I thank my supervisor, Lars Kai Hansen, for guidance and inspiration. I also thank Ole Winther for always being ready to talk about Monte Carlo methods and giving helpful advice. Kristoffer Hougaard Madsen for working with me in designing the stimulus used to generate data, for supplying and processing data for this project and for many constructive discussions. Ulla Nørgaard for unfailing administrative and practical help and advice. Thomas Deneux, for constructive discussions about hemodynamics and mathematics, and for enjoyable company. Egill Rostrup for supplying fMRI data from Hvidovre Hospital. Hermann Singer for his very kind and helpful correspondence during the implementation of the unscented Kalman filter.

I would also like to thank everyone in the ISP department for their encouragement and willingness to discuss on- and off-topic subjects, and for their pleasant company during the last three years.

Finally, I thank my family, May-Britt, Magnus and Anna Sofia, for their love and support.

Contents

Chapter 1

Introduction

"There is no scientific study more vital to man than the study of his own brain. Our entire view of the universe depends on it." - Francis H.C. Crick

There is a general consensus that the scientific study of the human brain is vital in every important sense. But it is only in the last few decades that the neuroimaging tools have become available that allow serious advances to be made in our understanding of how the brain works. Functional Magnetic Resonance Imaging (fMRI) is one of the most recently developed methods of neuroimaging, and arguably one of the most important for developing our understanding of brain function. This is especially true for the Blood Oxygenation Level Dependent (BOLD) variety of fMRI, a technique that has seen explosive growth in application since its invention in the early nineties [63],[62] - not so many years after the invention of MRI itself. The widespread adaptation of BOLD fMRI is due to the ability of these techniques to non-invasively measure spatially located signals in the brain that are closely related to local neural activity.

1.1 Contributions

This section gives an overview of the main contributions of this Ph.D. project.

1.1.1 Bayesian application of hemodynamic models

Non-linear, hemodynamic models are the focus of this work. Until now, learning of such models has been done using maximum likelihood (ML) or maximum a posteriori (MAP) approaches, see e.g. [23], [68]. In this project, approximations of the a posteriori distributions of the model parameters are sought. This yields more knowledge about the models and also allows more powerful predictive uses of the models. For one class of models, however, MAP learning has been used due to their associated high cost in computational time.

1.1.2 A framework for comparing hemodynamic models

Several different candidates models have been proposed and described in the literature, but very little work has been done to compare these models in a Bayesian sense. This is a crucial goal, as in any modelling domain, since it is the only way forward if better models are to be developed and if researchers are to know which model is best suited for different tasks.

A Bayesian model comparison framework has been developed in this project that takes into consideration both generalization ability and reproducibility, the latter measured in terms of the sensitivity of the posterior distributions (or MAP estimates) to changes in the data used for learning.

1.1.3 Model comparisons

Three different hemodynamic models have been compared using the above mentioned framework on two different real BOLD fMRI data sets.

1.2 Overview

Chapter 2 gives a brief introduction to the BOLD fMRI modality, the generation of the BOLD signal and describes the real data sets used in this project.

Chapter 3 then introduces the hemodynamic models to be investigated.

Chapter 4 describes the evaluation of the likelihood for models when there is no noise in the hemodynamic state space, which is the case for the original formulation of the hemodynamic models. It also describes the generation of synthetic data with these models.

Chapter 5 goes on to describe the evaluation of the likelihood for the models when noise is introduced into the hemodynamic state space, and describes the generation of synthetic data with these models.

Chapter 6 describes the Markov chain Monte Carlo methods used to learn the parameters of the models, in terms of obtaining approximate posterior parameter distributions. It also describes the simulated annealing method used to obtain maximum a posteriori (MAP) parameter estimates.

Chapter 7 describes the use of the learned posterior distributions or MAP parameters for prediction, and develops a framework for comparing model quality in terms of such predictions and in terms of the robustness of the learned parameters to changes in training data. This chapter also contains the main experimental results.

Finally, chapter 8 gives the concluding discussion, including an outlook on possible future research directions.

1.3 Origin of fMRI Images

Those images and figures in this thesis marked 'Courtesy of Scott Huettel' are taken with permission from
http://www.biac.duke.edu/education/courses/fall05/fmri/, and some of these are used in [39].

1.4 Nomenclature and symbols

Conventional mathematical symbols and are used throughout the thesis. In general, matrices are presented in uppercase bold letters (e.g. $\mathbf{A}$) and vectors are shown in lowercase bold letters (e.g. $\mathbf{x}$). Scalars are written in the normal typeface (e.g. x).

Probability density functions (PDF's) correspond to a stochastic variable, sometimes conditioned on another, and are evaluated at some point, i.e. numerical value. A complete notation could be for example $p_{x|y}(a|b)$ meaning: the PDF of the stochastic variable x conditional to the stochastic variable y evaluated at $x = a$ and $y = b$. Instead, the shorter $p(x|y)$ is preferred here to reduce clutter. It is unambiguous and only requires one to know that x and y stand for values (instances) of stochastic variables but also signify which PDF is referred to (here $x|y$).

List of main symbols

Mathematical symbols

$\mathbf{A}^T$	Transpose of the matrix $\mathbf{A}$.
$\mathbf{A}^-$	Pseudo-inverse of the matrix $\mathbf{A}$.
$\mu_{\mathbf{x}} \triangleq E[\mathbf{x}]$	Expectation of the stochastic variable $\mathbf{x}$.
$D(\mathbf{x}) \triangleq E[(\mathbf{x}-\mu_{\mathbf{x}})(\mathbf{x}-\mu_{\mathbf{x}})^T]$	Dispersion or 'variance' matrix of the stochastic vector $\mathbf{x}$.
$C(\mathbf{x},\mathbf{y}) \triangleq E[(\mathbf{x}-\mu_{\mathbf{x}})(\mathbf{y}-\mu_{\mathbf{y}})^T]$	Covariance matrix between stochastic vectors $\mathbf{x}$ and $\mathbf{y}$.
$\delta(x-\mu)$	The Dirac delta function centered on μ.
$\triangleq$	Definition, defining equation.

Physiological symbols

$v(t)$	Relative blood volume at time t.
$q(t)$	Relative deoxyhemoglobin content at time t.
$f(t)$	Relative blood inflow at time t.
$f_{out}(t)$	Relative blood outflow at time t.
$s(t)$	Stimulus signal at time t.
$a(t)$	Activation signal given to subject at time t.
$u(t)$	Neural activity at time t.
α	Inverse rigidity.
ϵ	Stimulus gain factor.
τ_0	Average transit time.
τ_s	Stimulus autoregulation time constant.
τ_f	Stimulus blood flow feedback regulation time constant.
E_0	Resting oxygen extraction fraction.
κ	Neural inhibition signal gain factor.
τ_u	Neural inhibition time constant.
τ_+	Balloon inflation time constant.
τ_-	Balloon deflation time constant.

Chapter 2

BOLD fMRI fundamentals

"More may have been learned about the brain and the mind in the 1990s - the so-called decade of the brain - than during the entire previous history of psychology and neuroscience." - Antonio R. Damasio

This chapter will give a brief introduction to the techniques and physics of BOLD fMRI. The key point is to show the BOLD signal's dependence on physiological variables, setting the scene for the hemodynamic models that are the main focus of this project.

2.1 MRI

Magnetic Resonance Imaging (MRI) is a relatively new medical imaging technique (the first commercial MRI scanners appeared around 1980). The basic principle relies on the quantum-mechanical behavior of hydrogen atoms - abundant in the form of water molecules in all brain tissue types - in the presence of controlled magnetic fields. The protons and neutrons of hydrogen atoms have a magnetic property called 'spin', with similar behavior to a dipole magnet[1]. The

[1] For an introduction to nucleic spins and quantum mechanics in general, see for example [33].

protons basically rotate, much like the earth rotates around the axis through its poles.

If an object is subjected to an external, uniform magnetic field B_0, then the spins of the hydrogen nuclei will align either in parallel or anti-parallel with the field, see figures 2.1A. These magnetic fields are typically 1.5 T (Tesla) or 3.0 T for scanners used for humans, while they can be stronger for scanners used on animals.

But the aligned spins can also be made to 'precess' around those directions, i.e. the spin axes can be brought to rotate (figure 2.1B). This 'wobbling' is similar to what happens when a spinning top toy looses momentum. All the spins have a characteristic precession resonance frequency ω_0 (around 64MHz for a 1.5T scanner) that depends on the strength of the external magnetic field as given by the Larmor frequency,

$$\omega_0 = \gamma B_0 \tag{2.1}$$

where γ is the so-called 'gyromagnetic ratio', a constant that depends on the atom ($42.58 \cdot 10^6$ $s^{-1}T^{-1}$ for Hydrogen).

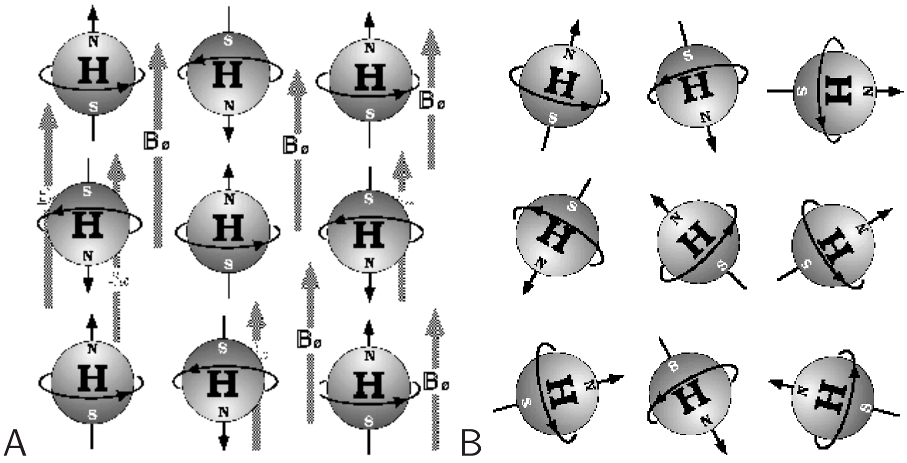


Figure 2.1: Spinning hydrogen protons. A: Spinning with spin axes aligned to the external magnetic field, B_0, after the application of a radio frequency pulse. B: Spinning with non-aligned spin axes.

If a radio frequency (RF) pulse with the same frequency is applied to the particles, the spins will absorb the energy[2] and precess coherently, i.e. the spins will be in phase.

After the application of the RF pulse, the absorbed energy then decays, the spins re-align with the external magnetic field, and the spins will dephase. The timing of all these events can be measured in terms of time constants. The realignment is measured as the so-called T1 signal, and the dephasing as the T2 signal. The signal of interest for fMRI is called T2* and measures changes in the T2 signal that are due to differences in magnetic susceptibility of the local tissue.

Through controlling the properties and timing of the radio pulses and magnetic fields, and the use of various signal processing techniques, it is possible to obtain 2-dimensional image slices of the tissue using any of these time constant signals (T1, T2 or T2*), which can then be combined into a 3-dimensional or 'volumetric' image of the entire object (e.g. a human brain). Perhaps the most widely used technique of generating the 2D fMRI images or 'slices' is 'gradient Echo Planar Imaging' (EPI); other methods are 'Spin Echo Imaging' and MPRAGE ('Magnetization Prepared Rapid Gradient Echo'). EPI has advantages of speed, contrast and relatively high signal-to-noise ratios (SNR) compared to other techniques. The drawback is that these images have rather low spatial resolution of around 3x3x3 mm.

The spatial orientation of 2D images is usually described by the terms axial, sagittal and coronal; see figure 2.2.

Two other fMRI scanner parameters are 'Field of View' (FOV) and 'Flip Angle' (FA). The FOV is the square 2D image area that contains the region of the brain to be scanned, given the location and orientation of the 2D slice planes. The FA is the angle to which the net magnetization is rotated when the RF pulse is applied.

An example of an MRI scanner is shown in figure 2.3.

[2] Incidentally, it is the large number of protons that allow for the generation of MRI signals, not the high energy of the radio pulses. Energy is proportional to frequency, and radio waves, with frequencies in the area of 10^7 s^{-1}, carry on the order of a trillion (10^12) times smaller energies than those used in X-ray or CT imaging. This is one of the advantages of MR imaging.

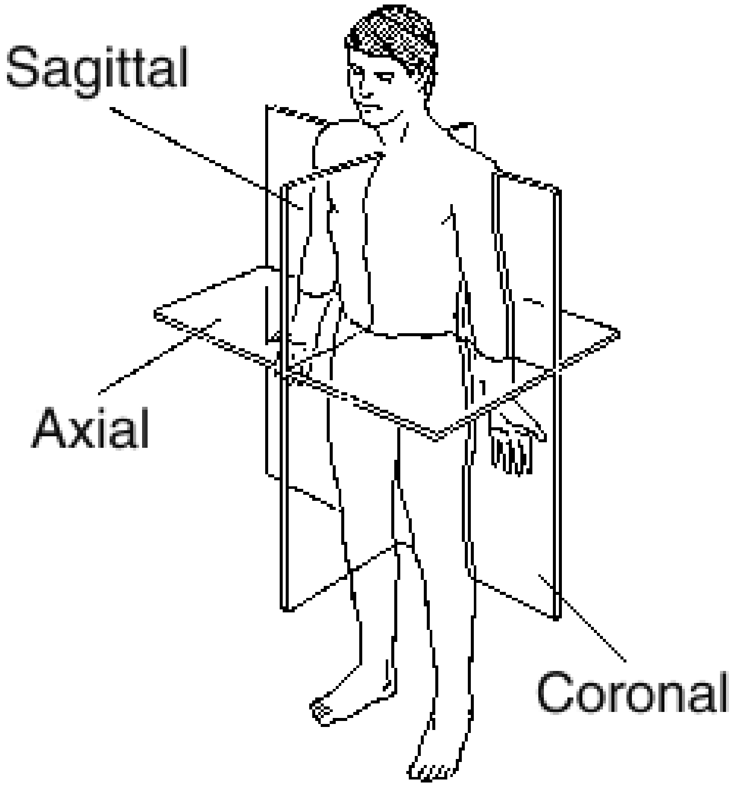


Figure 2.2: Illustration of the anatomical terms for plane orientation, defined relative to the human body.

2.2 BOLD fMRI

The discovery of Blood Oxygenation Level Dependent functional Magnetic Resonance Imaging - BOLD fMRI - in the early nineties by Ogawa and collegues [63],[62] was a major breakthrough in brain research. The key discovery is that when a region in the brain is activated, the local supply of oxygenated blood exceeds the increase in oxygen metabolism, resulting in an increase in oxygenation of the blood. The hemoglobin molecule that carries oxygen can exist in two states: oxyhemoglobin (oxygen is bound in the molecule) or deoxyhemoglobin (no bound oxygen). Very fortunately for brain imaging, the magnetic properties of these two states are different. Oxyhemoglobin is diamagnetic, and has very little influence on the local magnetic field. Deoxyhemoglobin, on the other hand, is paramagnetic and thus distorts the local magnetic field. This results in a shortening of the T2* relaxation time and a decrease in the MRI (T2*) sig-

Figure 2.3: Siemens Magnetom Trio MRI scanner. Photo courtesy of Siemens Medical Solutions.

nal. This signal is therefore called the 'Blood Oxygeneation Level Dependent' or BOLD signal. Therefore, with local brain activation, a decrease in deoxyhemoglobin means that the BOLD signal increases, which of course is nice from an intuitive point of view.

2.2.1 Physiological basis of BOLD fMRI

The BOLD signal in itself carries a lot of useful information, as it somehow relates to local brain activation. However, the relation between BOLD and brain activation is not clear, and neither is clear what 'brain activation' in this context precisely means. In this project, the relations that are sought are physiological, and it is therefore necessary to first give a physiological explanation of the BOLD signal and to relate it to underlying physiological processes.

In other words, a representation of the BOLD signal of the form

$$y(n) = g(\mathbf{x}(t_n); \theta, \mathbf{c}) \tag{2.2}$$

is desired, where $y(n)$ is the discretely sampled BOLD signal, $\mathbf{x}(t_n)$ is a vector of physiological variables evaluated at time t_n, θ is a vector of unknown parameters, and $\mathbf{c}$ a vector of known parameters or constants.

More than one version of (2.2) has been proposed. The first work is by Buxton et al. [17], but here an improved version developed by Obata et al. [61] is used. This is given by

$$y(n) = V_0[(k_1 + k_2)(1 - q(t_n)) - (k_2 + k_3)(1 - v(t_n)] \tag{2.3}$$

where the constants are given as

$$\begin{aligned} k_1 &= 4.3\nu_0 E_0 \mathrm{TE} \\ k_2 &= \epsilon r_0 E_0 \mathrm{TE} \\ k_3 &= \epsilon - 1 \end{aligned}$$

A detailed derivation of (2.3) is given in the appendix of [61], but the following may be noted. V_0 is the resting venous blood volume fraction, and is variously estimated at 0.02 and 0.03 (e.g. [23], [16]), and it should possibly be treated as a stochastic parameters. However, the choice here is made to consider it known and equal to 0.02 since the variance of the empirical estimates is very small. It enters into the equation because the BOLD signal is the sum of an intra-vascular and an extra vascular component, and the venules (small, randomly oriented collecting vessels) contain the most deoxyhemoglobin and are thus the most important intra-vascular BOLD signal source.

E_0 is the resting net oxygen extraction fraction, i.e. the fraction of oxygen extracted from the blood as it passes through the capillaries and venules at rest. It is considered here to be an unknown parameter. $v(t)$ and $q(t)$ are the physiological variables involved in the BOLD signal generation; $v(t)$ is the local blood volume of the venous compartment, and $q(t)$ is the total deoxyhemoglobin within this compartment, both relative to resting levels. ϵ[3] is the intrinsic ratio of the intravascular to the extravascular signal at rest, and is considered constant but depends on the scanner field strength.

TE is the 'echo time', a parameter of the EPI image formation technique, and is usually around 30-40ms. ν_0 and r_0 are quantities used in some linear approximations used to derive the BOLD equation and are also field strength dependent, see [61] for details. Values for the constants are given by Bandettini et al. in [8], and these values have been used in this project. They are - with the constants ν_0 and r_0 concatenated, but recalculated in order to keep the dependence on TE explicit,

$$\begin{aligned} k_1 &= 173.33 E_0 \mathrm{TE} \\ k_2 &= 47.67 E_0 \mathrm{TE} \\ k_3 &= 0.43 \end{aligned}$$

for 1.5 T field strength and

$$\begin{aligned} k_1 &= 346.67 E_0 \mathrm{TE} \\ k_2 &= 16.67 E_0 \mathrm{TE} \\ k_3 &= -0.5 \end{aligned}$$

for a field strength of 3.0 T.

It should be noted that the BOLD signal here is a percentage-wise change from a baseline, and not the absolute level.

[3] This ϵ is only used here and is not identical to the parameter ϵ used in the rest of the report.

2.3 Data sets

Two different real data sets have been used for the analysis in this project. In addition to these, synthetic data sets have been generated; these are described in chapters 4 and 5.

Each of the real data sets consists of a set of BOLD measurements or samples $Y^N = y_1,\ y_2, \ldots, y_N$ for a number of different voxels (brain locations corresponding to the spatial resolution of the scanner). The samples are recorded with a certain sampling time, in fMRI usually termed 'TR' (repeat time), that differs between the data sets.

Both data sets are focused on the regions of the brain that respond directly to visual stimulus, and are generated by presenting a subject in a scanner with a visual stimulus pattern on a display.

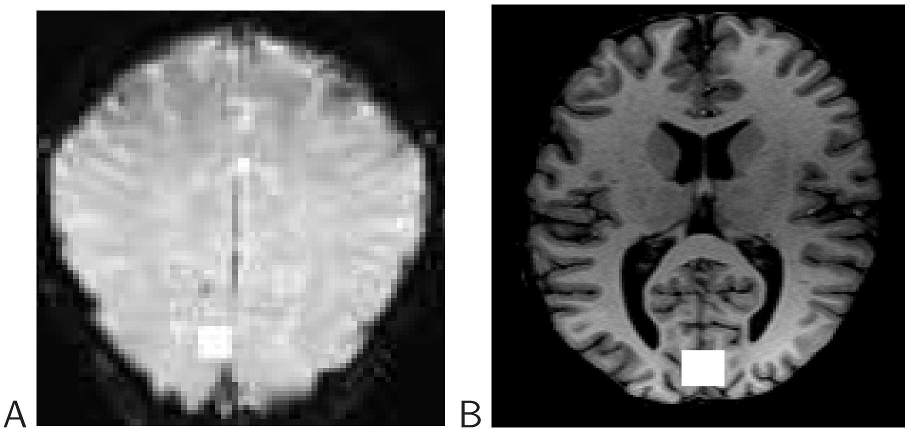


Figure 2.4: Regions of Interest, marked with white squares. Both images are axially oriented through the calcarine sulcus. A : Data set 1; B: Data set 2. For this data set, voxels from three adjacent slices were used.

From the raw data, regions of interest (ROI's) are selected as coherent collections of voxels that are seen to be activated by the stimulus given to the subject. For visual stimuli, these activations are robustly determined using the classical fMRI analysis tool, SPM2 (software available from http://www.fil.ion.ucl.ac.uk/spm/). Figure 2.4 shows the locations of the ROI's for the two data sets. The mean of the signals of all the ROI voxels is then used as the BOLD signal of each data set. This averaging increases SNR and is based on the assumption that for small ROI's, the BOLD signals are very similar, which is indeed found to hold through inspection.

The stimuli given to the subject are designed to include periods of rest before each activation (data set 1) or set of activations (data set 2). This allows for better preprocessing (see below), and has further significance for modelling. Each such stimulus-rest period is referred to as an *epoch*.

2.3.1 Data set 1

Data Set 1 was acquired by Dr. Egill Rostrup at Hvidovre Hospital on a 1.5 T Magnetom Vision scanner. The scanning sequence was a 2D gradient echo EPI (T2* weighted) with a 66-ms TE, a FA of 50°, a FOV of 230 mm and a sample time (TR) of 330ms. A single slice (2D image) was obtained in a para-axial orientation parallel to the calcarine sulcus. The calcarine sulcus, an anatomical structure in the occipital region of the brain, is shown in figure 2.5. It contains the primary visual cortex (V1). The visual stimulus consisted of a rest period of 20s of darkness (using a light fixation dot that helps the subject to fixate his eyes), followed by 10s of a full-field checker board reversing at 8 Hz, and ending by 20s of darkness. This flickering checker board stimulates the visual regions maximally. Ten separate runs were completed, and a total of 1000s recorded at each voxel. A ROI of 42 (7 by 6) significantly activated (as determined by SPM2 analysis) voxels from the visual cortex were selected.

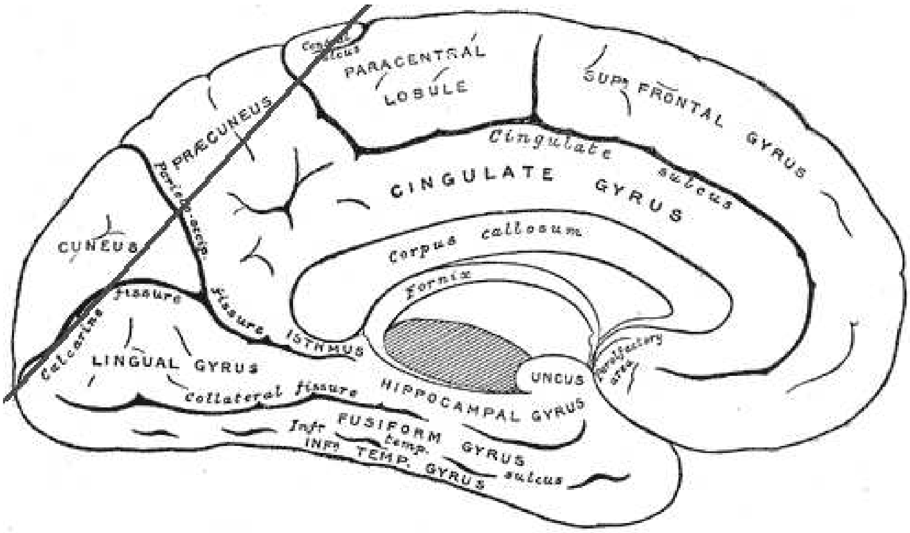


Figure 2.5: Schematic of a cross-section of the human brain, showing the location and orientation of the calcarine sulcus.

2.3.2 Data set 2

In order to 'challenge' the non-linear models, a random stimulus was designed with the purpose of generating a data set as non-linear as possible. Gamma-distributions were used to generate random stimulus durations (SD) and inter-stimulus intervals (ISI), see figure 2.6, and the signal is different for each epoch.

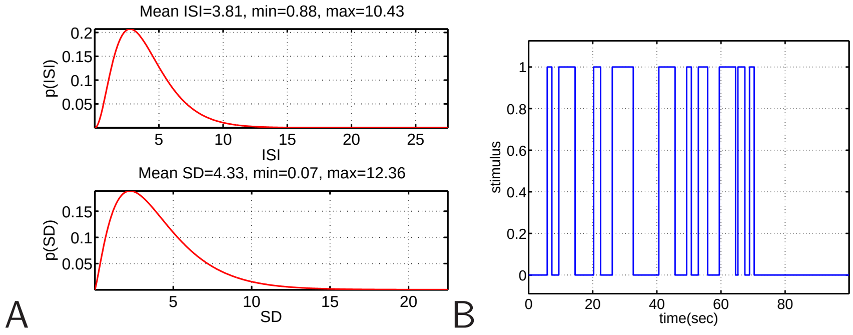


Figure 2.6: Stimulus design for data set 2. A: PDF's used to randomly generate the stimulus pattern, showing mean values and smallest and largest actual values. B: The resulting stimulus for the first epoch; note the resting period at the end.

The data was then acquired at Hvidovre Hospital, Denmark, using a 3T scanner (Magnetom Trio, Siemens). 1382 GRE EPI volumes each consisting of twelve 3mm slices oriented along the calcarine sulcus were obtained. Additional parameters where TR=725 ms, TE=30 ms FOV=192 mm, 64x64 acquisition matrix, FA = 82°. The stimulus consisted in a circular black/white flickering checkerboard (24 degrees horizontal, 18 degrees vertical) on a grey background. The checkers reversed black/white at 8 Hz. A ROI of 75 (25 from each of 3 slices) significantly activated (again as determined by SPM2 analysis), contiguous voxels in the visual cortex were selected, and the mean of these was used as the BOLD signal (see figure 2.4B).

2.4 Artifact removal: Preprocessing

When a BOLD signal is recorded, there are many different artifacts, i.e. unwanted signal components, in the raw recorded BOLD signal that must be coped with in some way. These artifacts are nuisances, because they correspond to variability in the data that is unrelated to the patterns of interest, i.e. the relation of the BOLD signal to local neural activity. The main physiological artifacts stem from heart beats, respiration and movement of the head. The scanner also has 'drift', i.e. a non-stationary additive component.

Fundamentally, there are two different approaches of artifact removal. The first is usually termed 'preprocessing' and consists of removing the artifacts before further modelling. The other is to include a model of the artifacts in a general model, so that the artifacts are handled simultaneously with the rest of the modelling. The two data sets used in this project was preprocessed in different ways. The advantage of the preprocessing approach is mainly simplification, in that the model needs no added complexity for built-in artifact handling. Also, it is possible to do very good preprocessing on the data so that most of the artifacts are eliminated while very little relevant information is lost from the signal.

A final piece of preprocessing that must be done if the data are to be used for hemodynamic modelling is normalization, i.e. the expression of the BOLD signal as a percentage-wise change in signal strength, since this is the target of these models.

2.4.1 Data set 1 preprocessing

The data were preprocessed according to the procedure described in [30]. A slight modification of the procedure was done in order to end up with signal values that correspond to percentage changes in the signal, see figure 2.7.

2.4.2 Data set 2 preprocessing

Data set 2 was preprocessed following the procedure described in [54]. The preprocessing consists of 2 separate steps: motion correction and nuisance effect modelling. Motion correction was performed using a 6 parameter linear (rigid body) transformation, which estimated movement parameters for each volume by minimizing the squared difference from the previous volume. To remove effects originating from scanner drift, movement and physiological noise, a procedure known as Nuisance Variable Regression (NVR) was used ([54]). This procedure aims to remove unwanted effects by modelling them using a multiple linear regression framework. The model consists of several nuisance effects (all with mean removed prior to estimation): discrete cosine transform (DCT) basis functions up a cut-off frequency of 1/128 Hz (a high-pass filter, 15 parameters), movement parameters and movement parameters squared to account for motion not corrected by rigid body realignment (12 parameters), movement parameters

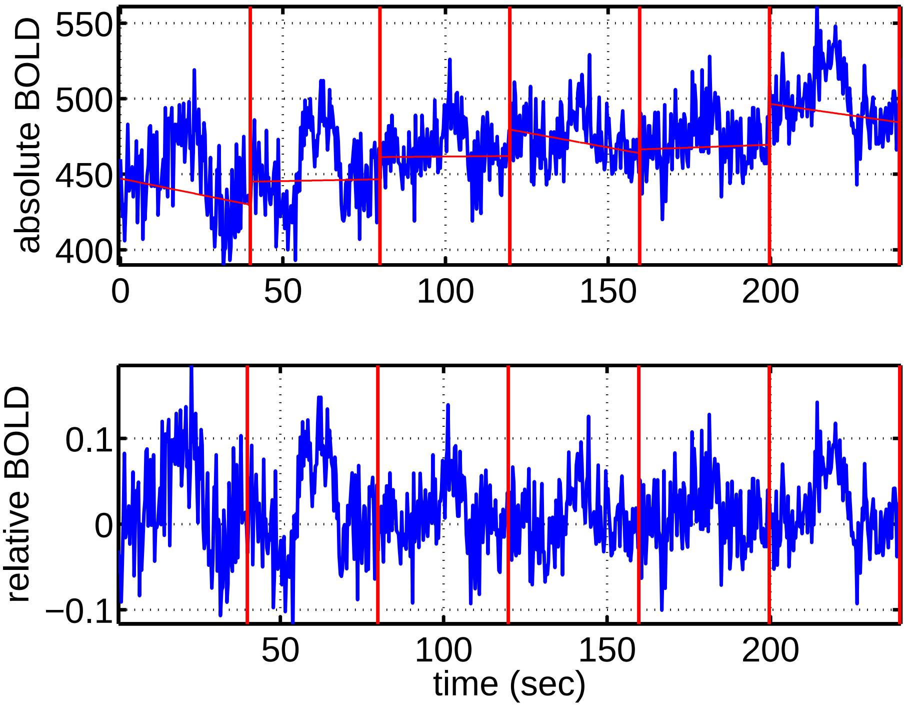


Figure 2.7: Preprocessing of data set 1. The top figure shows the raw data (first 6 epochs) after the initial 29 scanner saturation samples have been removed from each epoch. Also shown is the linear trend fitted for each epoch using only the resting-period samples. The bottom figure shows the result when the linear trend has been removed and the signal has been normalized to express the relative change in signal strength.

from previous volume and movement parameters from previous volume squared to account for spin history effects (12 parameters), Fourier expansion of aliased cardiac cycle parameters (10 parameters) and Fourier expansion of aliased respiratory cycle parameters (6 parameters). The preprocessing model thus has a total of 55 parameters, which were determined using maximum likelihood estimation assuming i.i.d. normally distributed noise (least squares estimation).

The entire BOLD signal (after preprocessing) is shown in figure 2.8. Note that the percentage wise change in the BOLD signal is only up to about 3 % compared to around 10 % for data set 1. This is due to the rapid stimulation; longer stimulus blocks create higher activations, and there was a concern about precisely this prior to scanning. However, due to a better scanner, the SNR is not much worse.

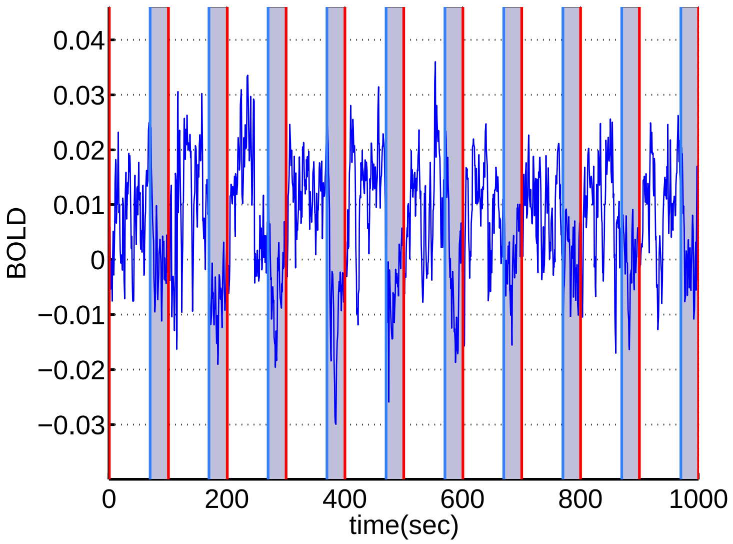


Figure 2.8: Data set 2 after preprocessing. The shaded areas correspond to the ends of the epochs when the stimulus is off.

2.5 BOLD fMRI and other brain imaging modalities

To round of this chapter, a short description of the relation of fMRI to some other important brain imaging modalities will be given to give some feel for the relative strengths and weaknesses of BOLD fMRI. These modalities are PET (Positron Emission Tomography), EEG (ElectroEncephaloGraphy), MEG (MagnetoEncephaloGraphy), single cell recording and optical imaging.

BOLD Based on blood oxygenation
Strength: High spatial resolution
Weakness: Low temporal resolution

PET Based on injected radioactive isotopes
Strength: Can measure various physiological functions
Weakness: Injection of radioactive isotopes, very low temporal resolution

EEG Measures electrical potential of cortical neurons
Strength: Very high temporal resolution

Weakness: Low spatial resolution

MEG Measures magnetic fields created by active brain regions
Strength: Very high temporal resolution
Weakness: Low spatial resolution

Optical Imaging Optically measures blood volume and blood volume changes
Strength: Very high spatial resolution at the scale of assemblies of neurons
Weakness: Generally can not be used on humans

Single cell recording Measures activity of single neurons using electrodes
Strength: Electrical response to stimuli of single cells
Weakness: Generally can not be used on humans

As can be seen from the above comparison, the power of BOLD fMRI is the ability to non-invasively measure a brain activity related signal with good spatial resolution, at the price of some temporal resolution. A good overview of different functional brain imaging methods is given in [37].

Chapter 3

Hemodynamic Models

"Present-day knowledge of the brain resembles in some ways earlier Europeans' knowledge of Africa. Explorers have mapped the coastline in detail, but the interior is mostly uncharted." - Douglas Tweed

In 1998, Buxton et al. proposed a model for the BOLD signal that was termed the 'balloon' model [17]. This model was later extended by Friston et al. [23], and again by Buxton et al. [16] in 2004 with new variants. These models all attempt to explain the BOLD signal in terms of underlying physiological processes. This chapter describes the various models and model variants. They share the basic property of being based on hemodynamics of the local brain tissue, i.e. the dynamics of physiological processes involved in blood volume and flow.

These models stand in contrast to the traditional linear models for BOLD fMRI (see e.g. [18] for an introduction), and may be seen as a more general, nonlinear description of hemodynamics than the traditional, linear 'hemodynamic response function' approach (see e.g. [31], [10]).

3.1 Model design and complexity

It is worth noting that the machinery of Bayesian analysis in itself does not tell one how to invent models (see [55], chapter 28, for a good discussion), but only how to use and compare given models. The models described here - both the structures of the models, the meaning of the parameters and the choice of non-linear functions - have been designed by researchers with knowledge of the relevant physiological processes. The level of complexity (loosely defined as flexibility to fit the data) of the models is determined by these design choices, so although the present models have quite few parameters, they are a priori expected to have a roughly suitable level of complexity for the modelling of BOLD fMRI signals.

3.2 Overview

The hemodynamic models have several components that are connected in a common way, see figure 3.1. First off is the stimulus function, $a(t)$. This is the stimulation that is given to the subject in the scanner. For both data sets used for this project, this is a visual stimulus (see section 2.3). The stimulus brings about neural activity, $u(t)$ (1). This neural activity affects the dynamics of the physiological state variables $\mathbf{x}(t)$ (2), creating a so-called hemodynamic response which is rather sluggish and non-linear. The state variables interact dynamically and non-linearly (3). There may or may not be internal noise in the physiological states (4); if included, such noise is a continuous time stochastic process, so it is referred to by its time increment, dW. The BOLD signal $y(t)$ is a function of the physiological states (5) with added measurement noise $v(t)$ (6).

Different hemodynamic models differ in some or all of these components and their connections, but the basic concept is the effect of increased neural activity on the blood supplied from the local capillaries. Figure 3.2 illustrates the profusion of these small blood vessels in brain tissue.

The internal interactions are defined in terms of parameterized ordinary differential equations, one equation corresponding to each variable, of the form

$$\frac{\delta x(t)}{\delta t} = f(\mathbf{x}(t), u(t); \theta) \tag{3.1}$$

where $x(t)$ may be any of the variables in $\mathbf{x}(t)$ and θ are the parameters. The BOLD signal then obtains as a function of $\mathbf{x}(t_n)$ ((2.2) repeated for convenience, ignoring the constants $\mathbf{c}$),

$$y(n) = g(\mathbf{x}(t_n); \theta) \tag{3.2}$$

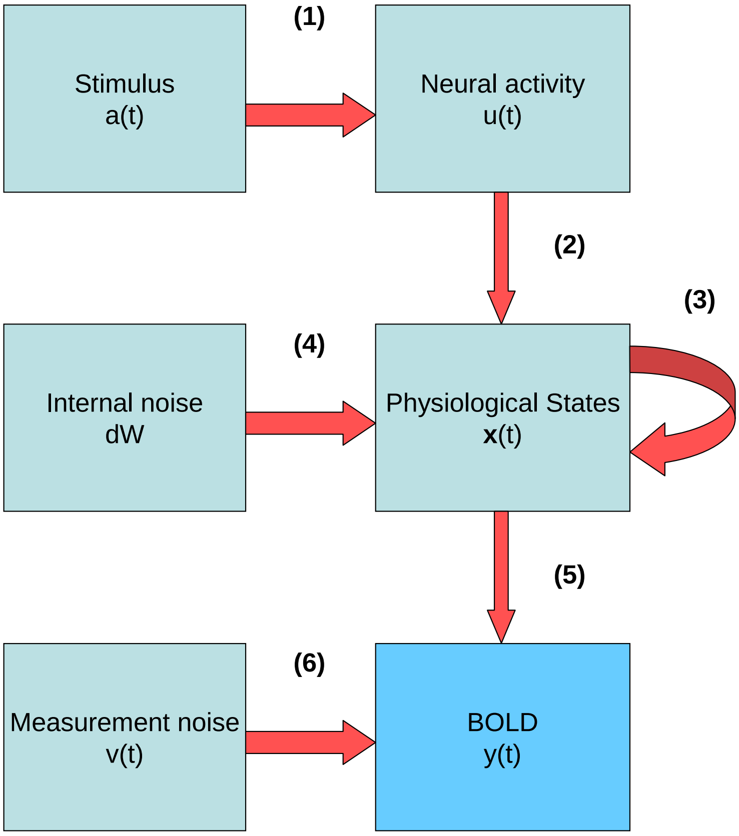
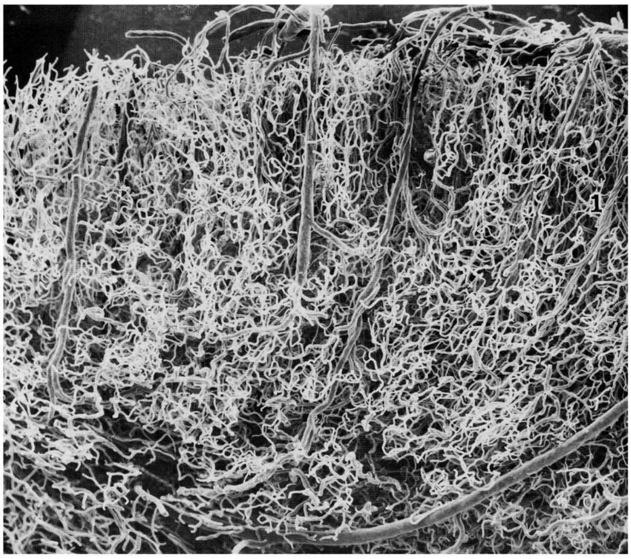


Figure 3.1: Overview diagram of hemodynamic models.

3.3 The standard balloon model

This model was originally developed in [17] and extended in [23]. It models the dynamics of the following physiological variables:



FUNCTIONAL MAGNETIC RESONANCE IMAGING, Figure 6.9 © 2004 Sinauer Associates, Inc.

Figure 3.2: An example image of arterioles and capillaries in the cortex of the human brain (courtesy of Scott Huettel).

- $v(t)$: Normalized venous volume
- $q(t)$: Normalized total deoxyhemoglobin content
- $s(t)$: Flow inducing signal
- $f(t)$: Inflow of blood

In addition to these, the neural activity $u(t)$ is assumed to be known, and is further assumed to be identical to the stimulus. The outflow of blood, $f_{out}(t)$, is an auxiliary variable that is given as a function of $v(t)$.

The specific differential equations for the standard balloon model are very well described and motivated in [23] and [16], so only a short description is given here. $v(t)$, the blood volume, of course depends on inflow and outflow of the 'balloon':

$$\frac{\partial v(t)}{\partial t} = \frac{1}{\tau_0} \left(f(t) - f_{out}(t) \right) \tag{3.3}$$

The parameter τ_0 is a time constant that equals the mean transit time for blood across the venous compartment. $q(t)$, the deoxyhemoglobin, is governed by a more complex equation:

$$\frac{\partial q(t)}{\partial t} = \frac{1}{\tau_0}\left[f(t)\frac{1-(1-E_0)^{1/f(t)}}{E_0} - v(t)^{(1-\alpha)/\alpha}q_t\right] \tag{3.4}$$

Here, the term

$$1-(1-E_0)^{1/f(t)} = E(t) \tag{3.5}$$

is the extraction fraction, the fraction of oxygen extracted from the blood as it flows through the balloon; it is an approximation given in [17] of the actual extraction fraction. The basic construction is that the first term in (3.4) is the increase in deoxyhemoglobin as new blood enters and has its oxygen extracted, while the second is the clearance of deoxyhemoglobin by the out flowing blood.

$s(t)$ is a somewhat artificial signal that is meant to subsume many neurogenic and diffusive signalling mechanisms that respond to neuronal activity, $u(t)$, and connect the latter to the hemodynamics. It is governed by

$$\frac{\partial s(t)}{\partial t} = \epsilon u(t) - s(t)/\tau_s - (f(t)-1)/\tau_f \tag{3.6}$$

The parameter ϵ thus controls the strength of the stimulus response to the neural activity. In addition there is negative auto-feedback in the second term, whereby $s(t)$ will oscillate towards zero if the neural activity ceases. The speed of this oscillation is controlled by the parameter τ_s. The final terms provides negative feedback from the inflow, controlled by the parameter τ_f.

The stimulus signal is assumed to directly control the outflow in that the time derivative of the latter equals $s(t)$,

$$\frac{\partial f(t)}{\partial t} = s(t) \tag{3.7}$$

The blood outflow $f_{out}(t)$ follows

$$f_{out}(t) = v(t)^{1/\alpha} \tag{3.8}$$

which is not a differential equation; as stated above, $f_{out}(t)$ may thus be considered an auxiliary variable in this model. This relationship is the basis of the 'balloon' term and means that the venous compartment expels blood faster when it is distended. α is an 'inverse stiffness' parameter, which is assumed to be between 0 and 1 (higher values would mean that the balloon expelled blood slower as it distended, but this invalidates the physiology behind the design of this function).

An overview of the structure of the standard balloon model is given in figure 3.3

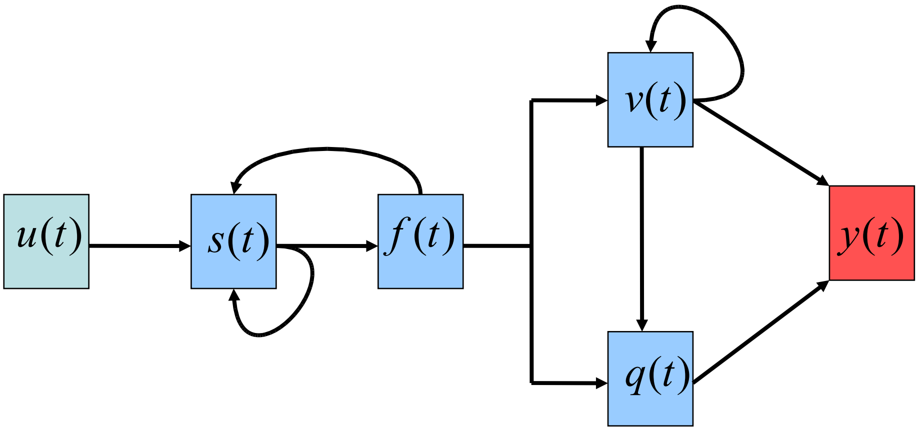


Figure 3.3: Diagram of the interactions in the hemodynamic models. The details of each of the interactions are described in the main text.

See also [74] for a good overview of some hemodynamic models.

3.4 A note on neural activity

The term 'neural activity' as used here is not a physiologically well defined concept. It has been shown ([52], [51], [53]) that the BOLD signal is closely related to the so-called local field potential (LFP) that reflects local processing of populations of neurons. The LFP is thought to be a weighted sum of the membrane potentials, both excitatory and inhibitory, of all the neurons in this population, mainly reflecting synaptic activity (resulting from input from other neurons) localized to dendrites and soma (see figure 3.4), although action potentials (information carrying electric waves travelling along the membrane) may also contribute to the LFP. This means that the neural activity as used in hemodynamic BOLD models ($u(t)$) can loosely be interpreted as reflecting the local, population-level processing of neural input rather than long-range communication with other brain regions.

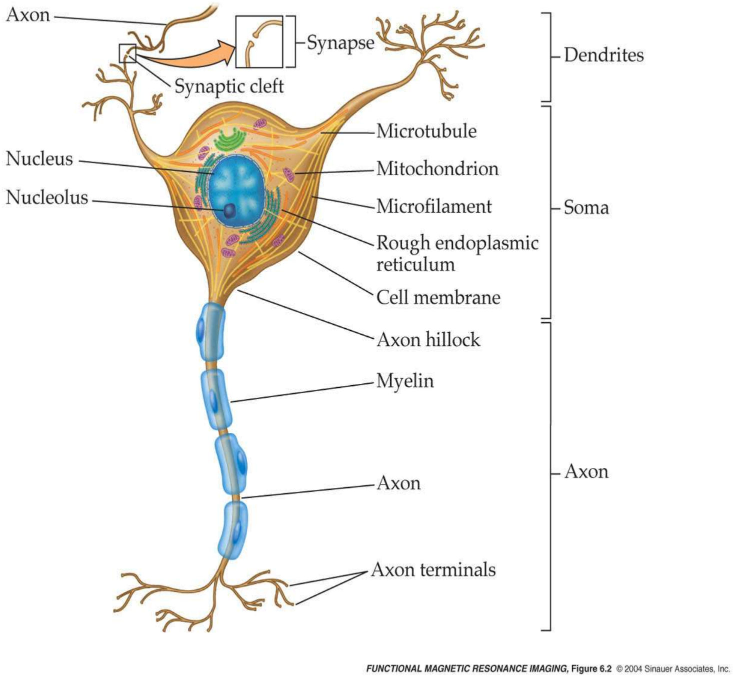


Figure 3.4: A neuron. The dendrites (top) receive input from other neurons, and the soma is the main body, containing the main 'machinery' of the cell (courtesy of Scott Huettel).

For more information on the relation between BOLD fMRI and neural activity, see [36], [9], [6], [49] and [5].

3.5 The augmented balloon model

Buxton et al. [16] have recently introduced an alternative dynamical model for the neural activity $u(t)$ and its connection to the stimulus $a(t)$, as well as a more complex relationship between blood outflow $f_{out}(t)$ and volume $v(t)$. The combination of these extensions with the standard balloon model is referred to here as the 'augmented balloon model'.

3.5.1 Non-linear neural activity

The neural activity is proposed to follow

$$\begin{aligned} u(t) &= a(t) - I(t) \\ \frac{dI}{dt} &= \frac{\kappa u(t) - I(t)}{\tau_u} \end{aligned} \tag{3.9}$$

where $a(t)$ is the square wave stimulus function and $I(t)$ is an inhibitory feedback signal. κ is a gain factor for the inhibition signal, and τ_u is a time constant that determines how quickly the neural activity is inhibited. This leads to an adaptive neural response to sustained stimuli. An example of $u(t)$ corresponding to a single one-second pulse of stimulus with $\kappa = 2$, $\tau_u = 1$ is shown in figure 3.5.

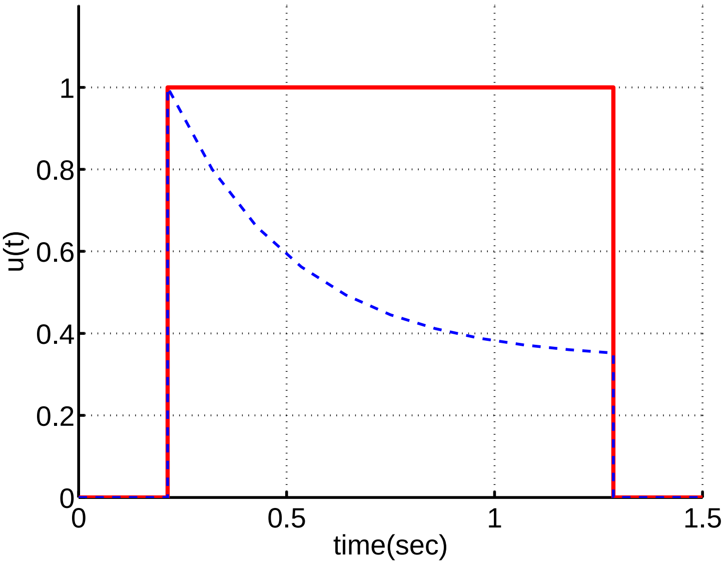


Figure 3.5: Response of the non-linear neural model $u(t)$ (dashed curve) to a one-second stimulus.

Note that the square pulse model used in the standard balloon model ($u(t) = a(t)$) obtains as a special case of this non-linear model as $\kappa/\tau_u \to 0$.

3.5.2 Visco-elastic outflow model

In addition to the non-linear neural model described above, it was proposed in [16] that the relation between outflow and volume in the standard balloon model (3.8) is based on steady-state conditions and could be modified for dynamic conditions. The proposed relation is

$$f_{out}(v(t)) = v(t)^{1/\alpha} + \tau \frac{\delta v(t)}{\delta t} \tag{3.10}$$

which means that the 'balloon' will transiently resist changes (for example due to so-called visco-elastic effects, hence the model label) and only after some time (controlled by τ) conform to the steady-state relationship (3.8). Also, τ is proposed to be potentially different during inflation and deflation:

$$\tau = \begin{cases} \tau_+ & \frac{\delta v(t)}{\delta t} \geq 0 \\ \tau_- & \frac{\delta v(t)}{\delta t} < 0 \end{cases} \tag{3.11}$$

Inserting (3.3) into (3.10) gives

$$\begin{aligned} f_{out}(t) &= v(t)^{1/\alpha} + \tau \frac{1}{\tau_0} \left[f(t) - f_{out}(t) \right] \\ &= \frac{v(t)^{1/\alpha} + \frac{\tau}{\tau_0} f(t)}{1 + \tau/\tau_0} \\ &= \frac{\tau_0 v(t)^{1/\alpha} + \tau f(t)}{\tau_0 + \tau} \end{aligned} \tag{3.12}$$

The problem with this is that to see if τ should be τ_+ or τ_-, it is necessary to know $\frac{\delta v(t)}{\delta t}$, but that in turn requires knowing $f_{out}(t)$. The solution to this coupling is to add $f_{out}(t)$ as a fifth state space variable. To obtain its derivative with respect to time, inserting (3.3) into (3.10) and differentiating,

$$\begin{aligned}\frac{\delta f_{out}(t)}{\delta t} &= \frac{1}{\alpha}v(t)^{1/\alpha-1}\frac{\delta v(t)}{\delta t} + \frac{\tau}{\tau_0}\left[\frac{\delta f(t)}{\delta t} - \frac{\delta f_{out}(t)}{\delta t}\right] \\ &= \frac{\frac{1}{\alpha}v(t)^{1/\alpha-1}\frac{\delta v(t)}{\delta t} + \frac{\tau}{\tau_0}\frac{\delta f(t)}{\delta t}}{1+\frac{\tau}{\tau_0}} \\ &= \frac{\frac{\tau_0}{\alpha}v(t)^{1/\alpha-1}\frac{\delta v(t)}{\delta t} + \tau\frac{\delta f(t)}{\delta t}}{\tau_0+\tau}\end{aligned} \tag{3.13}$$

Inserting (3.3) and (3.7) finally gives

$$\frac{\delta f_{out}(t)}{\delta t} = \frac{\frac{1}{\alpha}v(t)^{1/\alpha-1}(f(t)-f_{out}(t)) + \tau s(t)}{\tau_0+\tau} \tag{3.14}$$

When solving this new system, the sign of $f(t) - f_{out}(t)$ must first be tested to see if τ should be τ_+ or τ_-, so this is done whenever $\frac{\delta \mathbf{x}}{\delta t}$ is calculated for this model.

The augmented balloon model is somewhat more complex, with 4 additional parameters (κ, τ_u, τ_+ and τ_-), as well as an extra dimension in the hidden state space. The initial resting state is extended to $\mathbf{x}_0 = [1\ 1\ 1\ 0\ 1]^T$, i.e. blood outflow at resting level. An overview of the structure of the augmented balloon model is given in figure 3.6

3.6 A priori parameter distributions

In order to learn the parameters of these models, an a priori distribution $p(\theta)$ for the parameters must be chosen. There are many approaches to making this choice. Generally it is important that the priors are as non-informative as possible, and yet they should reflect any prior beliefs held about the parameters. Priors may also be designed with the purpose of limiting the complexity of the model ('regularizing priors'), but in the present case there actually exists prior physiological knowledge, so the choice is made to build the prior distribution on that knowledge (see [66] for a good discussion of the importance of priors).

The prior is assumed to factorize into a product of univariate priors,

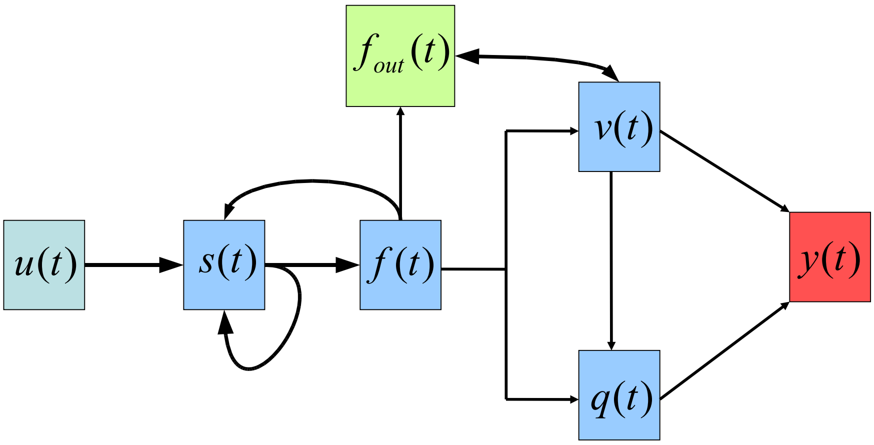


Figure 3.6: Diagram of the interactions in the augmented ballon model; note the additional variable, $f_{out}(t)$. The details of each of the interactions are described in the main text.

$$p(\theta) = \prod_{i=1}^{L} p(\theta_i)$$

where L is the number of parameters. This seems reasonable as there is little or no reason to believe - a priori - that the parameters are correlated.

3.6.1 Beta-distribution

For these parameters it is possible to specify more or less vague lower and upper limits for conceivable settings ([23],[16]). The family of scaled beta distributions[1] is therefore used for the priors of these parameters, as they are well suited to

[1] A standard beta-distribution with a scaled variable.

design appropriately flat distributions with upper and lower limits, and allow a natural control over the shape of the distribution. The scaled beta distribution has three parameters s, u_1 and u_2 that control its range, mode and shape:

$$p(\theta|s, u_1, u_2) = \frac{1}{Z(s, u_1, u_2)}(s\theta)^{u_1-1}(1 - s\theta)^{u_2-1}$$

with the normalizing factor

$$Z(u_1, u_2) = s\frac{\Gamma(u_1)\Gamma(u_2)}{\Gamma(u_1 + u_2)}$$

These parameters may be referred to as 'hyper-parameters', since they are parameters for the distribution of other parameters. See figure 3.7 and 3.8. The design of the priors is done by first choosing an upper range (all priors have a lower limit of zero). The scale is then the inverse of the range. The desired mode (peak) θ_{max} of each distribution is then set, followed by the 'peakedness', determined by u_2 and depending on how strong the prior belief is. u_1 is then given as

$$u_1 = \frac{\theta_{max}}{1 - \theta_{max}}(u_2 - 1) + 1;$$

Table 3.1 shows the prior parameters for all of the hemodynamic parameters.

3.6.2 Notes on the design of the priors

The parameter α is the inverse stiffness of the 'balloon' compartment modelling mainly the local venules. It is often simply set to 0.4 according to [16]. This indicates that large perturbations from this value are empirically and physiologically unexpected, and it is in any case constrained to lie between 0.0 and 1.0 (higher values would lead to the unphysiological effect of the volume increasing exponentially with flow increase). The closer it gets to 0.0, the stiffer the venules become, finally resisting any change in volume no matter how high the inflow of blood. The prior chosen for alpha reflects a rather strong belief that it should be close to 0.4.

ϵ may be termed the 'stimulus gain factor' and reflects both the amplitude of the local neural activity and the efficiency with which it is able to elicit a

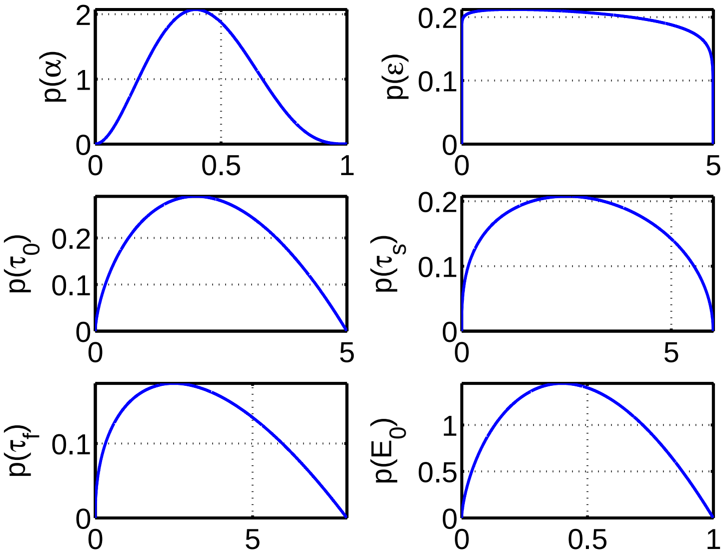


Figure 3.7: Prior distributions for the hemodynamic parameters. $p(\epsilon)$ is the least informative, and $p(\alpha)$ the most.

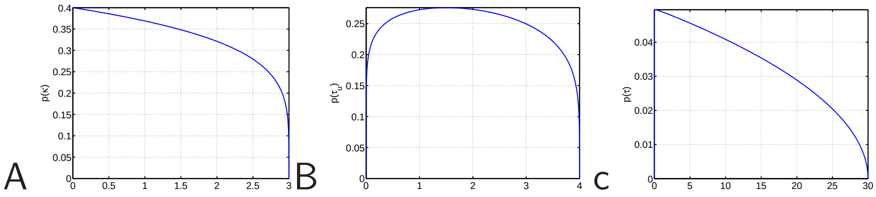


Figure 3.8: Prior distributions for the parameters of the augmented balloon model.

hemodynamic response. Its main purpose is to allow an appropriate scaling to a given data set. With the types of data that have been used in this project, a suitable range for ϵ is from close to zero up to around 2.0.

τ_0 is the average transit time of blood through a voxel. It is independent of voxel size, as the flow through a voxel also depends on its size. The value of τ_0 is determined by the blood flow (ml/min) and the resting venous blood volume fraction, V_0. As the flow is assumed to be around 60 ml/min in [16], flows less than half and more than double this value are considered very unlikely.

τ_s is the time constant for the autoregulation of the stimulus signal. The prior is based on the findings in [23], where τ_s is found to vary roughly from 1.2 to

2.2. Since this is based on certain voxels in a certain task, the prior is chosen to allow somewhat higher variation, with a cut-off at 6.0.

τ_f is the time constant for the feedback regulation from the blood flow on the stimulus signal. In [23], it was found to vary from around 2.0 to 3.2, and it is less abstract or contrived than τ_s. Following the damped oscillator argument given in [23], the resonance frequency of the feedback system is

$$\omega = 1/(2\pi\sqrt{\tau_f})$$

giving around $\omega = 0.1$ Hz at $\tau_f = 2.46$, the empirical mean. Allowing τ_f to vary from zero to 8 corresponds to a variation in this frequency from infinity (leading to instant suppression of $s(t)$ and thus of any BOLD response) to 0.056, around half of the empirical result in [23]. This range seems appropriate for the prior for τ_s.

E_0 is the resting oxygen extraction fraction. It is constrained to $0 < E_0 < 1.0$, and according to [23], known values vary between 0.2 and 0.55, whereas in [16], 0.4 is given as a typical value. The mode of the prior for E_0 is thus set to 0.4, and the shape chosen to correspond roughly to the normal variation.

In [16] ranges are given for κ and τ_u, but these ranges are not discussed in the text. A possible reason for limiting κ's upper bound at 2.0 might be that much higher values all lead to a neural activity shape that is basically square (only with lower amplitude), and thus carry no further information. As smaller values lead to the standard, square neural activity model, the mode is set very close to zero; the cut-off is set at $\kappa = 3.0$. The time constant, τ_u, is probably more physiologically based, as the expected time scale of any neural adaptation is not likely to be more than a few seconds. The cut-off for τ_u is set to 4 seconds, with a suitably flat shape reflecting the high level of prior uncertainty.

Interestingly, it was found that with the augmented neural activity models and uniform priors, κ and ϵ become highly correlated in the posterior, which is due to a mathematical invariance: increasing κ reduces the power of the neural pulses, in turn reducing the predicted BOLD signal; increasing ϵ mitigates this effect. The *beta*-priors actually used remove this correlation (see figures 3.7 and 3.8).

According to [16], values for the time constants during either inflation (τ_+) or deflation (τ_-) higher than 30 seconds are onsidered very unlikely, but no bid is given as to any values within the range ($0 - 30s$) that should be considered more likely than others, a priori. Since the behavior of $f_{out}(t)$ goes increasingly to that corresponding to the simpler standard balloon model as τ goes to zero, the prior is shaped as a monotonously decreasing function with a cut-off at 30 seconds.

For the observation noise variance no prior assumptions are made other than that it must of course be positive. A prior could be designed based on the fact that the variance of the whole BOLD signal is an upper limit for the observation noise variance, but as the noise variance has been consistently correctly estimated for synthetic data, there seems to be little need for such a bound. The prior for the observation noise is therefore simply set to a constant for positive values, $p(\sigma_w^2) = 1$ for $\sigma_w^2 > 0$ and $p(\sigma_w^2) = 0$ for $\sigma_w^2 \leq 0$.

θ	s	u_1	u_2	θ_{max}
α	1.0	3.0	4.0	0.4
ϵ	1/5	1.025	1.1	1.0
τ_0	1/5	1.67	2.0	2.0
τ_s	1/6	1.36	1.5	2.5
τ_f	1/8	1.45	2.0	2.5
E_0	0.33	1.67	2.0	0.4
κ	1/3	1.0	1.2	0.0
τ_u	1/4	1.12	1.2	1.5
τ	1/30	1.1	1.1	0.1

Table 3.1: (Hyper-) parameters for the Beta-distributed hemodynamic priors.

Chapter 4

Deterministic state space models

"It's choice - not chance - that determines your destiny." - Jean Nidetch

The likelihood of a hemodynamic model is defined as a function of its parameters,

$$\mathcal{L}(\theta) \triangleq p(D|\theta) \tag{4.1}$$

where the data consist of a set of BOLD samples, $D = Y^N = \{y_1, y_2, \ldots, y_N\}$, measured at discrete times $\{t_0, t_1, \ldots, t_N\}$. These samples may be further considered to be divided into independent epochs (as discussed in chapter 2), but this structure is omitted here for clarity.

The likelihood is central for learning and model comparison, and its evaluation is the subject of this and the following chapter.

To expound the structure of the likelihood, it is helpful to consider the hemodynamic model as a *state-space* model, so that the physiological state variables are referred to as *hidden variables*, in the sense that they are not measurable, and the BOLD signal samples are referred to as *observation variables* in the sense that they are measurable and are given as a function of the hidden variables. This may be written as an equation for the time dynamics of the hidden states,

$$\frac{\delta \mathbf{x}(t)}{\delta t} = f(\mathbf{x}(t), u(t), \theta) \tag{4.2}$$

and the observation function,

$$y_n = g(\mathbf{x}(t_n), \theta) + \epsilon \tag{4.3}$$

for the BOLD signal, where ϵ is considered to be distributed as $\mathcal{N}(0, \sigma_w^2)$ independently of time (i.e. the ϵ's are identically and independently normally distributed). A general state space model is illustrated in figure 4.1

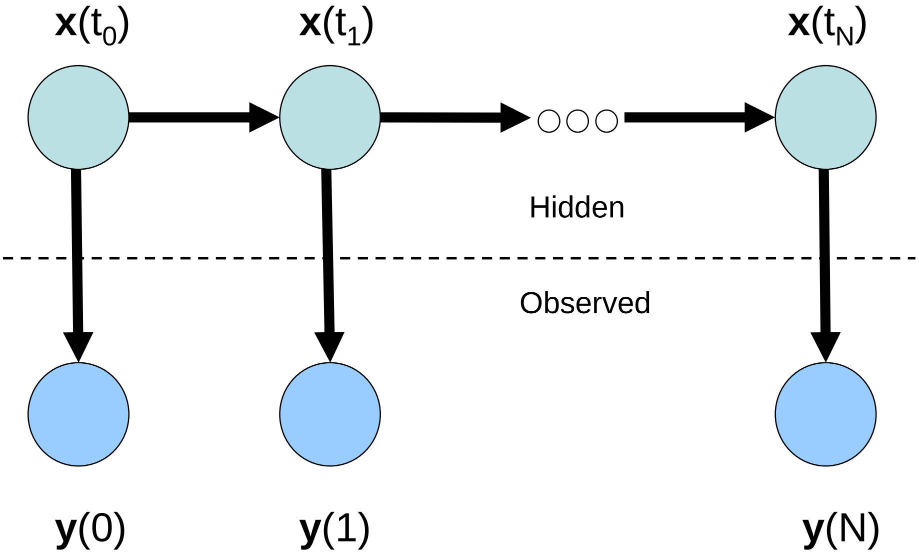


Figure 4.1: Graphical model diagram of a state space model.

The arrows in this diagram correspond to statistical dependencies, showing that the hidden state at time t_n, $\mathbf{x}(t_n)$ depends on the previous state $\mathbf{x}(t_{n-1})$, while the observation at time t_n, $y(n)$ depends on the hidden state at that time, $\mathbf{x}(t_n)$.

Equation (4.1) can thus be expanded as

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} p(y_n) = \prod_{n=1}^{N} \int p(y_n|\mathbf{x}(t_n)p(\mathbf{x}(t_n)|\mathbf{x}(t_{n-1})d\mathbf{x}(t_n)) \tag{4.4}$$

where $p(\mathbf{x}(t_1)|\mathbf{x}(t_0))$ is defined as the distribution of $\mathbf{x}(t_1)$ for convenience. If there is no noise in the evolution of $\mathbf{x}(t)$, then the hidden states are deterministic variables, which may be expressed by the corresponding probability density functions being delta functions, i.e.

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} \int p(y_n|\mathbf{x}(t_n)\delta(\mathbf{x}(t_n) - \hat{\mathbf{x}}(t_n))d\mathbf{x}(t_n)) \tag{4.5}$$

where $\hat{\mathbf{x}}(t_n)$ is the calculated value of $\mathbf{x}(t)$ at time t_n. This value is obtained by solving the ordinary differential equations, going in time from one observation time point t_{n-1} to the next, t_n, starting with the known initial condition $\mathbf{x}_{t_0} = \mathbf{x}_0$. This means that the likelihood factors in the following way:

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} \mathcal{N}(g(\hat{\mathbf{x}}(t_n)) - y(n); 0, \sigma_w^2) \tag{4.6}$$

where $g(\hat{\mathbf{x}}(t_n)) - y(n)$ are the residual errors of the model prediction.

4.1 Solving the ordinary differential equations

The only difficulty in the evaluation of this likelihood is then to obtain the values for the deterministic hidden states at the time point for the next observation. These values are given by solving the ordinary differential equations (ODE's) of the model (see the previous chapter). Since these form a coupled non-linear differential system, they must be solved numerically.

4.1.1 Runge-Kutta methods

The most basic method of solving an ODE is Euler's method. This simply assumes that the gradient $\frac{\delta \mathbf{x}(t)}{\delta t}$ remains constant between two observation time points, so that

$$y(t + \Delta t) = y(t) + \Delta t \frac{\delta \mathbf{x}(t)}{\delta t} \tag{4.7}$$

Unless the time interval Δt is very short, this assumption is unlikely to hold, which can yield very inaccurate results. To get accurate results with this method thus requires sub-division into suitably small time steps, which is inefficient. So-called mid-point methods and Heun's method refine Euler's method by evaluating the gradient at several time points, and this idea is generalized by the Runge-Kutta methods [1]. These predict $y(t + \Delta t)$ as a weighted sum of the gradients at intermediate time points, $t_k,\ k = 1..K$,

$$y(t + \Delta t) = y(t) + \Delta t \sum_{k}^{K} \gamma_k \frac{\delta \mathbf{x}(t)}{\delta t}\bigg|_{t=t_k} \tag{4.8}$$

where K is the order of the approximation. This is a much more powerful method than Euler's, but it is not clear how many steps should be taken between observations to obtain a desired accuracy. A variable step size method is therefore preferred that is able to speed up (increase step size) or slow down the integration depending on the behavior of the system. Here, an embedded Runge-Kutta method is used. An initial step size is chosen, and a prediction is made using both a 4th-order Runge-Kutta method and a 5th-order one. The difference between the estimates of these two methods is then used to update the stepsize: if the difference is very small, the 5th-order prediction is accepted and the step size increased. If it is larger than a set tolerance, the prediction is rejected and the step size reduced. In any other case, the step size is preserved and the prediction continues from the new point.

The values used for the constants, $\gamma_k,\ k = 1..K$ are those given in ([65], chapter 16), and the tolerances are set to different values for each dimension of $\mathbf{x}(t)$ (due to different scales of the dimensions) to give prediction errors smaller than 1 %; this was tested on various synthetic data with different parameter values (θ) and neural activity functions $u(t)$.

This variable step size method was compared with the basic Euler method and was shown to use average step size of around 250ms, while Euler's method used around 5 ms, a saving of computational time of around factor of 50. Figures 4.2 and 4.3 show an example of a solution for a simple stimulus.

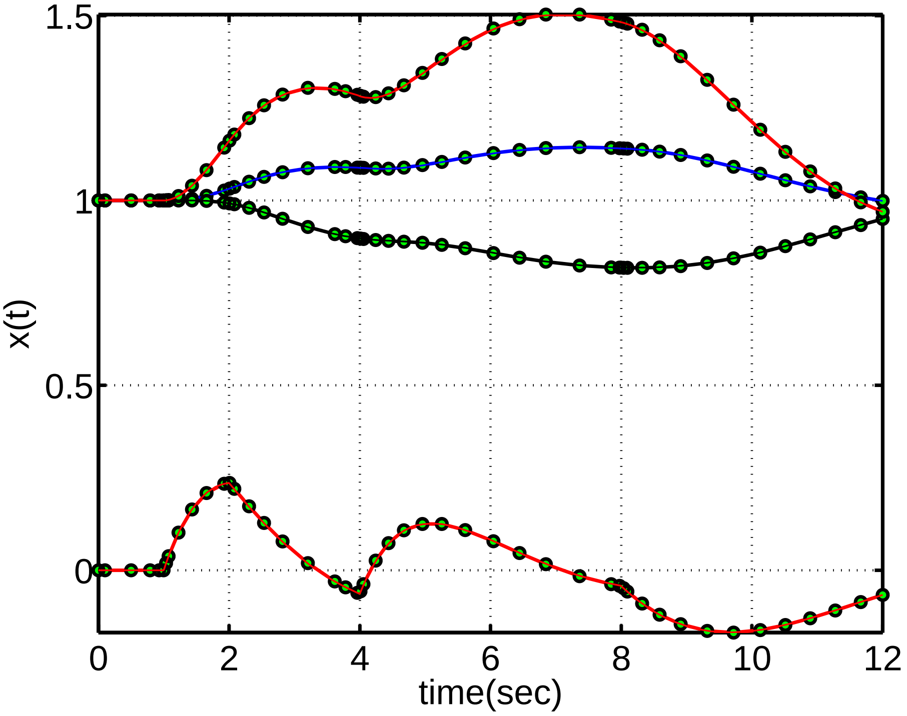


Figure 4.2: $\mathbf{x}(t)$ solution from the variable step-size Runge-Kutta solver (standard balloon model). The circles mark the solution points. From top to bottom: $f(t)$, $v(t)$, $q(t)$ and $s(t)$. The corresponding neural activity is seen in figure 4.3. Note the higher density of points as a function of higher curvature, and the non-linear effect of the second stimulus.

4.2 Handling discontinuities

The non-linear neural activity of the augmented balloon model (equation 3.9 in section 3.5.1) is discontinuous at all stimulus onset and offset times (see figure 4.3). However, the inhibition signal, $I(t)$ has no discontinuity after stimulus offset, and will inhibit the neural activity of the following pulse if the pulses are close together. Therefore, the ODE for this neural model is integrated within

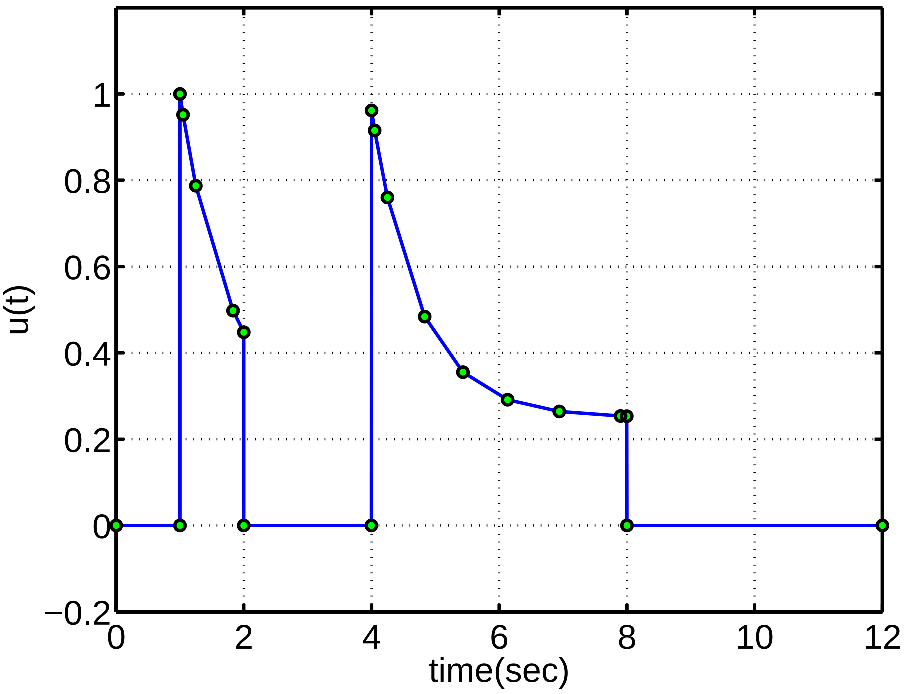


Figure 4.3: $u(t)$ solution from the variable step-size Runge-Kutta solver. The circles mark the solution points. The resulting hemodynamic states solution is seen in figure 4.2. Note the higher density of points as a function of higher curvature; also note that the second neural pulse starts slightly lower than $u(t = 4.0) = 1.0$ due to inhibition from the previous pulse.

each pulse, and $u(t)$ is set to 0 outside the pulses. The ODE for the inhibition signal $I(t)$ is continued between the pulses.

4.3 Interpolation

There are two occasions when simple, linear interpolation may be used to advantage in using the ODE solutions for the models. One concerns the non-linear neural activity model (equation 3.9 in section 3.5.1). The ODE for this model does not depend on the other physiological variables, so it may be solved separately. This means that the variable step size algorithm is free to find optimal step sizes for this integration, and these steps need not coincide with those used in the integration of the other physiological variables. But $u(t)$ enters into the ODE's for the latter, and therefore a simple linear interpolator is used to estimate $u(t)$ at arbitrary times, i.e.

$$\hat{u(t)} = \frac{u(t_2) - u(t_1)}{t_2 - t_1}(t - t_1) + u(t_1);$$

where t_1 and t_2 are the closest solution times to time t. With the chosen integration accuracy, these time points are generally close enough that linear interpolation adds very little error.

The second is in the solution of the ODE's for the physiological variables. The solution time points resulting from the use of the variable step size algorithm do not generally coincide with the sampling times for the BOLD signal, so here (multivariate) linear interpolation is also used, allowing the ODE solver to disregard the BOLD sampling times when choosing step sizes.

4.4 Simulation and synthetic data

Simulation means the generation of data from a model and a set of parameters. It is needed in order to produce synthetic data sets, which are valuable tools for the verification and analysis of methods and models.

The stimulus signal is the same as that used to generate data set 2 (see chapter 2). To justify the assumption that the BOLD signal is independent between epochs, the stimulus for each epoch is set to zero for at least 30 seconds; this also holds for the synthetic data set for the stochastic balloon model, and for data set 2. For data set 1, there is a period of 10 seconds of rest before and after each stimulus. This structure is helpful for preprocessing (e.g. removing low-frequency noise), in that such artifacts can be more accurately estimated using these 'resting' portions of data. Also, it allows us to assume a known, resting, physiological state ($\mathbf{x}_0$) at the start of each epoch.

For the deterministic models, data generation is very simple. The initial state is chosen, $\mathbf{x}_0$, and then the ODE's are solved giving $\mathbf{x}(t)$, $t \in [0; t_N]$. Finally, the measurements are obtained as $y(n) = g(\mathbf{x}(t_n), \theta)$, where the values of the hidden states are interpolated using the ODE solution points.

Figures 4.4 and 4.5 show the first two epochs for the standard and augmented balloon models. A distinct difference between the dynamics of the standard and the augmented balloon model is revealed in these figures, namely that with the augmented ballon model, the CBV (blood volume, $v(t)$) is forced to change more slowly than the CBF (the inflow, $f(t)$), resulting in a faster clearing of deoxyhemoglobin ($q(t)$) and an initial overshoot in the BOLD signal, relative to the standard balloon model. Also, the post-stimulus undershoot is seen to be much stronger in the augmented balloon model.

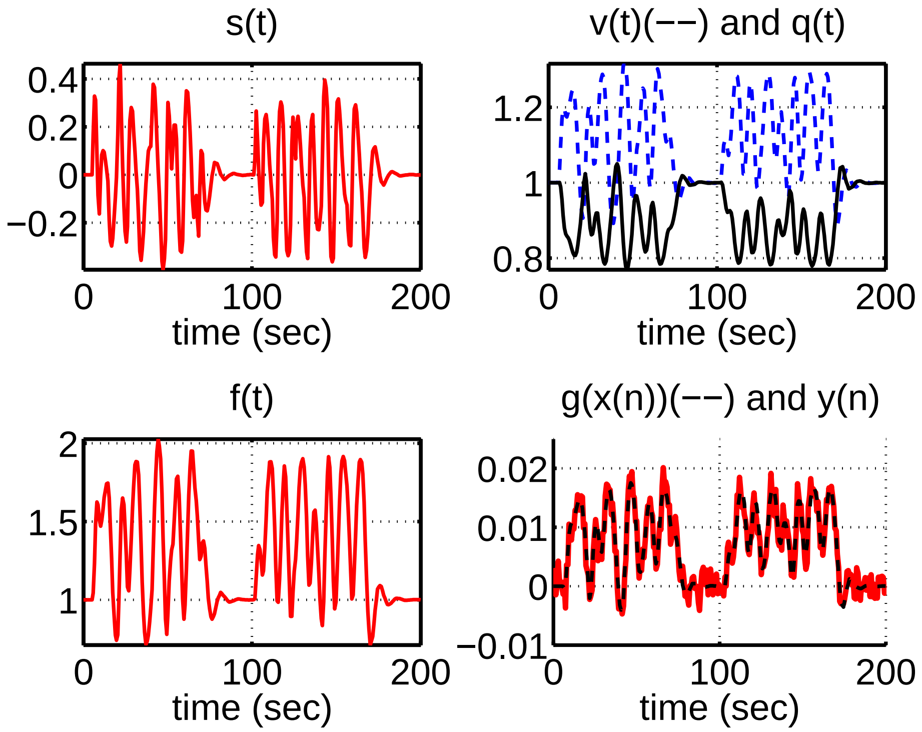


Figure 4.4: Synthetic data generated by the standard balloon model, $\theta = \left[\alpha\ \epsilon\ \tau_0\ \tau_s\ \tau_f\ E_0\ \sigma_w^2\right] = \left[0.4\ 0.5\ 2.0\ 2.5\ 2.5\ 0.4\ 1 \cdot 10^{-5}\right]$. Only the first two epochs are shown.

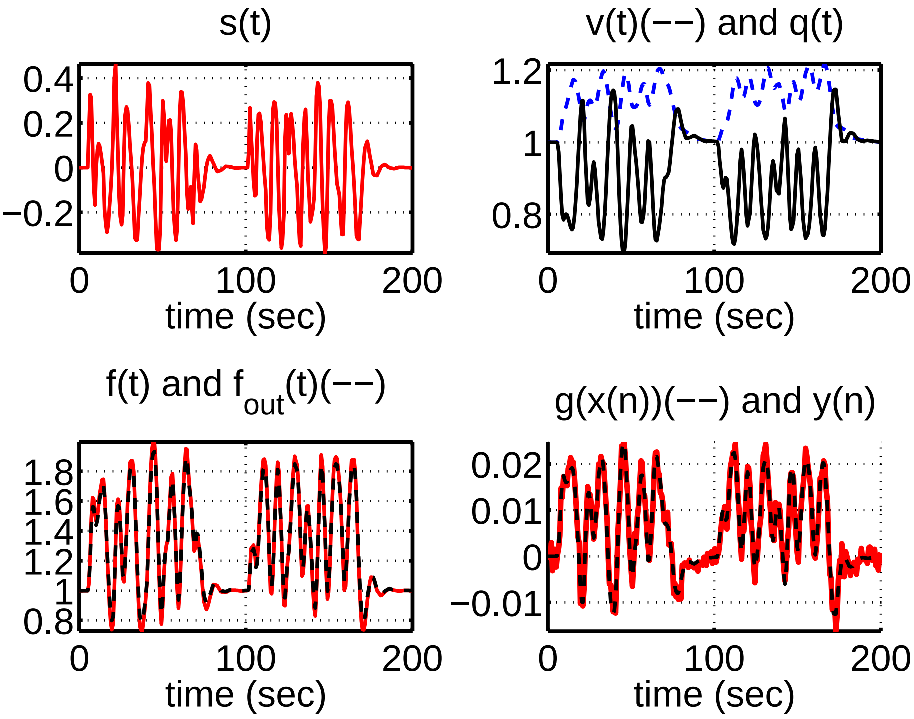


Figure 4.5: Synthetic data generated by the augmented balloon model, $\theta = \left[\alpha\ \epsilon\ \tau_0\ \tau_s\ \tau_f\ E_0\ \tau_+\ \tau_-\ \sigma_w^2\right] = \left[0.4\ 0.5\ 2.0\ 2.5\ 2.5\ 0.4\ 15.0\ 15.0\ 1 \cdot 10^{-5}\right]$. Only the first two epochs are shown.

CHAPTER 5

Stochastic state space models

"The amount of noise which anyone can bear undisturbed stands in inverse proportion to his mental capacity." - Arthur Schopenhauer

5.1 Non-linear, continuous-discrete state space models

Until now, the physiological state variables $v(t)$, $q(t)$, $f(t)$, $f_{out}(t)$ and $s(t)$ have been considered to be deterministic. This is based on the very strong assumption that the hidden states will follow deterministic trajectories from the initial condition for a whole epoch, with only the local neural activity influencing their course. It can easily be imagined that disturbances of various kinds are able to perturb the state variables from this deterministic course. Further, if the true neural activity is not exactly as assumed, namely equal to the known stimulus (or a deterministic function thereof), then the state variables will not be accurately estimated in this manner. And finally, no hemodynamic model will be a complete or perfect model of reality.

One way to relax this deterministic assumption is to consider the state variables to be stochastic rather than deterministic variables. Instead of a set of ordinary differential equations, a set of stochastic differential equations (SDE) is obtained that may be written as

$$d\mathbf{x}(t) = f(\mathbf{x}(t), u(t), \theta)dt + \mathbf{A}d\mathbf{w} \tag{5.1}$$

Here, $\mathbf{x}(t)$ is the r-dimensional hidden state vector, and $\mathbf{w}$ is an r-dimensional Wiener process, which is a stochastic process where the variance of the increments, $w(t) - w(s), t > s$, equals $t - s$. It introduces randomness into the system and makes $\mathbf{x}(t)$ a stochastic vector variable.

$f(\cdot)$ is parameterized by the parameter vector θ and also depends on $u(t)$, the neural activity function. $f(\cdot)$ is often referred to as the drift coefficient, since it causes a deterministic change in $\mathbf{x}(t)$ while the $\mathbf{A}$ matrix - of dimension $[r \times r]$ - is called the diffusion coefficient, since it controls the level of random perturbation coming from the Wiener process, in itself causing $\mathbf{x}(t)$ to 'diffuse' over time from any starting state.

In the following, $f(\cdot)$ is assumed to be time-invariant, since it is assumed to represent time-invariant physiological properties of local neural tissue. $\mathbf{A}$ is assumed to be a constant diagonal matrix, so that the parameters of θ that apply to $\mathbf{A}$ are simply its diagonal elements. The rationale behind this assumption is that the sources of randomness for the different variables are physiologically distinct entities. To give a hypothetical example, blood volume might be perturbed by movement of red blood cells, while the stimulus signal might be perturbed by irregularities in the supply of some chemical involved in the signaling pathway between neural activity and local blood supply; these two are not related. It is also seems reasonable to assume that the noise variance does not depend on the current state, although an investigation into this question is warranted as a future research goal. It is probably going to far to assume that $\mathbf{A}$ has the form

$$\mathbf{A} = \sigma_A^2 \mathbf{I}$$

as there is no prior knowledge supporting the idea that the noise variances for the different physiological variables should be identical. Therefore, $\mathbf{A}$ contains a different parameter for the variance of each state space dimension,

$$\mathbf{A} = \begin{pmatrix} \sigma_1^2 & \dots & \dots & \dots \\ \vdots & \sigma_2^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \dots \\ \vdots & \vdots & \vdots & \sigma_N^2 \end{pmatrix} \tag{5.2}$$

From a modelling point of view, this is also a very attractive assumption compared to a full matrix, meaning the difference of an additional $(D^2 - D)/2$ parameters, where D is the state dimensionality (e.g. 6 extra parameters for the 4-dimensional case).

Further, the choice is made to assume that the noise variance of the hidden states is stationary in time, as is the observation noise.

Note that it is not possible to divide by dt on both sides, as the Wiener process is nowhere differentiable (see for example [45]).

The one-dimensional BOLD measurements are made at discrete times,

$$\{t_0, t_1, \dots, t_N\},$$

and modelled as functions of the hidden state, exactly as before, repeated here for convenience,

$$y_i = g(\mathbf{x}(t_i), \theta) + \epsilon \tag{5.3}$$

$g(\cdot)$ is also assumed to be time-invariant and parameterized by θ. The measurement noise variables ϵ are assumed to be identically and independently distributed as the Gaussian $\mathcal{N}(0, \sigma_w^2)$.

The BOLD measurements up to and including time t_i are termed

$$Y^i = \{y_0, y_1, \dots, y_{t_i}\}.$$

For clarity, the various parameters are all included in θ, although only some of the parameters are used in $g(\cdot)$.

This form of a system of SDE's is not the most general one, but is based on the form of the relevant hemodynamic models and the above mentioned, informed assumptions. A graphical model diagram is shown in figure 4.1 (previous chapter).

The theory and applications of SDE's is a major research area, and several books have been written on the subject (see for example [45] and [47]). The present application is based on one particular approach [73], considering only the first two moments (mean and covariance) of $\mathbf{x}(t)$, leading to approximate numerical solutions.

This model is a generalization of the very well-known Kalman filter ([43],[80]), a major advance in signal processing. For discrete-time, non-linear versions, the most widely used approach for learning is some form of expectation-maximization (see e.g. [67], [4], [71], [28]). But such maximum-likelihood approaches are not useful here, as the task is to estimate the posterior distributions. However, the MCMC approach is equally applicable to stochastic state space models as deterministic ones; the main changes are in computational time and the structure of the likelihood function.

5.2 Likelihood structure

The likelihood function corresponding to (5.1) and (5.3) is no longer as simple as in (4.6), since the hidden states can not be deterministically calculated. The likelihood may instead be factorized as

$$\begin{aligned}\mathcal{L}(\theta) \triangleq p(Y^i|\theta) = p(y_0)p(y_1|y_0)p(y_2|y_1, y_0)\ldots \times p(y_N|Y^{N-1})\\ \triangleq \mathcal{L}_0(\theta)\prod_{i=1}^{N}\mathcal{L}_i\end{aligned} \tag{5.4}$$

where the $\mathcal{L}_i(\theta) \triangleq p(y_i|Y^{i-1})$ and $\mathcal{L}_0(\theta) \triangleq p(y_0)$. This factorization is of course valid for any multivariate stochastic variable.

In order to calculate the $\mathcal{L}_i(\theta)$ terms, it is necessary to obtain approximately the *predicted* hidden state distributions $p(\mathbf{x}(t_{i+1})|Y^i)$, as the mean and covariance of $p(y_i|Y^{i-1})$ are functions of these distributions. From the point of view of interest in the hidden states themselves, the *corrected* distributions (also called 'a posteriori' distributions), $p(\mathbf{x}(t_{i+1})|Y^{i+1})$ are also of interest, since they contain all information on the hidden state distribution given all observations up to time t_{i+1} (in the present application, the hidden states are treated as 'nuisance variables') [1]

There are various alternative approaches to obtaining these distributions. The classical method is the so-called extended Kalman filter (see e.g. [25]), but this has basic problems with convergence and accuracy. This is related to 'local linearization' (see e.g. [72]). A better approach is the continuous-discrete unscented Kalman filter, which is both more accurate and more stable (see [73] for a discussion of the various approaches in the continuous time case).

5.3 The continuous-discrete unscented Kalman filter

For linear stochastic differential systems, it is possible to calculate the corrected distributions $p(\mathbf{x}(t_{i+1})|Y^{i+1})$ exactly, using the so-called Fokker-Planck operator for the system. But for the present case, the continuous-discrete unscented Kalman filter offers an approximate solution. This is based on approximate solutions for the first and second moments of the a posteriori distributions. This is very similar to the discrete-time case unscented Kalman filter (see [40], [41]).

The Continuous-discrete Unscented Kalman filter algorithm is given by an initialization step, followed by iterations of predictive and corrective steps, from time t_0 to t_N. For derivation and details, see [73].

Initialization:

$$\begin{aligned} \mu(t_0|t_0) &= \mu_0 + C(x_0, g_0)/(D(g_0) + \sigma_w^2)(y(0) - E[g_0]) \\ \Sigma(t_0|t_0) &= \Sigma_0 - C(x_0, g_0)/(D(g_0) + \sigma_w^2))D(g_0, x_0) \end{aligned} \tag{5.5}$$

Prediction:

[1] 'Smoothing' estimates $p(\mathbf{x}(t_{i+1})|Y^N)$, i.e. using all observations. However, this is difficult and time-consuming to do for non-linear continuous systems and is not needed for the evaluation of the likelihood, which is the primary target in the present case.

$$\frac{d\mu(\tau|t_i)}{dt} = E\left[f(\mathbf{x}(\tau), u(\tau))|Y^i\right] \tag{5.6}$$

$$\begin{aligned}\frac{d\Sigma(\tau|t_i)}{dt} &= C\left[f(\mathbf{x}(\tau), u(\tau)), \mathbf{x}(\tau)|Y^i\right] \\ &+ C\left[\mathbf{x}(\tau), f(\mathbf{x}(\tau), u(\tau))|Y^i\right] \\ &+ \mathbf{A}\mathbf{A}^T\end{aligned} \tag{5.7}$$

Correction:

$$\begin{aligned}\mu(t_{i+1}|t_{i+1}) &= \mu(t_{i+1}|t_i) \\ &+ C(\mathbf{x}(t_{i+1}), g_{i+1}|Y^i)/(D(g_{i+1}|Y^i) + \sigma_w^2)\times \\ &(y_{i+1} - E[g_{i+1}|Y^i])\end{aligned} \tag{5.8}$$

$$\begin{aligned}\Sigma(t_{i+1}|t_{i+1}) = \Sigma(t_{i+1}|t_i) - C\left[\mathbf{x}(t_{i+1}), g_{i+1}|Y^i\right]/(D(g_{i+1}|Y^i) + \mathbf{A}) \\ \times C(g_{i+1}, \mathbf{x}(t_{i+1})|Y^i)\end{aligned} \tag{5.9}$$

The $(i + 1)$'th term of the likelihood is also calculated as

$$\mathcal{L}_{i+1}(\theta) = \mathcal{N}(y(i + 1), E\left[g_{i+1}|Y^i\right], D(g_{i+1}|Y^i) + \sigma_w^2) \tag{5.10}$$

and this is practically done alongside the filtering. The expectations conditioned on Y^i should be understood as being expectations with respect to the prior distribution of the hidden states, $p(\mathbf{x}(i+1)|Y^i)$ as approximated with the predicted moments, $\mu(t_{i+1}|t_i)$ and $\Sigma(t_{i+1}|t_i)$.

5.3.1 Notes

The initialization of the algorithm can be seen as a correction of an initial guess, given by μ_0 and Σ_0. $E[g_0]$ is calculated using the sigma-points corresponding to μ_0 and Σ_0.

The prediction step is given in terms of a system of ordinary differential equations for the a priori moments of $\mathbf{x}(\tau)$. The solution of this system is described below.

Comparing this likelihood with the likelihood for the deterministic model (4.6), there is an additional variance term, namely $D(g_{i+1}|Y^i)$. An intuitive way of comparing the two is that while the variance in the deterministic case would seem to be smaller, due to this additional term, the predicted measurement, $\hat{y}(i)$, is going to be further from the actual measurement $y(i)$, since there is no internal noise to aid in the prediction. Therefore, the σ_w^2 estimate can be expected to be higher for the deterministic model.

The correction step depends on the so-called 'theorem on normal correlation'. This is not elucidated in [73], so a brief explanation is given here.

Let the stochastic vector $(\mathbf{x}, \mathbf{y})$ be normally distributed with $E[(\mathbf{x}, \mathbf{y})] = (\mu_{\mathbf{x}}, \mu_{\mathbf{y}})$ and

$$D[(\mathbf{x}, \mathbf{y})] = \begin{bmatrix} D\mathbf{xx} & D\mathbf{xy} \\ D\mathbf{yx} & D\mathbf{yy} \end{bmatrix}$$

Then, according to the theorem on normal correlation, the conditional expectation is given by

$$E[\mathbf{x}|\mathbf{y}] = \mu_{\mathbf{x}} + D_{\mathbf{xy}} D_{\mathbf{yy}}^{-} (\mathbf{y} - \mu_{\mathbf{y}}) \tag{5.11}$$

and

$$D(\mathbf{x}|\mathbf{y}) = D_{\mathbf{xx}} - D_{\mathbf{xy}} D_{\mathbf{yy}}^{-} D_{\mathbf{yx}} \tag{5.12}$$

Letting $\mathbf{x}$ correspond to the hidden state at time t_{i+1}, and $\mathbf{y}$ to the new observation, $\mathbf{y}_{i+1}$, both also conditioned on the observations up to and including the last observation, Y^i, the combined mean and variance can be written as

$$E[(\mathbf{x}, \mathbf{y})] = (E[\mathbf{x}|Y^i], E[\mathbf{y}|Y^i])$$

and

$$D[(\mathbf{x}, \mathbf{y})] = \begin{bmatrix} D[\mathbf{x}(t_{i+1})|Y^i] & D[\mathbf{x}, \mathbf{y}|Y^i] \\ D[\mathbf{y}, \mathbf{x}|Y^i] & D[\mathbf{y}_{i+1}|Y^i] \end{bmatrix}$$

Inserting these expressions into (5.11) and (5.12) gives the needed result (see (5.8) and (5.9)).

The proof is somewhat technical and is given in [50], pages 56-57. However, the intuition regarding the employment of this theorem in the context of state-space filtering is straight forward: Take the predicted mean and correct it by some factor times the error between the predicted output and the measured. The "some factor" is the covariance between the state and measurement at time t_n times the (pseudo) inverse of the variance of the measurements. In other words, the higher the covariance between state and measurement, and the lower the uncertainty about the measurements (observation noise), the more the prediction should be corrected towards the actual observation. Similarly, the uncertainty of the state is reduced with a new observation, proportionally to the certainty about the observations, and again depending on the covariance between the state variables and the observation variables. The Kalman filter is therefore said to have a "predictor-corrector" structure, as the state is first predicted ahead in time, then corrected using the next observation, and so on until the entire observation time series has been processed.

A nice corollary is that

$$E[\mathbf{x}|\mathbf{y} = \mu_{\mathbf{y}}] = E[\mathbf{x}] \tag{5.13}$$

which can be seen directly from (5.11). Using this, it is also easy to see that in general for Gaussian variables,

$$E[(x - \mu_x)(y - \mu_y)] = E[x(y - \mu_y)] = E[y(x - \mu_x)] \tag{5.14}$$

This is used in the implementation of the filter, saving some computational time for the deduction of one of the means in calculations of covariances.

The correction step is also based on the optimal linear estimate ([73], see Lemma 14.1 in [50]) which gives the mean and variance of the posterior density as a linear function of the mean and variance of the prior. This also shows in the form of the correction formulas, which are identical in form to those obtained in the linear, discrete time Kalman filter case (see for instance [80], [27]), even though the non-linearites have not been linearized (as they would be in the EKF).

The key advantage of the unscented Kalman filter lies in the choice to approximate the probability density functions, which turns out to be give more accurate results than by approximating the non-linearities through linearization (EKF). This can be achieved with no additional computational cost.

5.4 The unscented transformation

The prediction ((5.6), (5.7)) and correction ((5.8), (5.9)) steps of the filtering algorithm - and most importantly the likelihood ((5.10)) require the calculation of expectations,

$$E[h(x)] = \int h(x)p(x)dx$$

where $h(\cdot)$ is some function. With the unscented transformation method [41], such expectations are estimated using empirical distributions on so-called 'sigma points' s_i, approximating the distribution as

$$p(x) \approx \sum_{i=-r}^{r} \omega_i \delta(x - s_i)dx \tag{5.15}$$

where $r = \dim(x)$.

These points are chosen so that the first two moments (μ and Σ) of the distribution are correct. With the non-linear differential equations of the hemodynamic models, the predicted distribution $p(\mathbf{x}(t_{i+1})|Y^i)$ will not be Gaussian, so (5.15) is going to be an approximation.

With this approximation, expectations and (co-) variances can be calculated as

$$E[h(x)] \triangleq \int h(x)p(x)dx \approx \sum_{i=-r}^{r} h(s_i)\omega_i \tag{5.16}$$

and

$$\begin{aligned} D(h(x)) &\triangleq \int (h(x) - E[h(x)])(h(x) - E[h(x)])^T p(x)dx \\ &\approx \sum_{i=-r}^{r} (h(s_i) - E[h(x)])(h(s_i) - E[h(x)])^T \omega_i \end{aligned} \tag{5.17}$$

Using a Cholesky factorization (lowering computational cost, see [78]), $\Sigma = \Gamma\Gamma^T$, and letting γ_i be the i'th column of Γ, one can use the $2r+1$ (r equalling the dimensionality of x) sigma points

$$\begin{aligned} s_0 &= \mu \\ s_i &= \mu + \sqrt{r+\lambda}\gamma_i,\ i = 1..r \\ s_{-i} &= \mu - \sqrt{r+\lambda}\gamma_i,\ i = 1..r \end{aligned}$$

with corresponding weights

$$\begin{aligned} \omega_0 &= \lambda/(r+\lambda) \\ \omega_i = \omega_{-i} &= 1/(2(r+\lambda)) \end{aligned}$$

λ is a scaling factor, and a good setting depends on the specific problem. For the hemodynamic systems investigated here, $\lambda = 3$ was found to work well so this value was used throughout.

Figure 5.1 shows a demonstration of the sigma points corresponding to a two-dimensional Gaussian distribution.

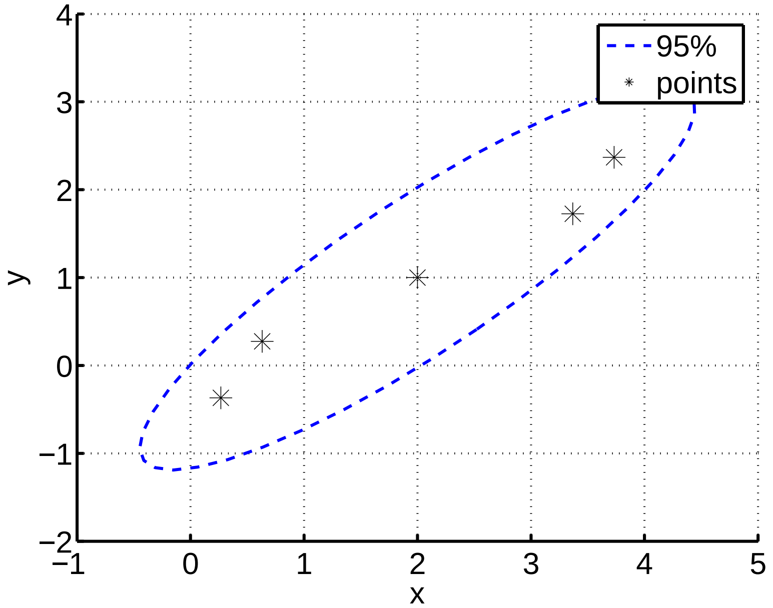


Figure 5.1: Demonstration of the location of the sigma points for a 2D Gaussian. Also shown is the 95% confidence interval ellipse for this distribution.

5.5 ODE solution

One very appealing aspect of the unscented Kalman filter is the opportunity to reuse whatever ODE solver has been implemented for the deterministic model counterpart. In the present case, the variable step size Runge-Kutta solver described in the previous chapter was reused.

The dimensionality of course goes from r to $2r+1$, as each sigma point is one dimension in the ODE system. Another speed bump is that the solver must provide a solution (prediction) at each observation time point, so that the correction step can be done before proceeding (the corrected state distribution must be known before the ODE solution is continued). By using the last time step size from the previous prediction solution, the ODE solver is still able to consistently use only three to five steps to predict a one-second interval with the models used here.

There are $2r+1$ sigma points (r being the dimensionality of the hidden state space). The differential equations for the first r of these are simply given by (5.6). The rest of the sigma points depend on the derivative of the hidden state dispersion matrix, $\Sigma(\tau|t_i)dt$. The derivative of the elements of the sigma points, x_i, can thus be found by using a Cholesky factorization of the derivative of $\Sigma(\tau|t_i))dt$,

$$\frac{d\Sigma(\tau|t_i))}{dt} = \dot{\Gamma}\dot{\Gamma}^T$$

and the derivatives of the sigma points are then given as

$$\begin{aligned}
\frac{ds_0}{dt} &= \frac{d\mu(\tau|t_i)}{dt} \\
\frac{ds_i}{dt} &= \frac{d\mu(\tau|t_i)}{dt} + \sqrt{r+\lambda}\dot{\gamma}_i, \;\; i = 1..r \\
\\
\frac{ds_{-i}}{dt} &= \frac{d\mu(\tau|t_i)}{dt} + \sqrt{r+\lambda}\dot{\gamma}_i, \;\; i = 1..r
\end{aligned}$$

where $\dot{\gamma}_i$ is the i'th column of $\dot{\Gamma}$.

5.6 Computational cost

Compared to the deterministic case, the estimation of the gradient (in the prediction step, equations (5.6) and (5.7)) now involves expectations using the unscented transformation, and the correction step is an additional computational load. An evaluation of the stochastic likelihood takes roughly 50 times longer than the deterministic version.

5.7 Simulation and synthetic data

Simulating from a system of stochastic differential equations is a subject of current research and of which whole books have been written (see e.g. [45]). The simplest simulation method is Euler's method, which works by starting at some state, $\mathbf{x}_0$, and simulating according to

$$\mathbf{x}(t+\delta t) = \mathbf{x}(t) + f(\mathbf{x}(t), u(t), \theta)\delta t + \mathbf{A}d\mathbf{w}$$

where δt is a very small time step, and $d\mathbf{w}$ is a random increment of a Wiener process with variance equal to δt, i.e. $d\mathbf{w} \sim \mathcal{N}(0, \delta t)$. If δt is chosen small enough, this approximation will be accurate. How small depends on the characteristics of $f(\cdot)$ and the diagonal values of $\mathbf{A}$. δt can be chosen by simply inspecting the results of varying it, but since data generation is not a time-critical task, this is usually not a problem. For the synthetic data generated in this project, it was set to $\delta t = 1e - 3$.

A synthetic data set was generated using the standard balloon model with the same hemodynamic parameters used for the creation of the synthetic data for the deterministic state space model, $\theta = \left[0.4\ 0.5\ 2.0\ 2.5\ 2.5\ 0.4\ 1 \cdot 10^{-5}\right]$, and the same stimulus function. The noise variances of all hidden variables was set to $1 \cdot 10^{-3}$. Figure 5.2 shows the first two epochs of the generated data.

A synthetic data set was also generated for the stochastic version of the augmented balloon model (not shown).

Figure 5.7 illustrates the local shape of the log likelihood function $\mathcal{L}(\theta)$ for this data by varying two of the parameters, τ_0 and τ_s while the other parameters are kept at their true values. Such inspections give a feel for the properties of the likelihood function, but of course do not rule out the existence of more than one modes with similar likelihood values.

5.7.1 Prior distribution for state-space noise variances

Setting the state-space noise variances to zero leads to the deterministic variant of the given model, and thus is the simplest version of the noisy model. Therefore, in order for the priors to reflect a belief leaning towards simplicity, the noise variance priors should assign decreasing probability density to increasing noise variance. As the elements of $\mathbf{A}$ reach around 0.1, the system of SDE's begins to fluctuate wildly, eventually drowning out the influence of the neural input. Hence it is reasonable to design a prior that assigns very small probabilities are to values higher than 0.1.

With these considerations in mind, the gamma-distribution is a suitable choice,

$$p(\theta|s,c) = \frac{1}{\Gamma(c)s}\left(\frac{\theta}{s}\right)^{c-1} \exp\left(-\frac{\theta}{s}\right) \tag{5.18}$$

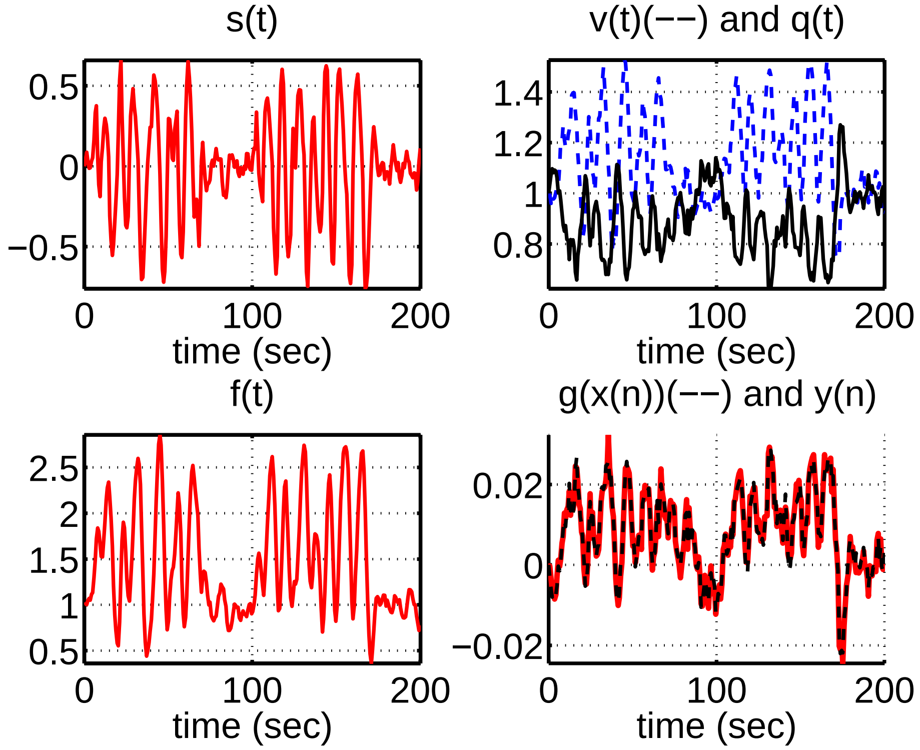


Figure 5.2: Synthetic data generated by a stochastic state space version of the standard balloon model. The first two epochs are shown.

where $\Gamma(\cdot)$ is the gamma function. This is a simple uni-modal distribution, with a scale parameter s and s shape parameter, c. These parameters were set to $s = 0.1$ and $c = 1.01$, see figure 5.4. The priors for all the hidden noise variances are set to the same distribution, as there is no prior belief that they should be different.

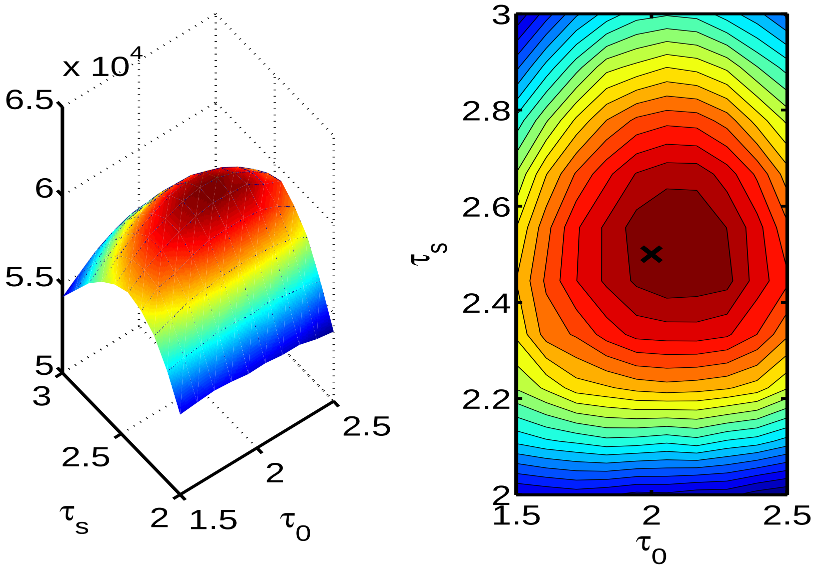


Figure 5.3: 3D surface and contour plots of the log likelihood surface of a stochastic version of the standard balloon model evaluated while varying two parameters around the neighborhood of the true known values (synthetically generated data). There is a clear peak around the true value of the parameter pair (marked with a cross on the right) for this set of parameters, the discrepancy being caused by noise.

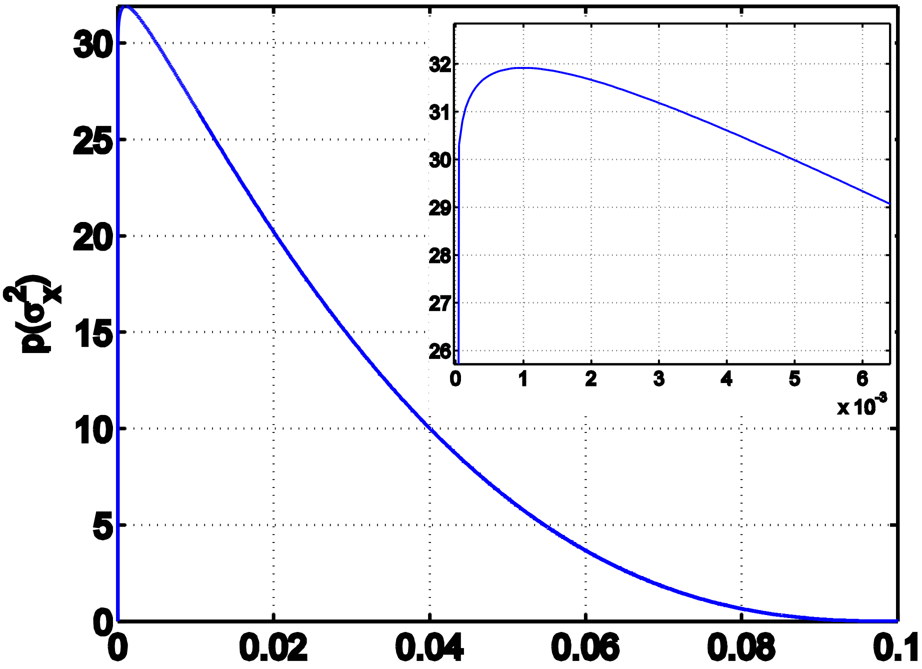


Figure 5.4: Prior distribution for the state space noise variances parameters; the inset shows a zoom of the mode. The same distribution is used for σ_v^2, σ_q^2, σ_f^2, and σ_s^2.

Chapter 6

Markov chain Monte Carlo learning

"The Bayesian 'machine', together with MCMC, is arguably the most powerful mechanism ever created for the processing of data and knowledge." - James O. Berger

'Learning' may be defined as obtaining the distribution of the parameters θ of a model conditioned on a data set D, i.e. the posterior distribution $p(\theta|D)$[1]. Due to the non-linear nature of the problems under consideration, it is not possible to arrive at an analytical form for the posterior for the model parameters, $p(\theta|D)$. It is therefore necessary to use some means of approximating this distribution.

There are many possible approaches to approximate learning, one of which is Markov chain Monte Carlo (MCMC) sampling[2]. The core idea in MCMC is to represent a target distribution by samples generated from it, $\theta_i \sim p(\theta|D)$. These samples may then be used to represent the distribution (e.g. histograms), and to calculate expectations with respect to it, including calculating its moments.

MCMC sampling is computationally costly, and for the stochastic state space models, not practically viable due to the cost of evaluating the likelihood func-

[1] This is a supervised learning definition, and other definitions are possible.

[2] A leading alternative is 'variational Bayes', see http://www.variational-bayes.org

tion (see equation (5.10)). In this case, a related method - 'simulated annealing' - is used to obtain an estimate of the maximum a posteriori parameter vector.

There are many good books and articles on MCMC sampling (see e.g. [58], [55]), and only a brief overview of the general theory is given here, the focus being on details relevant to the present application.

6.1 Estimating expectations

The fundamental idea in MCMC sampling is that expectations with respect to a distribution can be approximated through samples from that distribution,

$$E[h(x)] \triangleq \int h(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} h(x_i) \tag{6.1}$$

where $x_i \sim p(x)$.

The necessary samples can be generated in the form of a *Markov chain*, which in the present setting is a series of continuous-space, random variables $X^N = [x_0, x_1, \ldots, x_N]$ where the conditional probability density function of each variable is only dependent on the previous one,

$$p(x_i|X^N) = p(x_i|x_{i-1})$$

Defining the *transition probability function* $T(x, x')$ as the probability of making a transition from x to x', the PDF of each variable is thus

$$p(x) = \int p(x')T(x', x)dx'$$

where x' is the variable previous to x in the chain X^N.

The task is then to construct a transition probability function such that samples generated by first generating one sample from a chosen initial PDF, $p(x_0)$, and then repeatedly applying $T(x, x')$ will actually be distributed according to a target PDF, $\pi(x)$. For this to hold, $\pi(x)$ must be an *invariant* distribution of the chain, meaning that

$$\pi(x) = \int \pi(x')T(x', x)dx \tag{6.2}$$

A sufficient condition for invariance is that of *detailed balance*, which for continuous spaces is

$$\int_A \int_B \pi(x)T(x, x')dx'dx = \int_B \int_A \pi(x')T(x', x)dxdx' \tag{6.3}$$

i.e. the probability of making a transition from some point in A to some point in B is the same as the other way around. It is easy to see (through integration) that if (6.3) holds, (6.2) will hold.

The second condition is that the chain must be *ergodic*, which basically means that the starting point is inconsequential, so that the sampling may start in any place and still converge to generating samples from the target distribution.

The target distribution here is of course the conditional parameter posterior for a given hemodynamic model, $\pi(\cdot) = p(\theta|D)$. Bayes' rule allows us to rewrite the posterior in terms of the likelihood, $p(D|\theta)$, the prior, $p(\theta)$ and a normalizing factor, $p(D)$:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{6.4}$$

6.2 Metropolis-Hastings

The Metropolis-Hastings algorithm ([81], [55]) is one way of generating a Markov chain with detailed balance. It works by starting at an arbitrary state, θ_0, and then iteratively proposing small changes to generate the next sample in the Markov chain. This is done using a *proposal distribution*, $p(\theta'|\theta)$, where θ is the current state. Here, a Gaussian centered on the previous state has been chosen as the proposal distribution,

$$p(\theta'|\theta) = \mathcal{N}(\theta, \Sigma)$$

The proposal is accepted as the next discrete sample according to the *acceptance ratio*:

$$r = \frac{p(\theta'|D)}{p(\theta|D)} = \frac{p(D|\theta')p(\theta')}{p(D|\theta)p(\theta)} \tag{6.5}$$

The proposal distribution does not figure in this ratio, since it is symmetric, $p(\theta'|\theta) = p(\theta)|\theta')$; also note that the normalizing factor, $p(D)$, cancels out. If $r \geq 1$ the proposal is accepted, and the next sample generated is $\theta_{i+1} = \theta'$. If $r < 1$, the proposal is accepted with probability r. If the proposal is not accepted, the next sample is simply kept at the current value, $\theta_{i+1} = \theta$. In other words, the *acceptance function* is

$$A(\theta, \theta') \triangleq \min(1, p(\theta'|D)/p(\theta|D)) \tag{6.6}$$

The transition probability function of the Metropolis-Hastings algorithm may thus be expressed as a product of the proposal PDF and the acceptance function, $T(\theta, \theta') = p(\theta'|\theta)A(\theta, \theta')$.

It is easy to see that detailed balance holds for this transition probability function. Letting the integrand in (6.3) represent $\pi(x) = p(\theta|D)$,

$$\begin{aligned}\pi(x)T(x, x') &= \pi(x)p(x'|x)A(x, x')\\ &= p(x'|x)\pi(x)\min(1, \pi(x')/\pi(x))\\ &= p(x'|x)\min(\pi(x), \pi(x'))\\ &= p(x|x')\pi(x')\min(1, \pi(x)/\pi(x'))\\ &= \pi(x')T(x', x)\end{aligned} \tag{6.7}$$

since the proposal is symmetric.

In practice, of course, it suffices to compare a randomly generated variable that is uniformly distributed from 0 to 1 with r to determine wether or not the proposal is accepted, and no minimum need be taken.

The *acceptance rate* ρ is defined as the percentage of proposes samples that are accepted. Generally, for higher dimensional problems, it is necessary to use higher acceptance rates ([58]), but for the current problems of dimension around 10, an acceptance rate between $P_{min} = 0.2$ and $P_{max} = 0.5$ was found to give good 'mixing' (convergence).

6.2.1 Automated proposal generation

The shape and scale of the proposal distribution Σ is critical to the convergence of the algorithm, i.e. how many samples need be generated before the approximation (6.1) is sufficiently accurate. It is therefore normal to choose Σ by trial and error, manually. However, with the application at hand, where a lot of approximations must be done with different models and different data sets, this is not satisfactory. Therefore, an automated procedure for finding a good proposal was implemented.

The sampling is started with an arbitrary normal distribution of dimension $\dim(\theta)$. Then, several short 'scout' sampling runs are then performed, each with N samples. After each of these short runs, the covariance of the generated samples θ^N is used as the new proposal covariance,

$$\Sigma' = \frac{1}{N}\sum_{n=1}^{N}(\theta_n - \mu_\theta)(\theta_n - \mu_\theta)^T \tag{6.8}$$

where μ_θ is the empirical mean,

$$\mu_\theta = \sum_{n=1}^{N}\theta_n$$

After each update, it is necessary to scale this new proposal, i.e. to find a scaling σ so that $\Sigma = \sigma\Sigma'$ achieves the desired acceptance rate. This is a search problem, which has been solved by a simple bisection method, see figure 6.1. First, an upper bound σ_0 is found by doubling σ until the acceptance rate is less than P_{min}. Bisection is then carried out until the estimated acceptance rate is in the desired range.

After a set number (usually around 10) of these initial iterations, the main sampling run is performed keeping the proposal distribution constant (the samples of the initial runs are not used further).

It was found that 100 samples in each of these short runs was sufficient to find good proposals. As small a number as possible is of course desirable for speed, but too small numbers give too high of a variance in the covariance estimates (6.8) and may also lead to divergence during the bisection algorithm, as the variance of the estimated ρ is also increased with small sample size. Usually from 1 to 5 iterations are needed for each scaling.

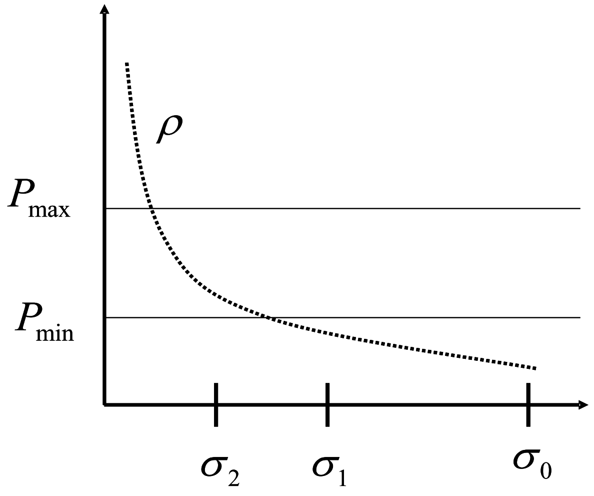


Figure 6.1: Bisection is used to find an appropriate scaling for the proposal distribution, so that the acceptance rate ρ lies between P_{min} and P_{max}.

6.2.2 Finding a good starting point

Another condition for reaching convergence as quickly as possible is to start the sampling from a 'good' starting point. 'Good' here means a point with a high posterior probability $p(\theta|D)$. Starting far from regions of high posterior probability means that it might take a long time to move to high probability region, and also the proposals found in such regions might not be optimal for sampling around the major modes, which is what the algorithm should be doing.

Several alternative policies exist for finding a good starting point. It is important that such a point can be found quickly, as otherwise the point of speeding up convergence is defeated. The method used here might be called an 'iterative univariate search', and is very quick and simple. Starting from parameter values that are expected a priori to be likely, each parameter in turn is iterated through to equally distributed values across a range covering the bulk of the corresponding prior distribution. It is then set to the value giving the highest likelihood. This is then repeated for the next parameter, keeping the optimal value for the previous parameter, until all parameters have been 'optimized' in this way. The whole procedure is then repeated, starting with the optimal values from the previous iteration. It was found that 2 iterations of this procedure with 4 points in each range was sufficient for finding a good starting guess across different data sets and models. This simple procedure could be improved in countless ways (e.g. refining the search range after each iteration), but serves its purpose very well in practice.

As an example of the benefit of this method, the log likelihood of the stochastic version of the augmented balloon model in typical runs was around 600 with the initial parameter setting, around 1500 after the starting point procedure, and around 2600 after 3000 simulated annealing iterations (data set 2).

6.3 Parallel tempering

Depending on the properties of the posterior, 'mixing', i.e. the ability of the algorithm to generate samples representative of the whole distribution, may be slow. The technique of parallel tempering (see e.g. [32]) may be employed as a quite straightforward enhancement of Metropolis-Hastings sampling. It works by sampling from several distributions $p_i(\theta|D)$ in parallel, each using a more or less 'flattened' version of the likelihood:

$$p_i(\theta|D) \triangleq \frac{p(D|\theta)^{\beta_i} p(\theta)}{\int p(D|\theta)^{\beta_i} p(\theta) d\theta}, \quad i = 1 \dots C \tag{6.9}$$

where $\beta_i \leqq 1$ are so-called *inverse temperatures*, $\beta_i = \frac{1}{T_i}$, and C is the number of 'chains'. At certain intervals (20 samples was used here), a proposal is made to swap the states of two adjacent chains, using an acceptance ratio somewhat similar to (6.5),

$$r_{PT} = \frac{p_i(D|\theta_{i+1})p_{i+1}(D|\theta_i)}{p_i(D|\theta_i)p_{i+1}(D|\theta_{i+1})} \tag{6.10}$$

where θ_i is the current state of the i'th chain (the priors cancel out). This proposal is made instead of a 'normal' Metropolis-Hastings proposal.

The chains must be spaced across temperatures at suitable intervals. To determine this dispersion it was found most logical to start with the first chain after the basic $T = 1$ chain, and set the temperature so high as to give a swap acceptance rate of around 0.2 to 0.5. The same can then be done to find a suitable temperature for the second-highest chain, and so on. The number of chains to use depends on at what point the 'heated' density becomes flatter than the priors, so to speak, because when that happens, the proposals in that chain will be largely rejected due to overstepping the bounds of the fixed-support priors, which is not productive. For the present data, using 6 chains from $\beta_1 = 1.0$ to $\beta_6 = 0.04$ ($T_6 = 25$) was found to give good results. Good parameters were found by experimenting with different settings (temperatures, samples between swap proposals, number of chains) on synthetic data, see figure 6.2. The parameters are interdependent, e.g. with many chains, it is necessary to suggest swapping moves more often.

Parallel tempering generally leads to better mixing in the same amount of computer time, since the high-temperature chains are able to move around quite freely, allowing the target distribution (at temperature $T = 1$) to 'escape' local minima, see figure 6.3.

6.3.1 Toy example: Mixture of Gaussians

In order to validate and get a feel for the MCMC algorithms, they were run on a toy example with a mixture-of-Gaussians likelihood of the form

$$p(x) = \sum_{i=1}^{K} P(i)\phi(x|\theta_i)$$

where each component ϕ_i is a Gaussian:

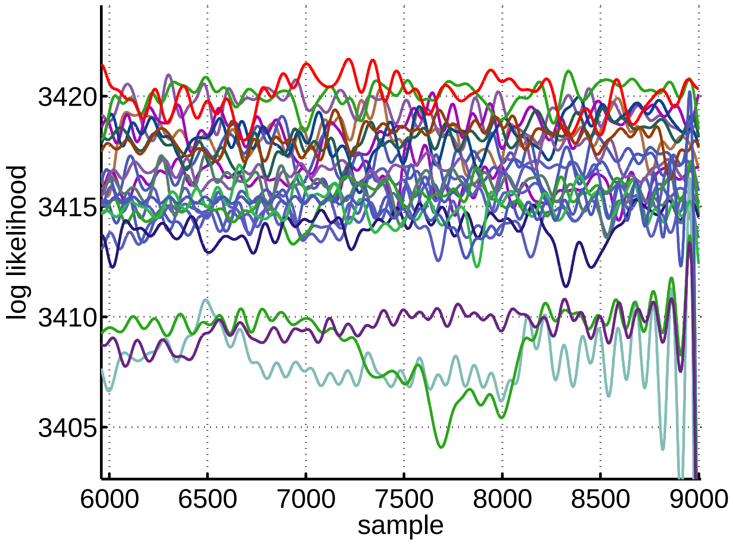


Figure 6.2: Log likelihoods for parallel tempering with different settings; final part of sampling run for synthetic data (generated by the standard balloon model, sub-sampled for clarity).

$$\phi_i(x) = \frac{1}{\sqrt{2\pi}^d} \frac{1}{\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right]$$

with the number of components $K = 2$, and the dimension of x, $d = 2$. By setting the means and variance such that the overlap between the two components is small, it is easy to illustrate the benefit of the parallel tempering algorithm. The running time of the parallel tempering algorithm relative to the standard Metropolis-Hastings algorithm is $\mathcal{O}(N_C)$, where N_C is the number of chains running in parallel. If parallel tempering is to be beneficial, it should therefore cover the target distribution more than N_C times faster (in number of iterations) than the standard MH algorithm. For the toy example, one only needs to remove the two components a certain distance from each other, depending on their covariances of course, to see the PT algorithm outperform the MH algorithm, see figures 6.4 and 6.5.

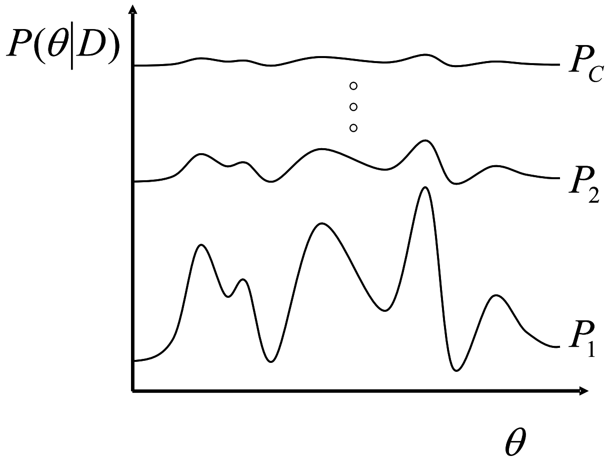


Figure 6.3: Parallel tempering. The high-temperature chains (top) allow the samples to move around, so samples 'trapped' in the lower 'colder' chains can escape.

6.4 Simulated annealing

For the stochastic model, evaluation of the likelihood takes too long (around 50 times longer than for the deterministic model) for the MCMC sampling approach to be practically viable. Instead, a point estimate approximating the maximum

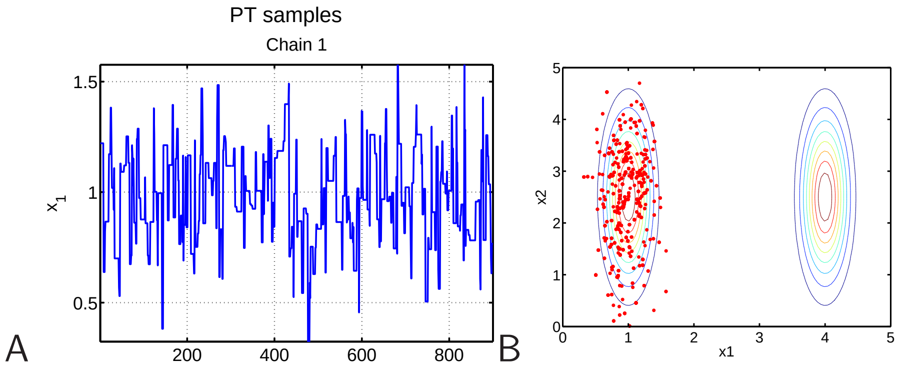


Figure 6.4: MCMC sampling with only one chain and 900 samples. A: The values for x_1 are confined to one mode of the distribution. B: Scatter plot of the samples overlaid on a contour plot of the distribution.

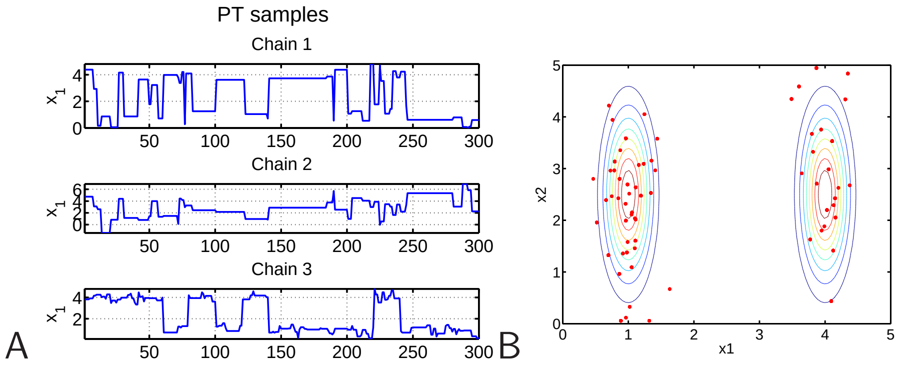


Figure 6.5: MCMC sampling with 3 chains, but only 300 samples. A: The values for x_1 find both modes of the distribution (chain 1 has the highest-temperatures). B: Scatter plot of the samples overlaid on a contour plot of the distribution.

a posteriori (MAP) is used instead,

$$p(\theta|D) \approx \delta(\theta - \theta_{MAP}) \tag{6.11}$$

where θ_{MAP} is defined as

$$\theta_{MAP} \triangleq \arg\max_{\theta} p(\theta|D) \tag{6.12}$$

This sample then represents the posterior distribution and may also be used to calculate approximate expectations.

The simulated annealing technique produces an approximation to θ_{MAP} and is related to both MCMC sampling and parallel tempering. The procedure is in fact identical to the Metropolis-Hastings sampling described above with the modifications that only the final sample is used as the estimate of θ_{MAP}, and the temperature (see (6.9)) is gradually decreased towards zero during sampling. This means that θ will move relatively freely around initially, but as the temperature decreases will only move in the direction of the gradient of $p(\theta|D)$, hopefully resulting in a final value close to θ_{MAP}.

The proposal is initially determined with the automatic procedure described for the Metropolis-Hastings algorithm, but as the temperature decreases, the scale of the proposal must be reduced in order to maintain an appropriate acceptance ratio. As discussed in [58], if one approximates a 'heated' likelihood $p(D|\theta)^{1/T}$ with a Gaussian, the standard deviation of that Gaussian will be $\sqrt{T}$, and thus a step size of around $\sqrt{T}$ will be appropriate. Therefore, the proposal is scaled at each step by $\sqrt{T/T'}$, where T is the temperature at the last step, and T' is the temperature at the current step.

6.4.1 Cooling schedules

Using synthetically generated data, a suitable starting temperature can be found be doing short sampling runs at increasing temperatures, until a temperature is found that allows rapid fluctuation of the parameters, enough so that they are able to converge to the vicinity of the true parameters relatively quickly (e.g. after a few hundred iterations), yet not so high as to continually overstep the prior bounds. The starting temperature should also not be set higher than necessary in this sense, because it will need to be gradually lowered to zero in the subsequent search for the MAP parameters, and this so-called 'cooling' must occur very gradually in order for the search to succeed ([58]) and it is desirable with regard to computational time to reach that point in as few steps as possible. The manner in which the temperature is brought down is referred to as a 'cooling schedule'

A practical approach is thus to experiment with the speed of cooling for increasingly long sampling runs, until a suitable cooling schedule has been found for the final sampling run that ends near the MAP point for synthetic data. With such experiments on synthetic data, it was found that an exponential decay gave good results,

$$T_i = T_0 \exp(-i\tau_T)$$

where T_0 is the starting temperature and T_i is the current temperature. τ_T is a time constant, set to around 1000-5000 depending on the length of the sampling run, to give temperatures close to zeros at the end of the sampling run. For synthetic data, it was found that using a starting temperature of $T \triangleq \frac{1}{\beta} = 10.0$ and slowly decreasing over a sampling run of 3000 samples consistently gives a MAP parameter estimate close to the true value.

6.5 Logarithm transformation

Instead of tracking the likelihood $p(D|\theta)$ and the posterior $p(\theta|D)$, the logarithm of these is used instead. This can be done since the logarithm is a monotonous function, and is also attractive for numerical reasons (the likelihoods are products of many small numbers, see 4.6 and 5.4).

6.6 Convergence analysis in MCMC

There are a great number of suggested methods to help determine whether or not a given Markov chain has converged (see [20] for a review). However, such methods are heuristics, in so far as there is in principle no way to tell whether convergence has occurred. What the various heuristics do is a test, and if the test is not successful, then one knows that convergence has *not* occurred. However, a successful test is no guarantee of convergence. Still, knowing that convergence *may* have occurred is of course better than knowing that is has *not* occurred.

Therefore, a set of heuristics have been employed to obtain indications that the final sampling estimate of the posterior distribution does indeed cover the true distribution.

Measuring the autocorrelations of the obtained samples gives an indication of how many samples are needed to obtain each independent sample. Although the samples used in (6.1) do not need to be independent, they must contain preferably a dozen or so ([55], p. 358) independent samples for the approximation to be good. A typical example of autocorrelations is shown in figure 6.8. This is related to maintaining an appropriate acceptance rate throughout each sampling run, and this can be verified after sampling, see 6.10.

Several sampling runs are also performed as part of the resampling procedure described in chapter 7. Comparing the distributions found for the various parameters across these runs is a test, since if the mixing is insufficient, the resulting distributions can be non-overlapping, whereas for good mixing, they should be highly overlapping. An example of this is the scatter plot of α and ϵ samples shown in figure 6.9. This may be seen as an intuitive relative of the $\hat{R}$ statistics convergence measure, see [42].

For synthetically generated data, it can also be verified that the distribution is sampled around the major peaks by comparing the samples of σ_w^2 with the known noise variance. If σ_w^2 is biased upwards, this indicates that the other parameters are off from the true values and have not converged.

The log likelihood is a very important indicator, and it can be inspected to determine if it seems to have converged or not, see for example figure 6.6.

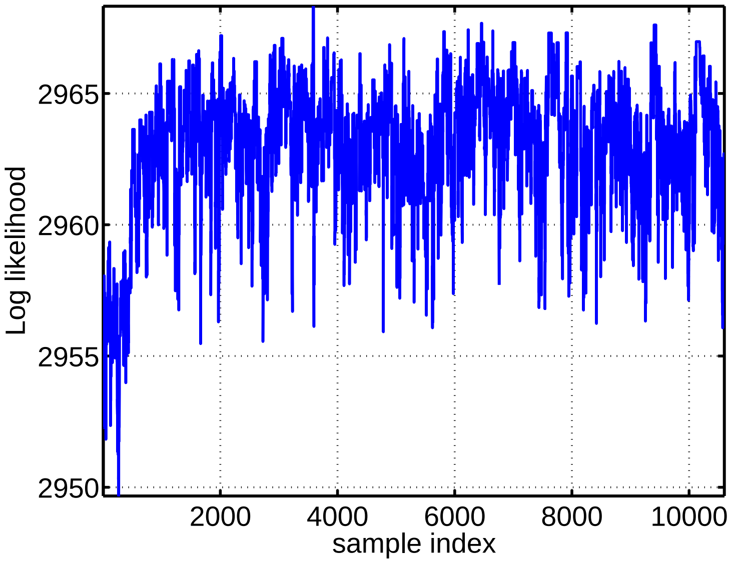


Figure 6.6: Log likelihood for one sampling run (synthetic data from the stochastic balloon model). After the first few hundreds of samples, the likelihood seems to have converged.

Most importantly, it is possible to confirm for synthetic data, that the true parameter values are contained within the obtained sampling distributions. The assumption that sampling with similar settings will yield a good approximation of the posterior distribution for real data is not accurate, since the real data were not actually generated by the model, but there is some confidence that some properties will be shared. More samples are then generated for real data than what was found necessary for the synthetic data, to take into account the likely discrepancy between real and synthetic data. This use of synthetic data seems to be one of the best indicators of how long the sampling needs to be. Parameter histograms are shown in chapter 7 together with the model comparison results. For simulated annealing, it is also possible to compare the MAP estimates with the true values, and figure 6.7 shows an example for the stochastic version of the standard balloon model where the MAP estimates for α and ϵ converges to the true values.

An example of MAP learning for the UKF is shown in figure 6.11, in which the estimate of the hidden variable $v(t)$ (blood volume) before and after learning is also shown.

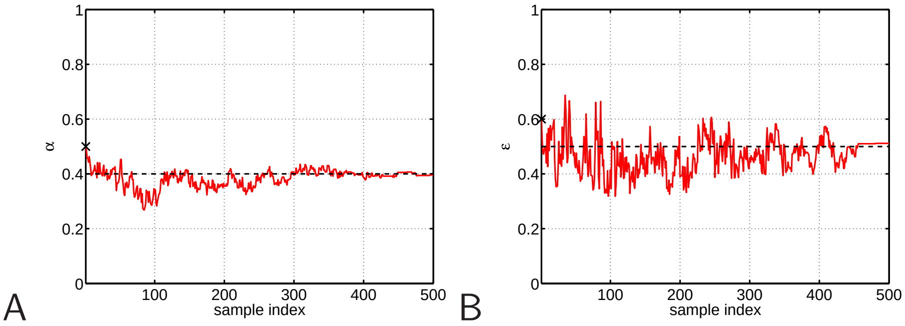


Figure 6.7: Convergence of the estimated parameters α and ϵ to the true values during simulated annealing. Synthetic data; only 500 samples from the 3000-long sampling run are shown for clarity. The dashed line shows the true value, and the initial values are marked on the y axis.

Finally, it is worth considering that even if not completely converged, the posterior approximation might still be a better approximation than those obtained using point estimate methods, such as maximum likelihood or maximum a posteriori.

6.7 Ergodicity

The sampling is constrained to be within the region defined by the priors, with the noise variance σ_w^2 being the only exception. There have been observed rare occurrences of certain combinations of parameters that lead to failure of the integration of the ordinary or stochastic differential equations. However,

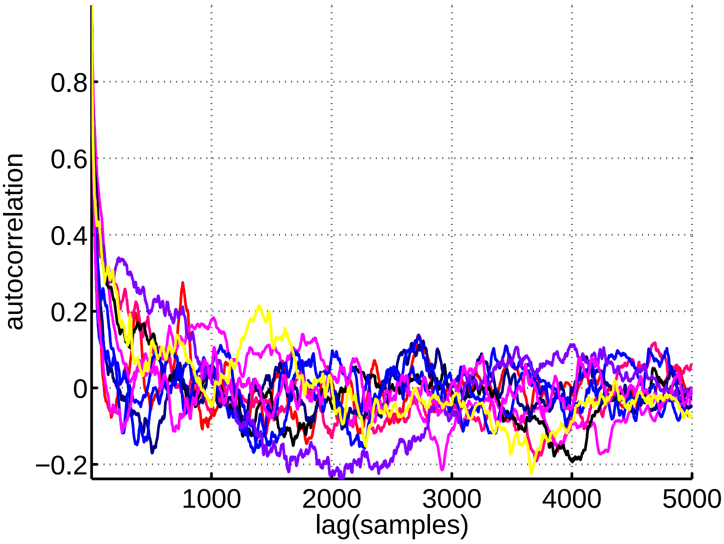


Figure 6.8: Autocorrelation for τ_0 as a function of lag, for several different sampling runs (synthetic data). This shows that around 200-1000 samples are needed for each independent sample.

as these are isolated points in parameter space, there are no barriers across the whole of the prior region that would make the sampling non-ergodic. It has been consistently found for all of the experiments with synthetic data, that the true parameters could be approximately retrieved, so ergodicity is not a practical problem with the present models.

6.8 Modifications of the stochastic augmented model

For certain hemodynamic states, the ODE's become ill-defined. For instance, if the inflow, $f(t)$ is a very small negative number, the extraction fraction becomes a very large negative number (see (3.5)), leading to unphysiological behavior and numerical problems in the solution of the system.

Also, negative in- and out-flows are not physiologically acceptable in themselves, but may occur as there is no prevention mechanism for this in the models. The models assume that inflow is solely controlled by a stimulus signal, but it could

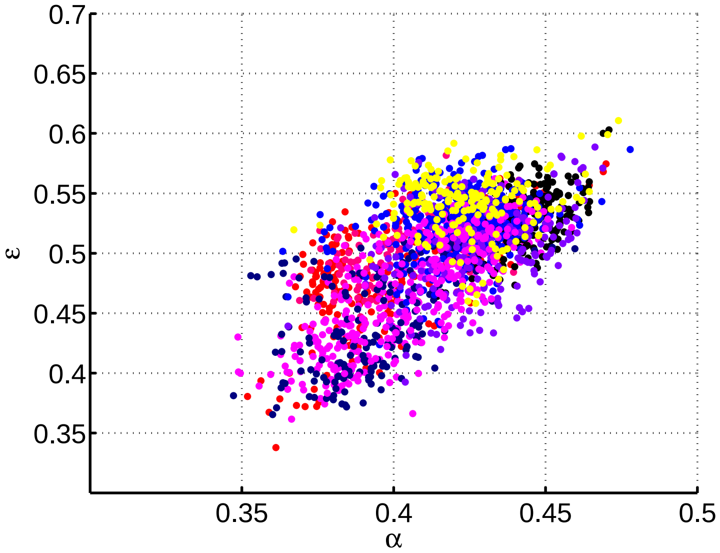


Figure 6.9: Scatter plot of α and ϵ parameter samples obtained from several sampling runs (synthetic data from the standard balloon model); notice the high degree of overlap. The true parameter values are $\alpha = 0.4$ and $\epsilon = 0.5$.

be argued that no amount of negative stimulus from neural activity could drive the blood flow to actually reverse direction. The relative blood volume, $v(t)$, may also become negative, which is not even physically conceivable.

These issues did not lead to problems for the deterministic models or the stochastic version of the standard balloon model. For the stochastic version of the augmented balloon model, however, they did, for both real data sets. Although the predicted mean states rarely entered these critical regions, some of the sigma points were found to move into them.

In order to resolve these issues, the following modifications were made for the stochastic version of the augmented balloon model. To the ODE's for $f(t)$, $f_{out}(t)$ and $v(t)$ were added terms of the form

$$\frac{c}{x(t)^2}$$

where $c < 1.0$ is a suitably small number and $x(t)$ represents any of the relevant states. This term only becomes significant in the vicinity of $x(t) = 0.0$.

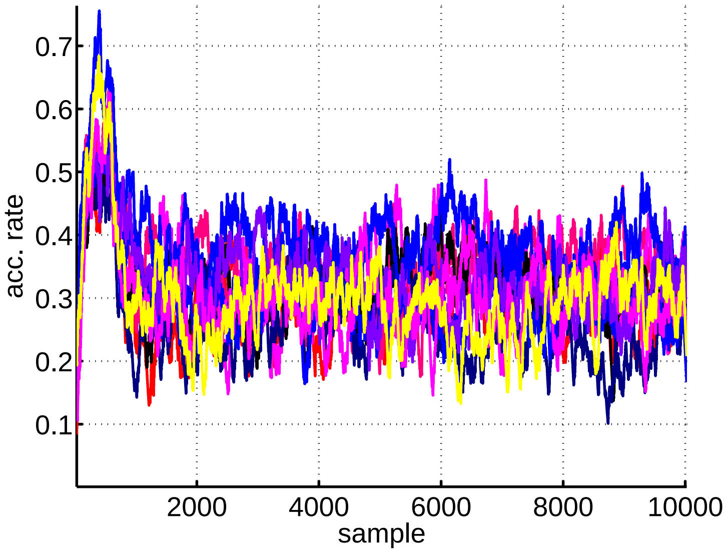


Figure 6.10: Acceptance rates for several sampling runs (synthetic data from the standard balloon model). The rates are consistently in the appropriate range.

However, even with these modifications, the stochastic version of the augmented balloon model was unable to converge (using simulated annealing) to a good fit of the BOLD signals in either of the two real data sets. Despite extensive work to find a reason for this problem, no progress was made. Due to time constraints, this model variant was not investigated further.

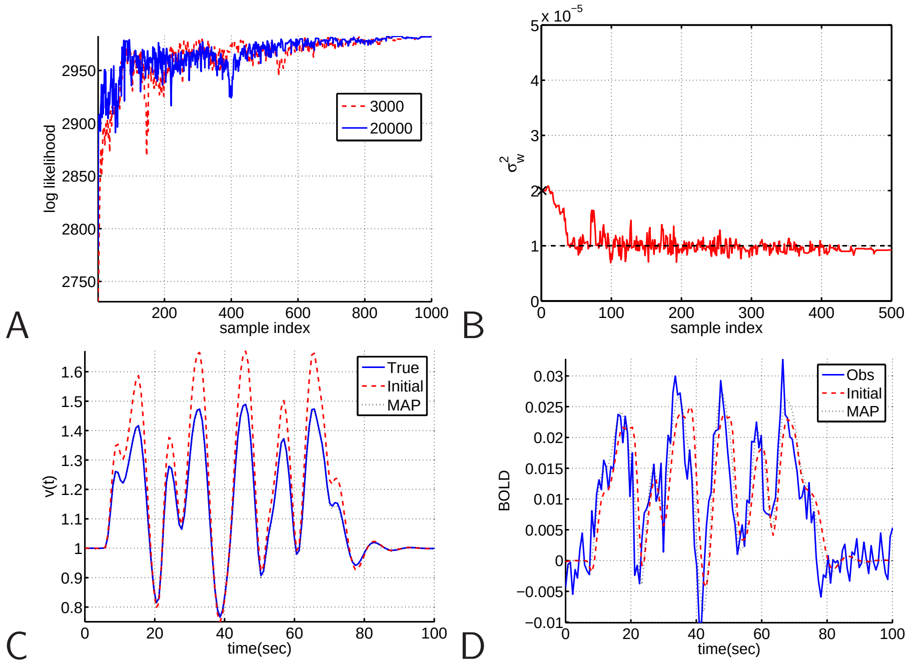


Figure 6.11: Example of MAP learning for the UKF, using synthetic data generated by the stochastic balloon model. A: Log likelihoods for two different sampling runs, one with a length of 20000 samples, the other with 3000; only 1000 samples are shown from either run for clarity. B: MAP estimate of the observation noise variance σ_w^2, is found quickly (500 samples shown; dashed line shows the true value. C: Closeup of $v(t)$ estimates. D: BOLD prediction (on training data) before and after learning.

Chapter 7

Model comparison

"All models are wrong, some are useful" - George Box

The main goal of this work is to build a framework for the evaluation and comparison of the quality of different hemodynamic models. Such a framework is important as it links the model design process, based on physiological and mathematical assumptions, to actual fMRI data - the 'real world'. This link is necessary to encourage progression towards better models and thus increase our knowledge about the brain, in particular as seen through the BOLD fMRI modality.

7.1 Model evaluation and selection

Sometimes the purpose of model evaluation is to select one model over others for a particular use. In this work, the focus is on building a framework for the evaluation of models, and not so much the actual model selection itself. The latter is context dependent, but the present model evaluation and comparison framework should allow for informed model selection in cases where model selection is desired.

7.2 Bayes factors versus prediction

The basic, classical Bayesian approach to model evaluation and comparison is the so-called 'evidence framework', in which the posterior probability (density) of each model,

$$p(M_i|D) = \frac{p(D|M_i)p(M_i)}{p(D)} \tag{7.1}$$

is considered as the measure of a model's quality. Comparison of two models is then expressed through the *Bayes factor*, $p(M_1|D)/p(M_2|D)$.

$p(D|M_i)$ is called the *evidence* for model M_i, and $p(M_i)$ is the prior for that model (usually considered equal for all models, since there is no a priori belief that one model should be better than others). The evidence is obtained by marginalization of the model parameters,

$$p(D|M_i) = \int p(D|\theta, M)p(\theta|M)d\theta \tag{7.2}$$

where the model-dependent parameter prior $p(\theta|M)$ is the same as the $p(\theta)$ used in previous chapters (only with the model dependency implicitly given by the context).

A very good review of Bayesian model selection methods is given in [12] (see also [11]), and there are extensions (especially the 'intrinsic Bayes factor') of this basic Bayesian model evaluation approach that in some circumstances can mitigate some of the drawbacks, which are mainly:

The true model must be included If the true model is not included, the result can generally not be trusted (cf. the opening quote in this chapter).

The priors must be proper If not, the numerical value of the Bayes factor is arbitrary and thus uninterpretable.

However, these drawbacks still basically characterize this approach and the Bayes factor approach is not wholly 'trustworthy' due to these difficulties. Based on these considerations, the alternative prediction framework, described in the following, is chosen as the tool for model evaluation and comparison in this work (see [79] for a similar approach, although reproducibility is not considered in this reference).

7.3 Prediction

It is natural to consider the *prediction* ability of a model that has learnt a posterior parameter distribution $p(\theta|D)$ on a training data set, and is then evaluated on a test set D^*, marginalizing (averaging) over the posterior,

$$p(D^*|D, M) = \int p(\theta|D, M)p(D^*|\theta, M)d\theta. \tag{7.3}$$

where M is the model. This integral unfortunately cannot be solved analytically due to the non-linearities involved in these models. For deterministic state space models where MCMC samples are available, the MCMC approximation

$$\int p(\theta|D, M)p(D^*|\theta, M)d\theta \approx \frac{1}{L}\sum_{i=1}^{L} p(D^*|\theta(i), M) \tag{7.4}$$

can be used, since it approximately holds that $\theta(i) \sim p(\theta|D, M))$ (L being the number of samples). For the stochastic models, the MAP approximation can instead be used to give

$$p(D^*|D, M) \approx p(D^*|\theta_{MAP}, M) \tag{7.5}$$

This prediction measure is an objective cost function that does not depend on the right model being included in the group of models to be compared, and neither do improper priors pose any difficulty. One might say that in this approach, the Bayesian idea is used as the optimal way of predicting using the posterior distribution of the parameters. The Bayesian averaging involved in the prediction is known to be optimal, in some cases even when the chosen prior is not identical to the 'true prior', i.e. the prior that is assumed to have generated the data (see [35]).

This measure can also be thought of as measuring *generalization*, i.e. the ability of a model to learn on one data set and generalize what has been learned to a different data set. Generalization is a central goal of all machine learning (see [48], so this measure is intuitively pleasing as well. The *generalization ability* for a given model, training and test set is formally defined as the logarithm of the predictive distribution,

$$G(D^*, D, M) \triangleq \log p(D^*|D, M). \tag{7.6}$$

7.4 Reproducibility

A second goal of machine learning is *reproducibility*. This is a more recent concept than generalization, and is well described in [76] and [75]. It is related to, but different to generalization.

Reproducibility concerns the sensitivity of what is learned to the particular data set used for training. This is highly relevant in model comparison, because a model that generalizes well with parameters that vary greatly depending on the particular training data set used might be less attractive than a model with a slightly lower ability to generalize, but that produces more robust posterior parameter distributions. This is particularly so in the case of 'physiological' models where the parameters carry physiological meaning. The weight one assigns to generalization and reproducibility is therefore context dependent. For instance, if de-noising of the BOLD signal is the objective, generalization will be important; if one wants to understand the mechanisms underlying the BOLD signal, reproducibility could be more important.

There is a natural way to measure reproducibility when the generalization approach is used, and that is by considering that the posterior parameter distribution, $p(\theta|D, M)$ contains the information needed on what the model has learnt. A measure is then needed of the similarity of posteriors obtained conditioned on different data sets, and a natural candidate is the Kullback-Leibler distance ([19]) between the distributions. This can be estimated when MCMC sampling approximations of the posterior are available, but not when only MAP estimates are provided. In the latter case, a simple percentage-wise deviation measure is used instead, as described below.

7.4.1 Kullback-Leibler reproducibility measure

A given split of the data into a training and a test set can be inverted by considering the training set as the test set and vice versa. By learning parameters on both sets, in this context just called D_1 and D_2, it is possible to measures reproducibility by the negative Kullback-Leibler (KL) distance between the two,

$$R(M, D_1, D_2) \triangleq - \int p(\theta|D_1, M) \log \frac{p(\theta|D_1, M)}{p(\theta|D_2, M)} d\theta \tag{7.7}$$

Higher KL distance of course equals *less* reproducibility, hence the minus sign.

Since this depends non-trivially on the dimensionality and shape of the distributions, the KL distance is averaged over dimensions, i.e. the KL distance is calculated for each dimension seperately, and the mean is then taken.

7.4.2 Kernel density estimators

Since only samples of θ are available, 7.7 cannot be calculated directly. Instead, some form of probability density estimate must be used. The traditional method is to use the histogram approximation, but this has several drawbacks, so a kernel density estimator is used instead. This is a non-parametric method where identical Gaussian distributions are used as kernels, centered on each sample, so that the posterior distribution is approximated as

$$p(\theta|D_i, M) \approx \frac{1}{L} \sum_{i=1}^{L} K(u) \tag{7.8}$$

where

$$u_i = \frac{(\theta - \theta_i)^T S^{-1} (\theta - \theta_i)}{h^2} \tag{7.9}$$

is a distance measure between θ and the i'th sample.

The covariance matrix S is calculated from all the samples, and h - the 'kernel bandwidth' - is defined as

$$h = \left\{ \frac{4}{(d+2)} \right\}^{1/(d+4)} L^{-1/(d+4)} \tag{7.10}$$

where $d = \dim \theta$. For further details, see [57].

It was found that using 100 uniformly distributed samples from the entire MCMC sample was enough to get a good estimate of the KL distance.

7.4.3 Percentage deviation reproducibility measure

As an estimate of reproducibility for the models learned through MAP parameter estimates is chosen the negative of the percentage-wise difference between the parameter estimate from each split of the training data and the mean of the two estimates,

$$R(M, D_1, D_2) \triangleq -\frac{|\hat{\theta}_1 - \mu_\theta|}{\mu_\theta} \tag{7.11}$$

where $\hat{\theta}_i$ is θ_{MAP} for data set i and $\mu_\theta \triangleq (\hat{\theta}_1 + \hat{\theta}_2)/2$ is the mean of the two estimates.

When one (deterministic) model has been learnt as MCMC samples, and another (stochastic) as a MAP estimate, the mean of the approximated MCMC posterior can be used as a representative point estimate of the distribution,

$$\hat{\theta} = \frac{1}{L}\sum_{i=1}^{L} \theta_i \tag{7.12}$$

and the percentage-wise measure of reproducibility can then be used for both models.

7.5 Relation to AIC and BIC

Several model comparison methods exist that rely on asymptotic assumptions, meaning that they can only be relied on to give accurate comparisons when the samples size is very large (see e.g. [44]). Since in the present case this is not the case, they are not really applicable.

However, it is interesting to note that the Bayesian Information Criterion is based on Bayes factors, whereas the Aikaike Information Criterion is based on expected likelihoods over all possible data sets. This means that the AIC is closer in principle to the present approach, and the resampling method described below.

7.6 Resampling

The generalization measure for a given model depends on a particular choice of training and test data sets (see (7.6)). But a better generalization estimate would be independent of these choices, and so they should be marginalized. First, the mean predictive generalization over test sets is

$$G(D,M) \triangleq \langle G(D^*,D,M)\rangle_{p(D^*)} = \int [\log p(D^*|D,M)]p(D^*)dD^*.$$

From this it can be seen that apart from an additive constant (the entropy of the true distribution), the mean generalization so defined is equal to minus the Kullback-Leibler distance between the 'true' distribution of the data, $p(D^*)$, and the model distribution, $p(D^*|M)$.

This measure still depends on the choice of training data set. Marginalizing again, the average generalization over training data sets is

$$\begin{aligned} G(M) \triangleq \langle G(D,M)\rangle_{p(D)} = \left\langle \langle G(D^*,D,M)\rangle_{p(D^*)} \right\rangle_{p(D)} = \\ = \int \left[\int \log p(D^*|D,M)p(D^*)dD^* \right] p(D)dD \end{aligned}$$

These integrals can not be computed with the present models. However, using a cross-validation approach (see e.g. [13], [69]), they can be estimated using a *resampling* procedure. Resampling is a method of obtaining statistical estimates when the sample size is small and traditional large-sample methods are inapplicable ([3]). Here, the method of 'split-half resampling' is used, mainly because of its suitability for estimating reproducibility (see [76], [75]). The available data (BOLD data divided into independent epochs) is split randomly into two halves multiple times, in this way 'sampling' from the simultaneous distribution of test and training data sets, $p(D,D^*)$. This gives an approximate generalization measure,

$$G(M) \approx \frac{1}{K}\sum_{i=1}^{K} \log p(D_i^*|D_i,M) = \frac{1}{K}\sum_{i=1}^{K} G(D_i^*,D_i,M). \tag{7.13}$$

With 10 quasi-independent epochs available, each split of the data uses 5 each for the test and training sets. Each of the resulting estimates is an unbiased estimate of the model generalization. The mean over all splits is a convex combination and thus also an unbiased estimate, but with a reduced variance. In fact, each split allows either half-set to be treated as training and test set, so there are actually 2 estimates of the generalization with each split. The maximal size of K depends on how many epochs are available, but only $K = 20$ have been used here.

In parallel with the estimation of generalization (7.13), the KL distance is measured between the two posterior parameter distributions of each split half, and the reproducibility is then the average over all splits

$$R(M) = -\frac{1}{K}\sum_{i=1}^{K}\int p(\theta|D_i^1, M)\log\frac{p(\theta|D_i^1, M)}{p(\theta|D_i^2, M)}d\theta \tag{7.14}$$

where D_i^1 and D_i^2 are the two data sets in split i.

Similarly, the percentage-wise reproducibility measure is estimated as

$$R_i(M) \triangleq -\frac{1}{K}\sum_{i=1}^{K}\frac{|\hat{\theta}_{i1} - \mu_\theta|}{\mu_\theta} \tag{7.15}$$

where $R_i(M)$ is the estimate for the current split and $\hat{\theta}_{i1}$ is the parameter estimate for the first half (or second, as the result would be the same) of the current split, and μ_θ is the mean of the two.

Resampling is a widely applicable technique, not least the so-called 'bootstrapping' approach, see e.g. [3], [34], [22].

7.7 Generalization and reproducibility tradeoff

Generalization and reproducibility performance estimates present the model user with a tradeoff. A model that is incapable of fitting the data well may still learn very similar parameters (or distribution thereof), independent of which data are presented to it. But most likely, such a model will be useless because it cannot 'explain' the data. On the other hand, a more complex model may learn the data well, measured in terms of generalization ability, yet may give very different answers for the parameters depending on the data set it learns from. This could be due to internal invariances, i.e. the likelihood function $p(D|\theta)$ could be very flat in some regions, or have multiple optima of similar values. Thus, the MAP parameter estimate would change with small random perturbations such as would occur with changes in the data set, and the posterior distribution would reflect these invariances to some degree (depending on how accurate an approximation is obtained). Such a model may be useless in another way, namely if the parameters have some meaning and the information they carry is to be interpreted, such as with the physiological parameters in the hemodynamic models. It is therefore up to the users to select which model is the best for the task at hand. Sometimes, of course, a certain model might have both the highest generalization ability and the highest reproducibility, and the choice would then be clear; but this is not typically the case.

7.8 Experimental results

Results are presented here in order of data sets, starting with the synthetic data. First, a brief description of the types of results shown is given.

7.8.1 Parameter histograms

For the deterministic models, histograms are shown for the MCMC sampling of the parameters. The initial 'burn-in' samples, including those used to estimate the proposal distribution, but also those samples deemed to precede convergence, are disregarded. The samples from all the resampling splits are pooled together. Some split-half sampling or simulated annealing runs may not have converged (as determined by the convergence criteria outlined in section 6.6), and those are excluded prior to further analysis.

7.8.2 Mean BOLD predictions

For each iteration of the split-half resampling, a predicted mean BOLD signal on the test data results. This allows a mean of the mean prediction to be shown, together with confidence intervals for the mean.

The prediction of the mean BOLD signal for any sample in a test data set is given by

$$E[g(y_n)] = \int g(y_n)p(y_n|D) \tag{7.16}$$

where D is the training set. This mean prediction is of course closely related to the likelihood, and it is approximated using the MCMC samples in the same way,

$$\int g(y_n;\theta)p(\theta|D) \approx \frac{1}{N}\sum_{i=1}^{N} g(y_n;\theta_i) \tag{7.17}$$

For the MAP estimates (stochastic model), the approximation is

$$\int g(y_n;\theta)p(\theta|D) \approx g(y_n;\theta_{MAP}) \tag{7.18}$$

7.8.3 Empirical confidence intervals

For the MAP parameter estimates and the mean BOLD test predictions (7.18), there are only as many estimates as there are splits. For these, empirical confidence intervals can be calculated, without needing to assume normal distributions. For a desired $f \cdot 100$-percent confidence interval, the empirical confidence interval can be found by simply sorting the N samples and removing $M/2$ of the lowest and highest values, where $M = (1-f)N$. The interval is then defined by the lowest and highest of the remaining samples.

However, since the number of splits used is never greater than 20, the variance of the empirical confidence interval limits is high. Therefore, the choice was made to use the interval bounded by the lowest and highest of the samples, corresponding to a 100 % confidence interval. It should be noted, of course, that the limits of these intervals have significant variance, but they do capture some information as to the uncertainty of the estimate.

The BOLD predictions and their bounds shown in all figures use these empirical confidence intervals. It must be kept in mind that they apply to the variance of the mean prediction, and that the total variance of the BOLD signal is composed of this mean variance and the observation noise variance, modelled by σ_w^2.

7.8.4 Convergence of parameters

As en example of the convergence of the parameters to the true values for synthetic data, the standard balloon model was learnt on the synthetic data generated by itself. The histograms are shown in figures 7.1 and 7.2.

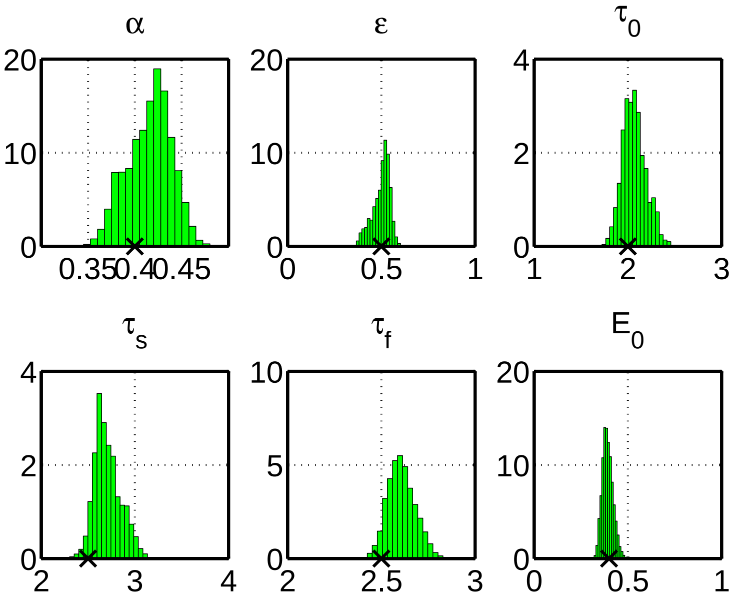


Figure 7.1: Histograms of the hemodynamic parameters of the standard balloon model, learned from a synthetic data training set generated by itself. The true parameters are marked on the x axis of each figure.

As the figures show, the true parameters are generally contained within the posterior distribution and quite close to their mean. The corresponding prediction on test data is shown in figure 7.3. This procedure was also carried out for the augmented balloon model and the stochastic version of the standard balloon model, and convergence was also found (not shown).

For the deterministic models learning on synthetic data, 15.000 MCMC samples (after burn-in) were found to be sufficient, while 3.000 simulated annealing samples were enough for convergence in the case of the stochastic model. For the real data sets, 40-50.000 samples were generated for the deterministic models, and 4.000 simulated annealing steps where done for the stochastic model.

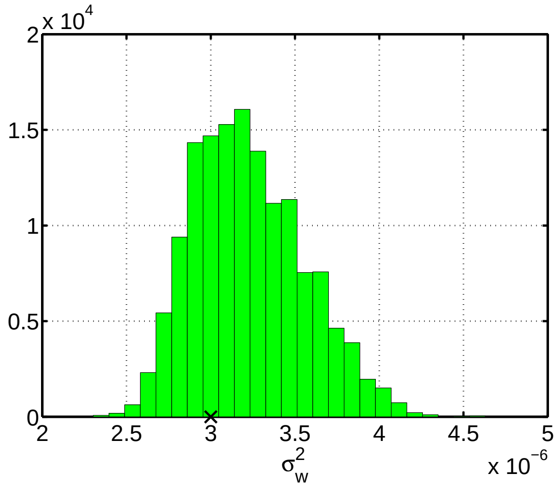


Figure 7.2: Histograms of the observation noise variance of the standard balloon model, learned from a synthetic data training set generated by itself. The true parameter is marked on the x-axis.

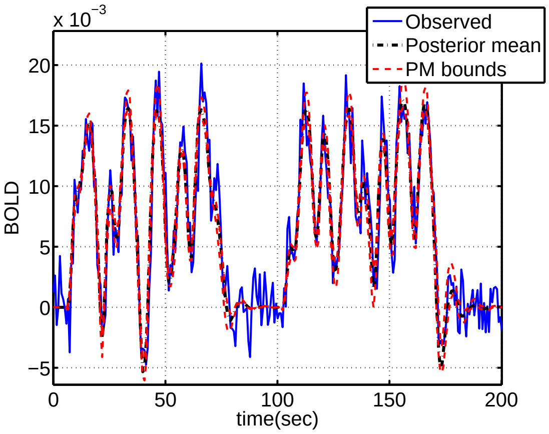


Figure 7.3: Test prediction of the first two epochs for the standard balloon model, trained on synthetic data generated by itself.

7.8.5 Deterministic models - synthetic data

The generalization and reproducibility of the deterministic models were first compared on the synthetic data set generated by the standard balloon model.

Figure 7.4 shows the result.

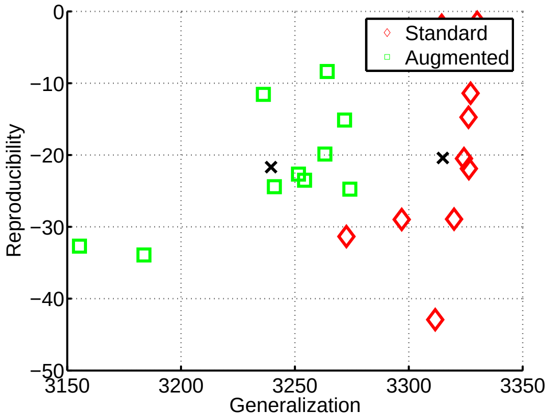


Figure 7.4: Generalization and reproducibility for the two deterministic models, evaluated on the synthetic data set generated by the standard balloon model.

When data are generated by the simpler standard balloon model, then - as might be expected - both the generalization ability and reproducibility are seen to be higher for the simple model, although reproducibility is not significantly higher. But when data are generated by the augmented balloon model, the situation is more complex: the true (augmented) model is able to generalize significantly better, but the deterministic model is still more reproducible, see figure 7.5. This relatively poor reproducibility is probably due to the added complexity of the model, and means that either model could be chosen as the 'best' one, depending on the intended use of the model. It is possible that with higher variance in the hidden state noise, the true model would outperform the simpler model in both reproducibility and generalization.

7.8.6 Deterministic models - data set 1

Figures 7.6 and 7.7 show the histograms of the parameters of the standard balloon model, when it is trained on data set 1. The most outstanding feature is the high range for α, a priori expected to be close to 0.4, but here has a posterior mean of 0.92. This corresponds to a much reduced stiffness, but it is hard to state wether or not this is very surprising. As noted in [23], the effect of changing α is not very marked. τ_f is also floating around the maximum of the prior ($\tau_f = 8$) and it might be interesting to see the effect of increasing the upper bound in the prior. For the other parameters, the distributions are more as expected.

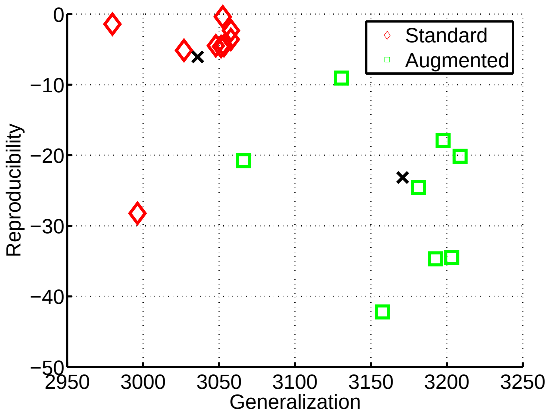


Figure 7.5: Generalization and reproducibility for the two deterministic models, evaluated on the synthetic data set generated by the augmented balloon model.

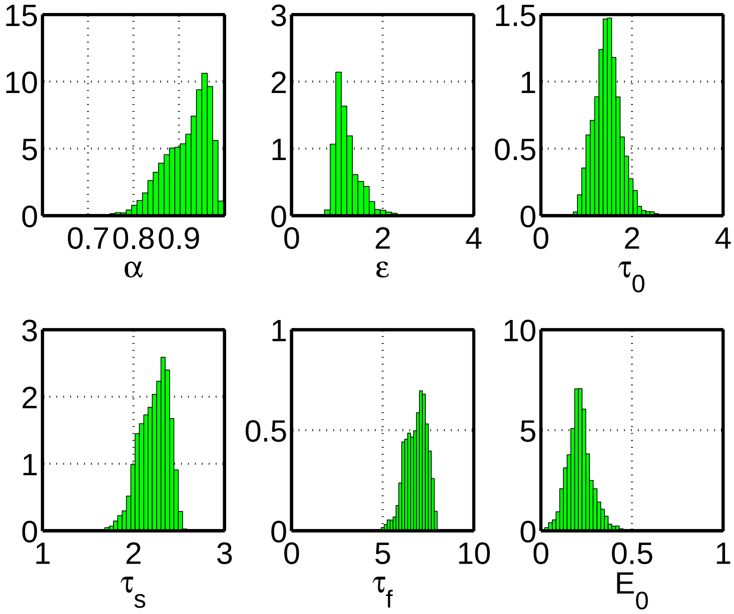


Figure 7.6: Histograms of the hemodynamic parameters for the standard balloon model, learnt on data set 1. Note the high values of α, the inverse stiffness parameter.

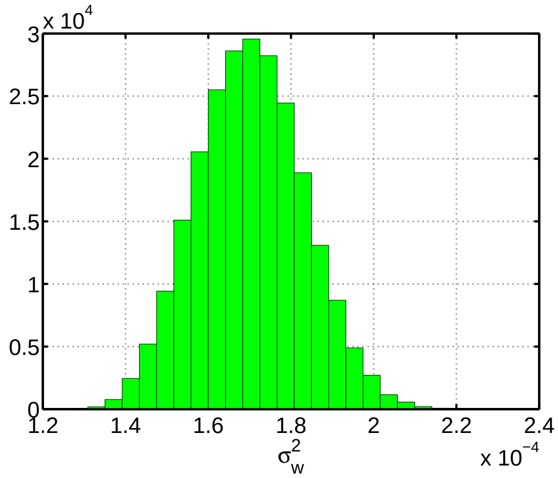


Figure 7.7: Histograms of the observation noise variance parameter for the standard balloon model, learnt on data set 1.

Figures 7.8, 7.9 and 7.10 show the histograms for the augmented balloon model for this data set. Here, α is more in the expected range, with the exception of a few samples in the 0.9 area, as with the standard model. E_0, however, is very high and very close to 1.0 for many samples. This may indicate a bad fit of this model to this data, since extraction fractions above 0.55 (see [23]) are unusual. The distributions of κ and τ_u correspond to a very square-like neural activity function, which makes sense if the standard model is indeed the best model for this data. Together, these results do seem to indicate that a square-pulse neural activity function is a good approximation for this data set. τ_+ and τ_- are not identical, but beyond that it is best left up to physiological experts to interpret the distributions, although it may not be very relevant if the augmented model is not suitable for this data set. σ_w^2 is slightly higher than for the standard model, confirming a relative inability to fit the data for the augmented model.

The BOLD predictions on test data for the two models are shown in figure 7.11, together with the generalization and reproducibility comparison. The performance diagram shows very clearly that the standard balloon model is the best choice for this data set. The BOLD test prediction for the augmented balloon model is particularly bad for the first epoch (figure 7.11C), but is more similar to that shown for the second epoch for the other epochs (not shown).

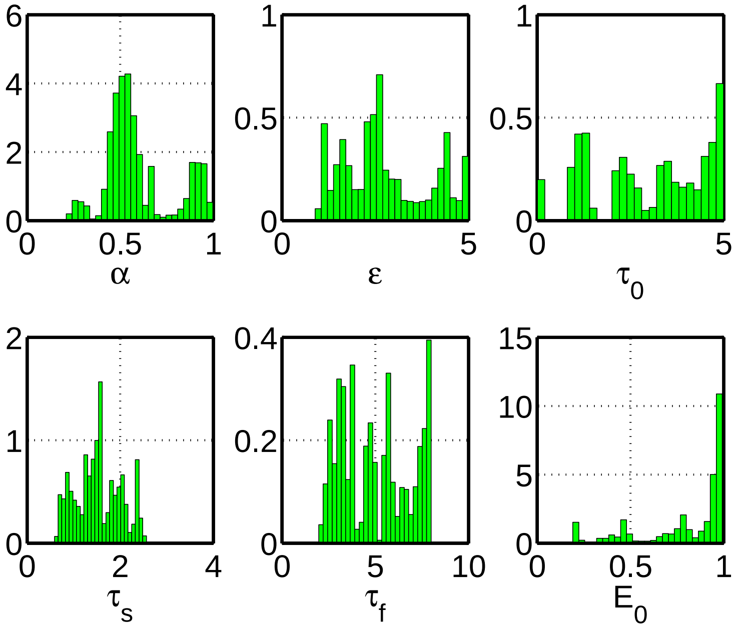


Figure 7.8: Histograms of the observation noise variance parameter for the augmented balloon model, learnt on data set 1. Note the extremely high value of E_0.

7.8.7 Deterministic models - data set 2

The parameter histograms for the standard balloon model are shown in figures 7.12 and 7.13. Compared to the parameter distributions found for data set 1 (figure 7.6), α is distributed around 0.5 and is thus much more in accordance with physiological expectation. It would also seem that the marginal distributions for τ_f and τ_s are bi-modal, corresponding to two different 'explanations' of the signal being given by the model. As for the previous histograms, it may also be noted that the posterior distributions generally have a much lower variance than the prior distributions, confirming that learning has indeed taken place.

The augmented balloon model parameter histograms are shown in figures 7.14, 7.15 and 7.16. The parameters that are shared with the standard balloon model are not all that differently distributed, except for τ_0 and E_0. The neural activity parameters are not indicating a square-pulse neural activity shape for this data set, and this is interesting considering the very different stimulus signal used to generate this data set. In [15] it was found that for long stimulus pulses (> 3

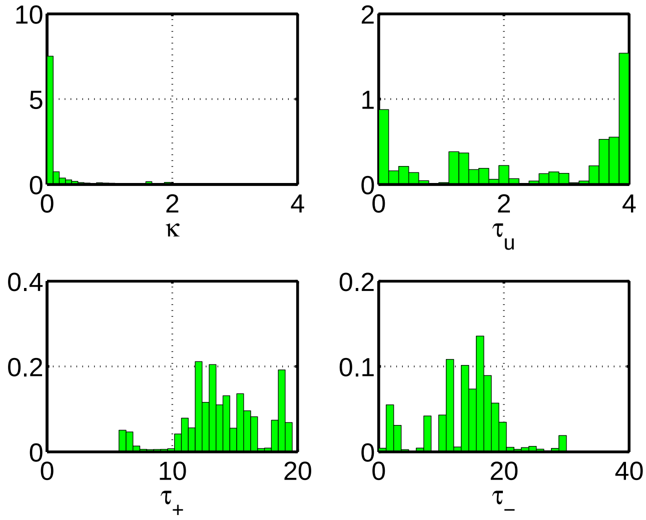


Figure 7.9: Histograms of the additional parameters of the augmented balloon model, learnt on data set 1.

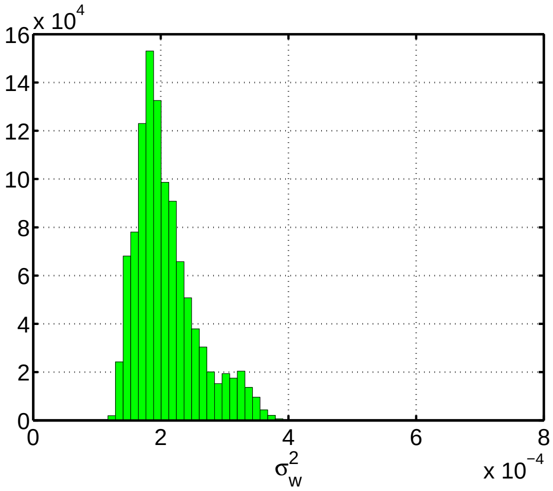


Figure 7.10: Histograms of the observation noise variance parameter for the augmented balloon model, learnt on data set 1.

seconds), linear models were adequate, while for 3-second pulses, they were not. [15] also report from [14] and [2] that neural adaptation occurs after 0.5 to 7 seconds after prolonged visual stimulation, which can be compared to the mean

τ_u found here of around 2.0 seconds.

Interestingly, the distributions for τ_+ and τ_- are similar for both data sets. The distribution of σ_w^2 for both models are very similar (figure 7.13 and figure 7.16), indicating that both models attain roughly the same error on the training data.

The performance results are less clear-cut than for data set 1, see figure 7.17. Still, the standard balloon model does seem to be a better model both in terms of generalization and reproducibility, as for data set 1. The BOLD test predictions are hard to distinguish visually, giving an indication of the sensitivity of the generalization and reproducibility metrics.

7.8.8 Standard and stochastic balloon models - synthetic data

As the augmented balloon model was somewhat overshadowed by the standard model for both real data sets, the next step taken was to compare the standard balloon model (also referred to here as the 'deterministic model') to the stochastic version of the standard balloon model (also referred to as just the 'stochastic model'). Also, as noted in section 6.8, the stochastic version of the augmented balloon model exhibited problems with convergence.

The results for the synthetic data set generated by the standard balloon model are shown in figure 7.18. Although the BOLD test predictions look very similar, the generalization-reproducibility scatter plot reveals the better performance of the true model.

For the synthetic data generated by the stochastic model, the result is not quite the opposite, see figure 7.19. As before, the more complex model has a lower reproducibility, even when it is the true model. However, the generalization ability is clearly best for the true, stochastic model.

7.8.9 Standard balloon model and the stochastic version - data set 1

Figure 7.20 shows that the deterministic model has both a higher reproducibility and a higher generalization ability than the stochastic model for this data set. This is a very clear result, again demonstrating a very high degree of suitability of the standard balloon model for this (type of) data. The BOLD test predictions also reveal that the stochastic model's mean predictions are less certain.

7.8.10 Standard balloon model and the stochastic version - data set 2

Figure 7.21 shows that for this more 'complex' data set, the stochastic model proves to have a higher generalization ability, although again, the simpler deterministic model is more reproducible. The ability of the stochastic model to express greater variation in the mean BOLD signal through the addition of noise in the hidden state space and thus fit to more complex BOLD signals seems to be rewarded for this data set. It seems that yet again, increased flexibility comes at a price, namely a reduction in reproducibility.

The mean values of the MAP parameter estimates of the stochastic model are shown in table 7.1, together with the smallest and highest values across all split-half resampling estimates. Compared to the distributions obtained for the deterministic models, the mean values are very different, but the variances are roughly similar. The much lower estimate of σ_w^2 is deceptive, since there are other noise sources in the stochastic model, not present in the deterministic one. For these hidden state noise sources it is interesting to see that the standard deviations for $v(t)$ and $s(t)$ are roughly double that of $q(t)$ and $f(t)$. This might indicate that the components of the model corresponding to the former states are most lacking in accuracy, leading to the possible interpretation that these two components should be modified.

	Mean	Smallest	Highest
α	0.2105	0.0987	0.376
ϵ	0.2698	0.1493	0.766
τ_0	2.7675	2.1667	3.301
τ_s	1.2212	0.8889	2.166
τ_f	3.0462	2.0935	3.971
E_0	0.5473	0.3337	0.763
σ_v	0.1090	0.0617	0.202
σ_q	0.0448	0.0107	0.090
σ_f	0.0368	0.0041	0.102
σ_s	0.0815	0.0300	0.160
σ_w^2	1.24e-005	4.90e-006	3.76e-005

Table 7.1: MAP parameter estimates for the stochastic version of the standard balloon model, trained on data set 2.

7.9 Discussion

The results presented here indicate that the simpler deterministic model - the standard balloon model - is better than both the augmented balloon model and the more complicated stochastic state space model for the real data based on a straight forward block stimulus design (data set 1). When compared on data generated with a more complex stimulus function, the stochastic model is shown to be better able to capture the structure of the resulting BOLD signal. The price seems to be a reduced reproducibility, so that the physiological interpretation of the stochastic state space model is less clear. These results indicate the crucial importance of considering the context and task when doing model comparisons. In [8], the conclusion (based on visual and motor tasks) was that probably both the neural and the hemodynamic activity were non-linear. The results presented here seem rather to indicate that the non-linearity is first and foremost in the hemodynamics, yet again underlining the dependence of the results on the exact task and setup (see also [64]).

With more data, generalization and reproducibility would increase for both models, and a different comparison result could be obtained. It would be of great interest to see similar comparisons for other amounts and types of data - such as across different parts of the brain and different stimuli - and for other models not investigated here (e.g. [46], [7] and [83]).

7.9.1 Sources of variability

The variance in the performance estimates will come from at least four different sources.

The first is the result of mathematical invariances in the model itself. For instance, as mentioned before (see 3.6.2), increasing κ and decreasing ϵ together can produce similar BOLD responses, as the increase in κ leads to weaker neural pulses, for which the increase in the 'gain' factor ϵ can compensate.

The second type of variance is from the data itself - the likelihood for a given data set is not sharply peaked and thus gives a natural variance in the posterior. There may also be several regions in parameter space with similar likelihoods (local optima).

Both of these sources result in flat regions or ridges in the likelihood and probably also multiple modes, that are reflected in the posterior.

Third, the MCMC method does not give the same result on every run. Indeed, this realization is the very reason for doing multiple MCMC runs, so that one can with reasonable confidence believe that the variance from the algorithm has been exhausted and further runs will not reveal significant new information on the posterior distribution.

Finally, the resampling framework introduces variance of its own, in that each split of the data will lead to slightly different posterior approximations.

From a physiological viewpoint, only the first of these sources of variation is a nuisance - the others all reveal informative variation present in the data with respect to the model under investigation. The mathematical invariances, however, should preferably be relatively small, or else the physiological, interpretative usefulness of the model comes in question. The physiological priors hopefully help in dampening these invariances, as was seen with κ and τ.

Figure 7.22 shows the likelihoods of the split-half simulated annealing runs for the stochastic model (learning on data generated by itself) and the corresponding convergence of one of the parameters (α). There is some bias from the truth ($\alpha = 0.4$) across the splits, and this is most likely mainly due to the noise in the data. However, the split-half farthest from the truth is also the one with the lowest likelihood, illustrating that some of the bias may be from incomplete convergence of the simulated annealing algorithm.

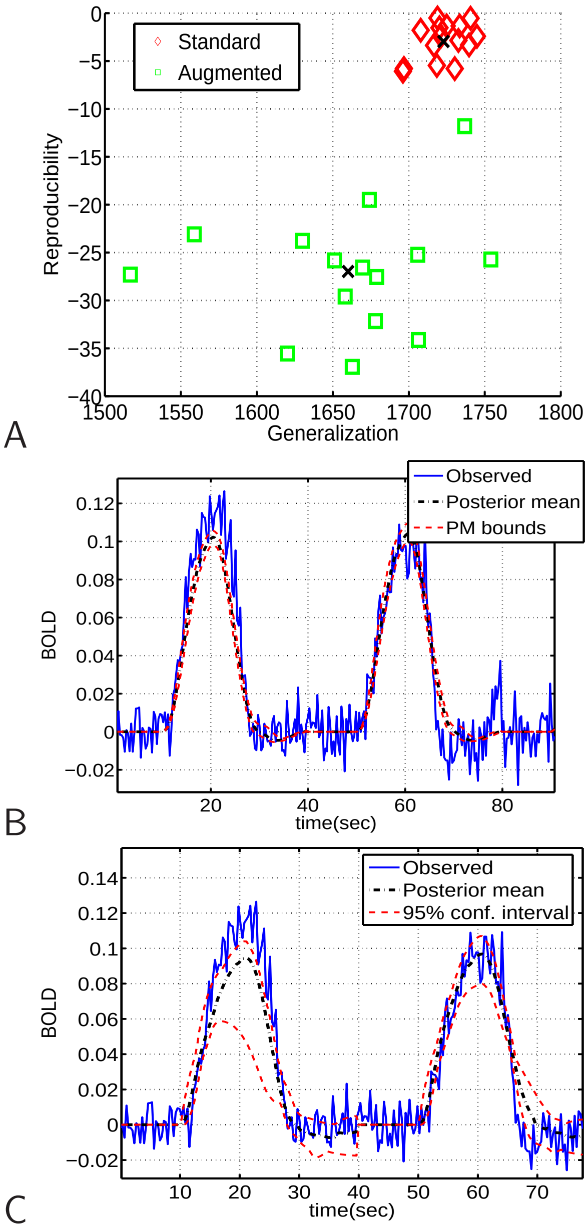


Figure 7.11: Standard and augmented balloon model results for data set 1. A: Comparison of generalization (G) and reproducibility (R) of the standard and augmented balloon models. B: Prediction of the first two epochs of two test epochs by the standard balloon model. C: Prediction of the first two epochs of two test epochs by the augmented balloon model.

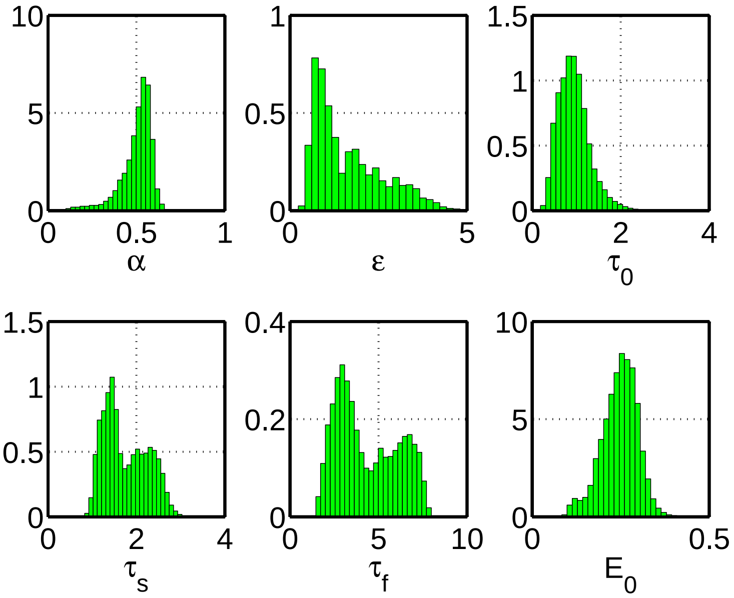


Figure 7.12: Histograms of the hemodynamic parameters for the standard balloon model, learnt on data set 2. Note the bi-modal appearance of two of the time constants, τ_s and τ_f.

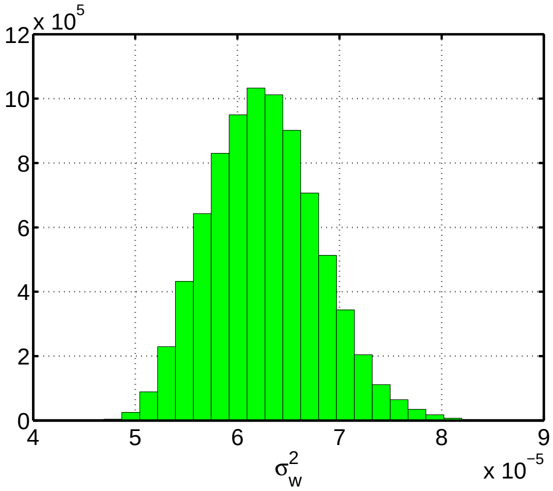


Figure 7.13: Histograms of the observation noise variance for the standard balloon model, learnt on data set 2.

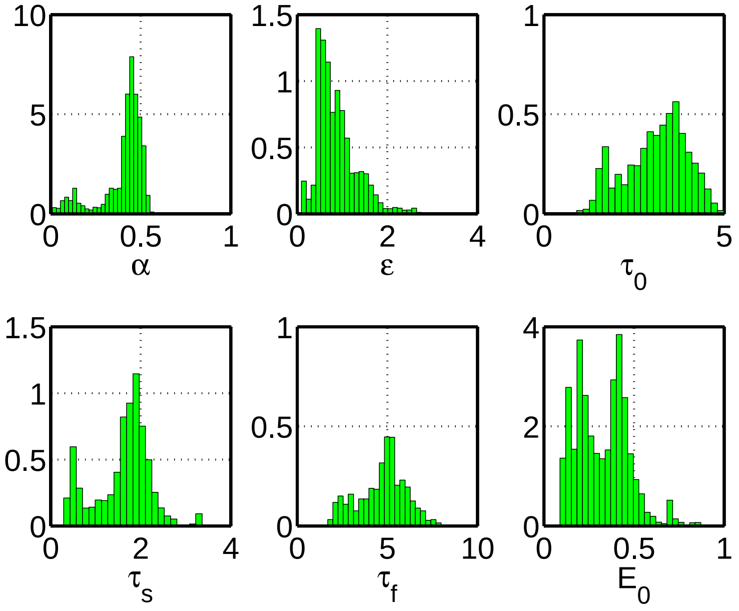


Figure 7.14: Histograms of the hemodynamic parameters for the augmented balloon model, learnt on data set 2.

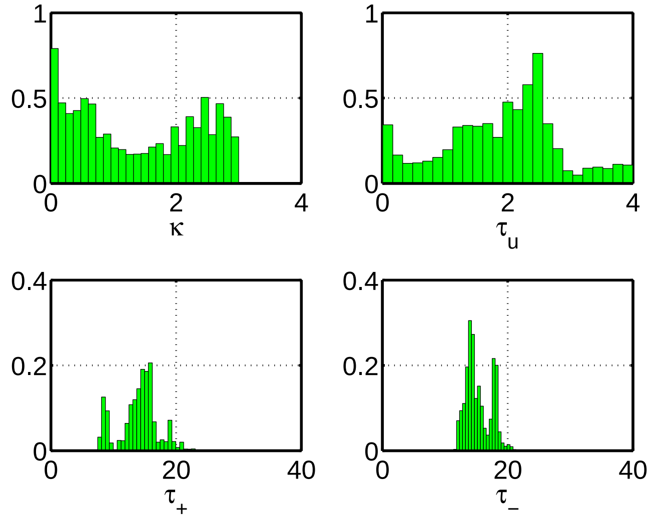


Figure 7.15: Histograms of the hemodynamic parameters for the augmented balloon model, learnt on data set 2.

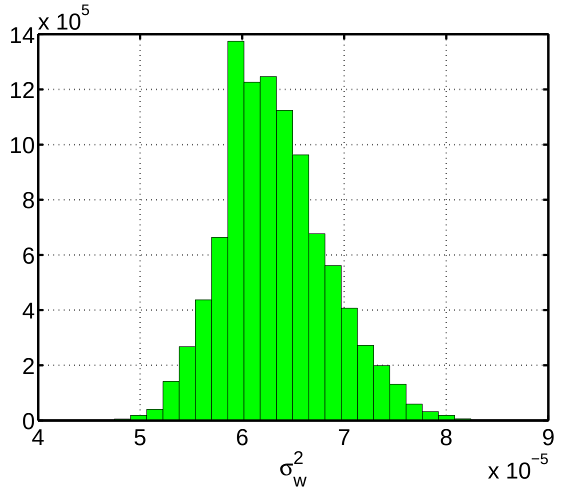


Figure 7.16: Histograms of the observation noise variance for the augmented balloon model, learnt on data set 2.

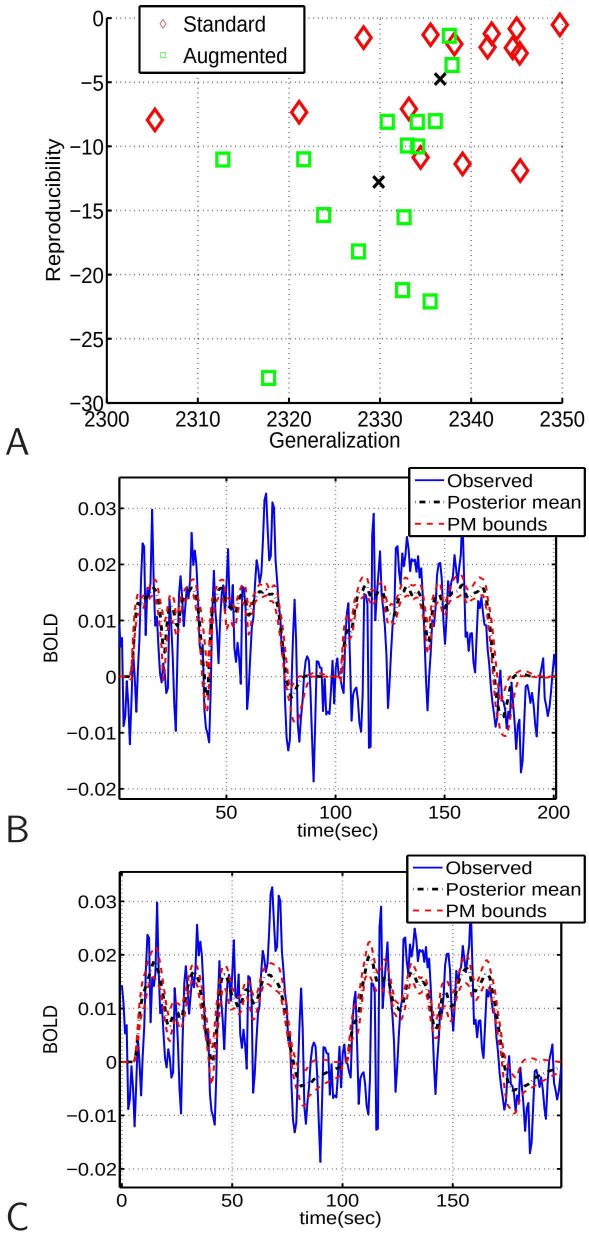


Figure 7.17: Standard and augmented balloon model results for data set 2. A: Comparison of generalization (G) and reproducibility (R) of the standard and augmented balloon models. B: Prediction of the first two epochs of two test epochs by the standard balloon model. C: Prediction of the first two epochs of two test epochs by the augmented balloon model.

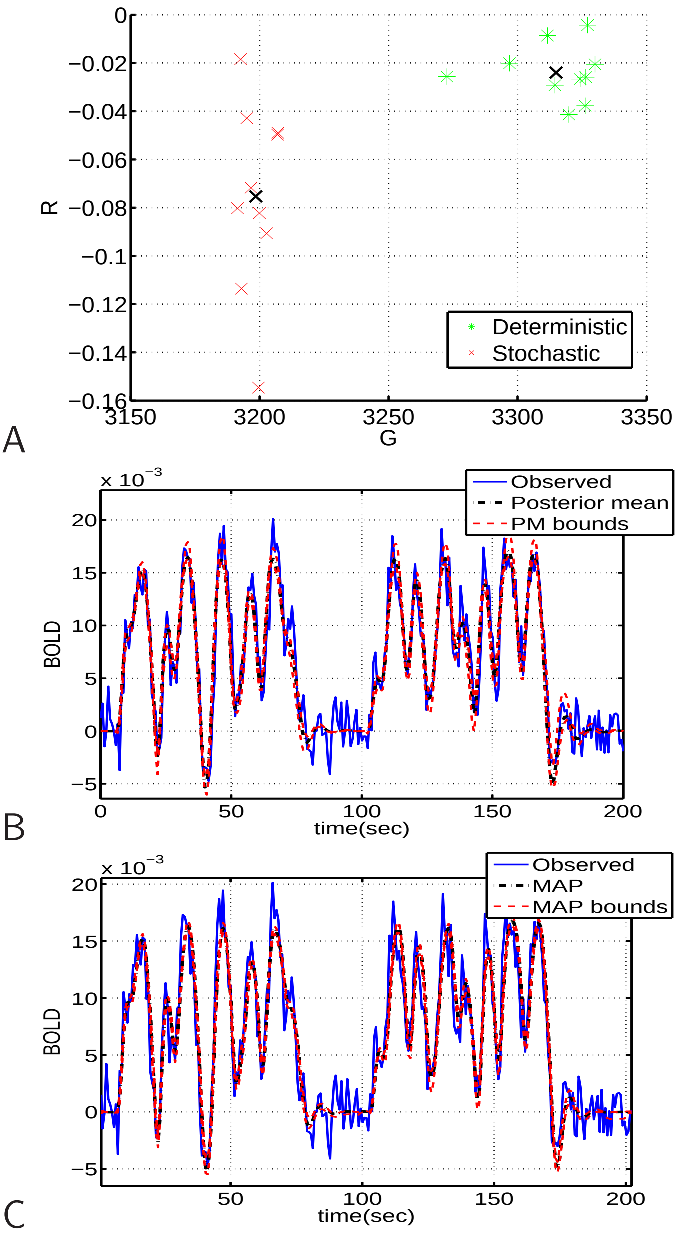


Figure 7.18: Results for synthetic data, generated by the deterministic model, and learnt by that and the stochastic model. A: Comparison of generalization (G) and reproducibility (R) of the deterministic state space and the stochastic state space models. B: Prediction of the first two epochs of two test epochs by the deterministic model. C: Prediction of the first two epochs of two test epochs by the stochastic model.

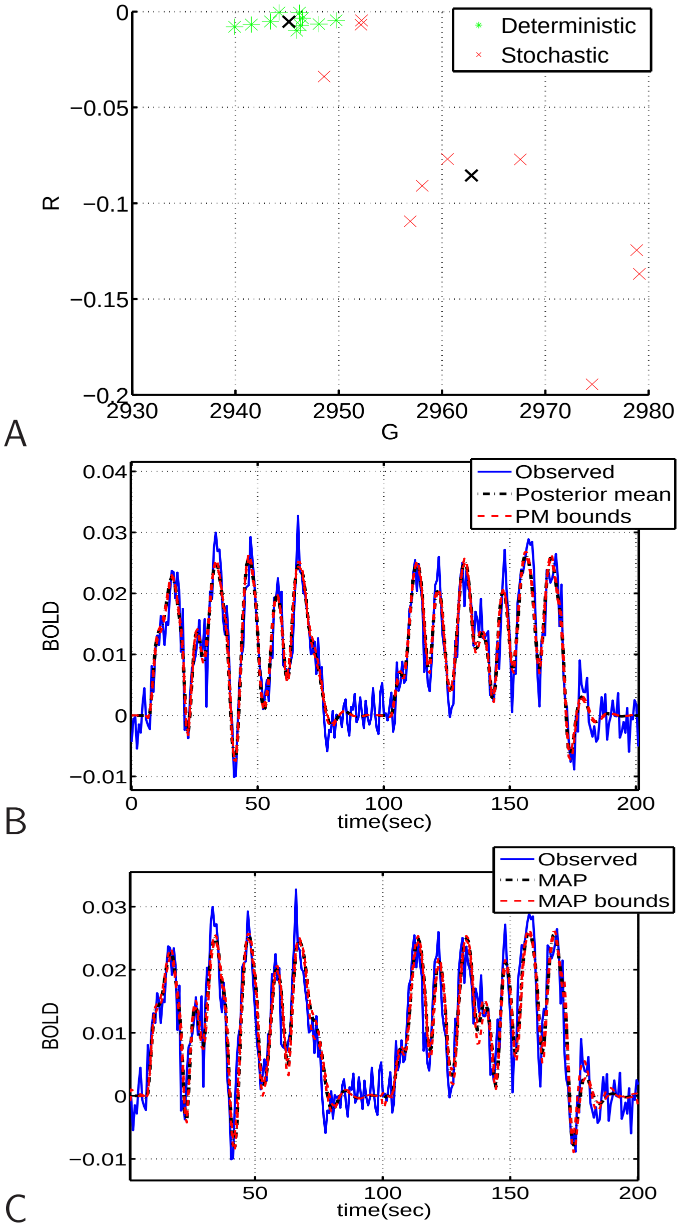


Figure 7.19: Results for synthetic data, generated by the stochastic model, and learnt by that and the standard balloon model. A: Comparison of generalization (G) and reproducibility (R) of the deterministic state space and the stochastic state space models. B: Prediction of the first two epochs of two test epochs by the deterministic model. C: Prediction of the first two epochs of two test epochs by the stochastic model.

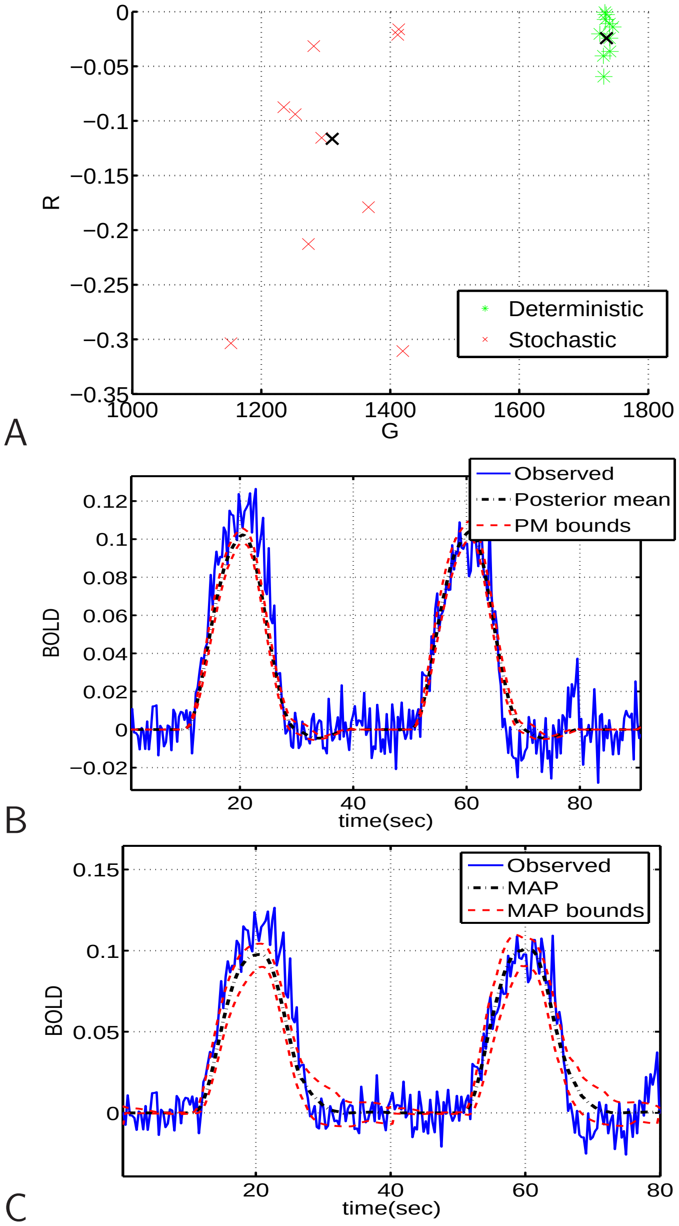


Figure 7.20: Deterministic and stochastic model results for data set 1. A: Comparison of generalization (G) and reproducibility (R) of the deterministic state space and the stochastic state space models. B: Prediction of the first two epochs of two test epochs by the deterministic model. C: Prediction of the first two epochs of two test epochs by the stochastic model.

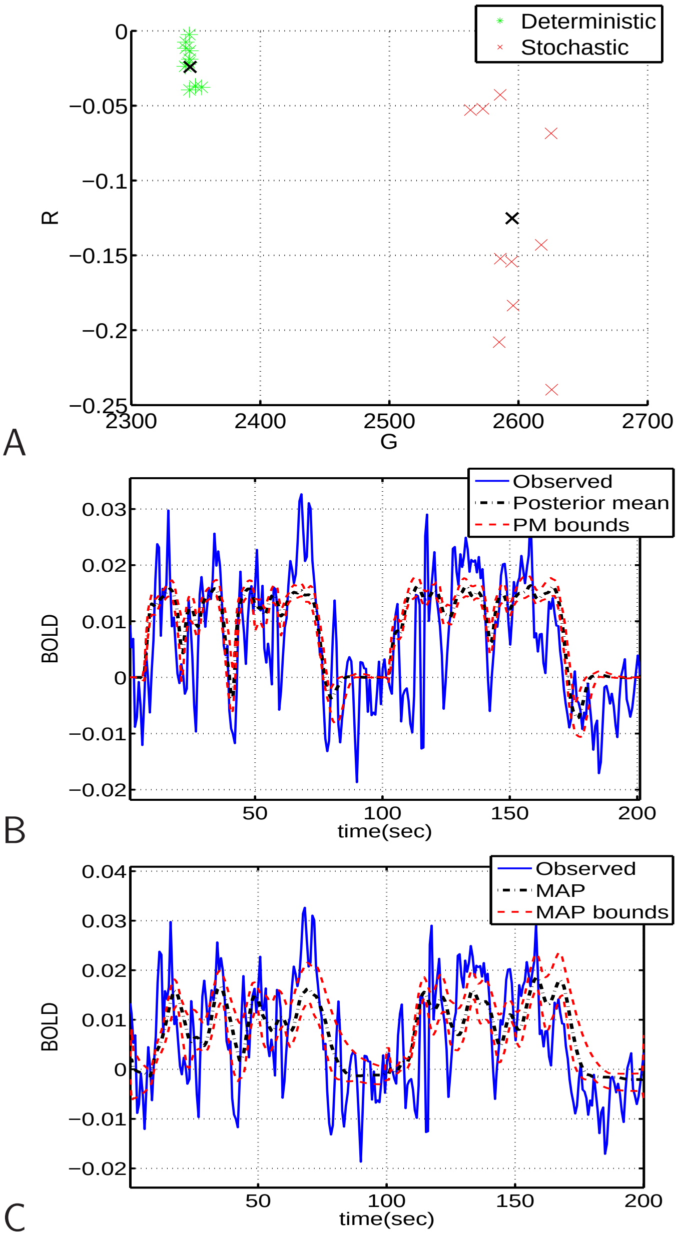


Figure 7.21: Deterministic and stochastic model results for data set 2. A: Comparison of generalization (G) and reproducibility (R) of the deterministic state space and the stochastic state space models. B: Prediction of the first two epochs of two test epochs by the deterministic model. C: Prediction of the first two epochs of two test epochs by the stochastic model.

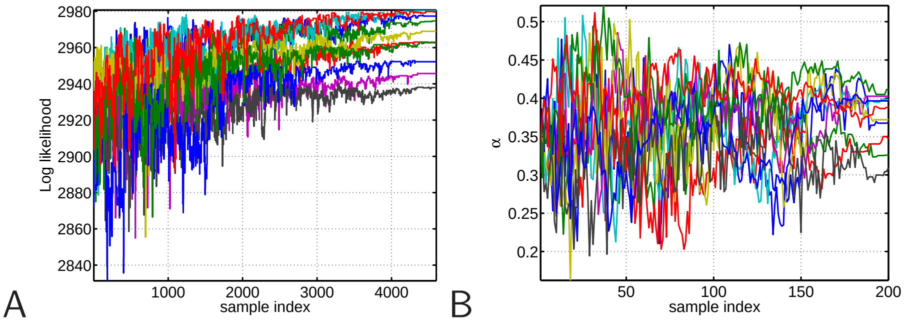


Figure 7.22: A: Simulated annealing likelihood increasing as samples are generated. As the temperature decreases, fewer of the random proposals are accepted. Synthetic data generated by the stochastic balloon model (noisy hidden states). 10 runs are shown based on as many different splits with 5 epochs in each training set. B: Corresponding convergence of the α parameter (true value is 0.4); only 200 evenly spaced samples of the original 4600 samples are shown here for clarity).

Chapter 8

Conclusion

”After all is said and done, more is said than done.” - Aesop

In this Ph.D. work methods have been developed that are able to successfully learn the parameters of non-linear models for fMRI, both in a classical formulation, and in a stochastic state space formulation. A framework has also been developed for comparing models in terms of their ability to generalize, and also in terms of reproducibility. The latter is measured as the robustness of the learned parameters to changes in training data. Comparison of models can then be done in a principled manner, by considering both these measures together. It is important to apply such comparison methods to current and new models to determine their relative merit for use in different contexts.

This framework for evaluating and comparing the quality of the models was used on real data to reveal significant differences in the suitability of the different models for different data. In particular it was found that for the block-design data set used here, the standard balloon model was well suited and outperformed other models. For the data set created by rapid, random, event-related stimuli, however, it was shown that the complexity of the stochastic state space model is a worthwhile addition. The model comparison framework was further verified by results using synthetic data sets, showing that it is possible to correctly distinguish the performance of the models.

It must be noted that these conclusions are based on the 'philosophical' choice of using fixed, physiologically based prior distributions for the parameters. If the priors instead were represented by parameterized functions (using 'hyper-parameters') that were fit to the data, it is possible that different conclusions would be reached.

The research in BOLD fMRI is a very active field, and a great deal of work has been done to investigate the relationships between stimuli and neural activity, between neural activity and the hemodynamic response, and the dependence of the BOLD signal on different types of stimuli (see e.g. [56] and [70]).

However, the widespread adaptation of non-linear models of the type described in this thesis has not yet quite happened. It may be hoped that methods for meaningful estimation of model quality, such as those developed in this work, will further the application of these model families.

8.1 Future research directions

The experiments have shown that the assumptions of stationarity across time of all model parameters may not generally hold. Both the phase and amplitude of the BOLD signals seem to vary slightly across epochs. Although some of this could be attributed to artifacts not completely removed during preprocessing, there seems to be room for improvement of the models. This can either be done by adding specific model components that can explain such variation, or by considering whether some parameters, such as $\epsilon(t)$, are better thought of as time-varying variables. Related models could also be subjected to similar analysis, such as [46] and [83], and compared to the models investigated here.

On the learning front, there are alternatives to MCMC approximations, for example variational Bayes (see e.g. [26], [77]); see [21] for an exciting combination of variational Bayes and MCMC. There are also many possible refinements of MCMC (other than parallel tempering), e.g. [60], [59]. An obvious research area is the comparison of various learning methods for these models.

These models may be inverted to produce estimates of neural activity as indicated in the work of Riera et al. [68]. This task is often referred to as *deconvolution*, because the BOLD signal is seen as a convolution of the neural activity signal with a (linear) hemodynamic response function (see e.g. [29]). In [68] a regularized radial basis function set is used, with parameters estimated using a likelihood based approach which leads to rather smooth activation estimates. Using our Bayesian sampling approach from an augmented posterior distribution including parameters of the neural activity time course (such as stimulus onset times etc.) may be a way to let data determine the level of regularization, hence, potentially lead to more crisp estimates of non-trivial neural activation sequences. This would be of particular interest in more complex activation designs involving different stimulus activation conditions within epochs.

Deconvolution would also allow whole-brain estimation of model parameters, since there would be no requirement to know the neural activity in advance. The properties of the parameters across the brain (individual, spatial variances), and across subjects, could then be investigated. Deconvolution is a very important application of these models, and it is hoped that more work will be done to attempt deconvolution with hemodynamic models.

A major limitation of the current models is found in the spatial dimension. The common assumption is that the region of interest is in essence one big vascular 'balloon', and neither internal or external spatial interactions are considered. With voxels getting smaller as scanners improve, the assumption of no spatial interaction does not hold. It is well known that activation in one location will affect the BOLD signal in surrounding area because of capillary recruitment etc. Therefore, an obvious improvement to the hemodynamic models is their extension to take spatial interactions into account. On the level of regions (several voxels), dynamic causal models ([24]) offer an interesting spatial extension of non-linear hemodynamic models. On the voxel level, an interesting approach was presented at the Human Brain Mapping conference in 2006 (HBM2006) [38]; this is a continous-space continous-time physical model that is still in development. A Bayesian discrete temporal-spatial approach is given in [82]. A continuous-space, continuous time formulation of a hemodynamic model is a natural next step for this model family.

Appendix A

Publication: ISBI 2006

This appendix contains the article *Identification of non-linear models of neural activity in BOLD fMRI.*, in *3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano, 2006*, pages: 952-955. Author list: Daniel J. Jacobsen, Kristoffer Hougaard Madsen and Lars Kai Hansen. Own contribution approximately 90%.

IDENTIFICATION OF NON-LINEAR MODELS OF NEURAL ACTIVITY IN BOLD FMRI

Daniel J. Jacobsen, Kristoffer Hougaard Madsen, Lars Kai Hansen

Intelligent Signal Processing
Technical University of Denmark

ABSTRACT

Non-linear hemodynamic models express the BOLD signal as a nonlinear, parametric functional of the temporal sequence of local neural activity. Several models have been proposed for this neural activity. We identify one such parametric model by estimating the distribution of its parameters. These distributions are themselves stochastic, therefore we estimate their variance by epoch based leave-one-out cross validation, using a Metropolis-Hastings algorithm for sampling of the posterior parameter distribution.

1. INTRODUCTION

Neuroimaging has made major contributions to our understanding of the relation between behavior and distribution of brain information processing. The richest neuroimaging modality is fMRI based on the so-called BOLD effect, involving the hemodynamic response which is rather sluggish, non-linear, and non-local. With the long term goal of increasing the spatio-temporal resolution of BOLD fMRI, we are interested in models linking subject behavior, neural activity, the so-called hemodynamic response, and fMRI BOLD observations, see e.g., [1, 2, 3, 4].

We will examine the model proposed by Friston et al. [2] which consists of a set of ordinary differential equations (ODE's) that model the evolution in time of four basic physiological state variables: The blood volume $v(t)$, blood inflow $f(t)$, amount of de-oxyhemoglobine $q(t)$ and a so-called 'flow inducing signal' $s(t)$, collected in the state vector, $\mathbf{x}(t) = [v(t)\ q(t)\ f(t)\ s(t)]^T$. The flow inducing signal is driven by an underlying neural activation function $\nu(t)$ - a time function describing the local neural activity.

The measured BOLD signal y_n is then modeled as a nonlinear function of 'snapshots' of the continuously evolving states, with additive white Gaussian noise w_n; subscript indices are used for these variables to emphasize the discrete 'sampled' nature.

$$\begin{aligned} \frac{\partial \mathbf{x}}{\partial t} &= f(\mathbf{x}(t), \nu(t)) \\ y_n &= g(\mathbf{x}(t_n)) + w_n \end{aligned} \tag{1}$$

The BOLD signal is measured with a sampling interval denoted *TR*. The model has seven parameters: σ_w^2, the variance of w_n, and six physiological parameters[1], combined in $\theta = \left[\alpha\ \epsilon\ \tau_0\ \tau_s\ \tau_f\ E_0\ \sigma_w^2\right]^T$. E_0 is the so-called 'resting net oxygen extraction fraction in the capillary bed'.

In addition, we assume that the states evolve from an initial known resting state $\mathbf{x}_0 = [0\ 1\ 1\ 1]^T$. The latter assumption is reasonable if a suitably long resting period precedes stimulation sequences. The dynamical model thus consists of the set of non-linear differential equations

$$\frac{\partial v(t)}{\partial t} = \frac{1}{\tau_0}\left(f(t) - v(t)^{1/\alpha}\right)$$

$$\frac{\partial q(t)}{\partial t} = \frac{1}{\tau_0}\left[f(t)\frac{1-(1-E_0)^{1/f(t)}}{E_0} - v(t)^{(1-\alpha)/\alpha} q_t\right]$$

$$\frac{\partial s(t)}{\partial t} = \epsilon\nu(t) - s(t)/\tau_s - (f(t)-1)/\tau_f$$

$$\frac{\partial f(t)}{\partial t} = s(t)$$

Finally, the BOLD observation model involves the nonlinearity,

$$\begin{aligned} g(\mathbf{x}(t)) = V_0[&(k_1+k_2)(1-q(t)) \\ &-(k_2+k_3(1-v(t)))] \end{aligned} \tag{2}$$

with a set of empirical constants taking values $V_0 = 0.02$, $v_0 = 40.3, TE = 0.03, r_0 = 25, \epsilon = 1.43$, $k_1 = 4.3 E_0 v_0 TE, k_2 = E_0 \epsilon r_0 TE, k_3 = \epsilon - 1$ for BOLD imaging at $1.5T$ [5]. The BOLD fMRI measurements are spatially sampled in volume elements (voxels). Experiments are typically ivided temporally in quasi-independent baseline-activation stimulus 'epochs'.

2. STATISTICAL MODELING

For given hemodynamic parameters and neural activity, the likelihood of an epoch is straightforward to set up. First, the hidden states will evolve deterministically according to (1) driven by the given neural activity $\nu(t)$. We use a variable step-size 4th/5th-order embedded Runge-Kutta method to solve these [6], with the starting condition $\mathbf{x}(t = 0) =$

[1] we assume the resting blood volume fraction, V_0, to be constant at 0.02 [2]

$\mathbf{x}_0$, the initial (relaxed) state, $\mathbf{x}_0 = [1\ 1\ 1\ 0]^T$ (all values relative to resting state). This gives a sequence of states, $\mathbf{x}_{1:N} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, corresponding to the sampling times $\{t_1, t_1 + \mathit{TR}, \ldots, t_1 + N \cdot \mathit{TR}\}$, where t_1 is the starting time of the epoch. The mean BOLD signal is given by

$$y_n = g(\mathbf{x}_n; \theta)$$

with the observed output modelled as

$$y_n^* = y_n + w_n(\theta) \tag{3}$$

As the residuals are assumed normal i.i.d., $y_n^* \sim \mathcal{N}(y_n, \sigma_w^2)$, the likelihood becomes

$$p(y_{1:N}^*|\theta) = \prod_{n=0}^{N} p(y_n^*|\theta) = \prod_{n=0}^{N} \mathcal{N}_{y_n^*}(y_n, \sigma_w^2).$$

The neural activity function $\nu(t)$ is traditionally assumed to be a simple square wave signal representing signal 'on' during stimulus and signal 'off' during baseline [2]. Buxton et al. [3] have recently introduced an alternative dynamical model representing neural activity which we will use here. This model posits

$$\begin{aligned} \nu(t) &= a(t) - I(t) \\ \frac{dI}{dt} &= \frac{\kappa\nu(t) - I(t)}{\tau_1}, \end{aligned} \tag{4}$$

where $a(t)$ is the square wave stimulus reference function and $I(t)$ is an inhibitory feedback signal. The values of the parameters are unknown a priori, although in [3], ranges are given as $\tau_1 \in [1; 3]$, $\kappa \in [0; 3]$. We refer to them jointly as $\phi = [\kappa, \tau_i]$.

Note that the square wave model obtains as a special case of this non-linear model as $\kappa/\tau_i \to 0$.

In this report we will apply this model to real fMRI data and investigate the posterior distribution of κ and τ_i. By way of leave-one-out cross-validation, the uncertainty on these distributions will further be described.

3. GENERALIZATION AND ESTIMATION PROCEDURES

The target for this investigation is the *posterior* distribution of the parameters of the non-linear neural model, ϕ:

$$p(\phi|D) = \frac{p(D|\phi)p(\phi)}{p(D)} = \frac{p(D|\phi)p(\phi)}{\int p(D|\phi)p(\phi)d\phi}$$

i.e. the distribution of the parameters conditioned on the observed BOLD data, D. This distribution cannot be obtained analytically; instead, we employ a Metropolis-Hastings Markov-chain Monte Carlo (MCMC) method to generate samples from the posterior.

3.1. Markov-chain Monte Carlo sampling

We use a Metropolis-Hastings (MH) algorithm [7] starting at an arbitrary state, ϕ_0, and at each step proposing small changes in ϕ from a *proposal distribution*, in our case a Gaussian centered on the 'current' state $p(\phi_{n+1}|\phi_n) = \mathcal{N}(\phi_n; \Sigma)$. The parameters of the hemodynamic model are sampled simultaneously, but as they are not the focus of this report, they are not discussed further and ignored for clarity (also the approximate marginal distribution of κ and τ_i is just the sampled values of these, so no further work is required to obtain the desired distribution).

The MH method produces samples from the true posterior distribution in the limit of large number of samples. We employ a set of heuristics to ensure convergence before averaging, e.g., inspection of saturation of the training set likelihood and stabilization of the actual parameter values.

A good proposal distribution is major determinant for succes of the Metropolis-Hastings algorithm. Again we invoke heuristics: Starting with a spherical normal distribution of dimension $\dim \phi$, we perform several (short) scout sampling runs. After each of these, the covariance of the generated samples is used to adapt the covariance of the proposal, Σ, scaled to give an acceptance rate of around 0.3. This procedure greatly speeds up the final sampling run (the samples of the initial runs are not used in averaging).

For most of the parameters, we use simple uniform priors ($p(\phi)$) over positive parameters; for the oxygen extraction fraction E_0 we use a uniform distribution over the interval $[0, 1]$.

These samples ($\phi_n,\ n = 1..N$) can then be used as an empirical approximation of the target distribution, e.g. for prediction of a BOLD signal y:

$$\begin{aligned} p(y|D) &= \int p(y|\phi)p(\phi|D)d\phi \\ &\approx \frac{1}{N}\sum_{n=1}^{N} p(y|\phi_n), \quad \phi_n \sim p(\phi|D) \end{aligned}$$

The resulting approximate distribution clearly depends on the particular data set D. To obtain information on the uncertainty of the posterior, we employ epoch-wise leave-one-out cross-validation. With K quasi-independent epochs available, each split of the data leaves out one epoch for a test set, which can be used to validate the ability of the model to prediction test data, while $K - 1$ are used for a training set to give an estimate of the posterior distribution.

To get an estimate of the uncertainty of the posterior distribution approximations, we fit a normal distribution to the samples for each cross-validation split. This yields a distribution of means and variances that can be used to illustrate the variance of the distribution. Each split yields an unbiased estimate of the true mean and variance of the distribution, and the mean over all splits is a convex combination thus also an unbiased estimate, but with a reduced variance.

4. EXPERIMENTAL EVALUATION

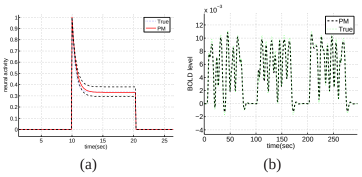
The described method was evaluated on a synthetically generated data set and on a real BOLD fMRI data set.

4.1. Synthetic data

The synthetic data was obtained by simulating the hemodynamic model with parameters set to the maximum likelihood values reported by Friston et al. [2],
$\theta = \left[0.330.540.981.542.460.34\ \sigma_w^2\right]^T$,
with σ_w^2 set to produce a desired SNR (signal-to-noise ratio, measured as the ratio of the de-noised BOLD and observation noise signals) close to 5.0 dB, which is similar to real recording conditions. κ and τ_i were set to 2.0 and 1.6 respectively. The model is initialized in $\mathbf{x}_0$, the states are evolved using a Runge-Kutta solver, and observations are made, adding Gaussian white noise with the prescribed variance. Each epoch contains 100 samples with sampling interval $TR = 1.0s$.

To justify our assumption that the BOLD signal is independent between epochs, the stimulus for each epoch is set to zero for the last 30 seconds. This is helpful for two further reasons. Preprocessing (e.g. removing low-frequency noise), is aided in that such artifacts can be more accurately estimated using these 'resting' portions of data. Finally, it allows us to assume a known, resting, physiological state ($\mathbf{x}_0$) at the start of each epoch. The stimulus is designed to evoke non-linear behavior in the model; this is achieved by inter-stimulus intervals (ISI) and stimulus durations (SD) being sampled from a suitable gamma distribution.

Figure 1 shows parts of the reconstructed neural and BOLD signals. The posterior mean of κ and τ was 1.92 ± 0.51 and 1.32 ± 0.22 respectively, resulting in very close reconstructions.



(a) (b)

Fig. 1. (a): One pulse of the reconstructed and true neural activity. The shape of the signal is retrieved, although the slope is slightly off, corresponding to a bias in the τ_i estimate. (b): Reconstructed and true BOLD signal (first 3 epochs).

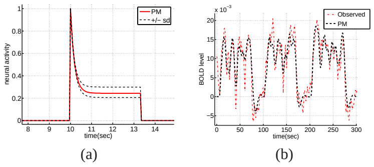
4.2. Real data

The data was acquired at Hvidovre Hospital, Denmark, using a 3T scanner (Magnetom Trio, Siemens). We obtained 1382 GRE EPI volumes each consisting of twelve 3mm slices oriented along the calcarine sulcus. Additional parameters where TR=725 ms, TE=30 ms FOV=192 mm, 64x64 acquisition matrix, FA = 82. The stimulus consisted in a circular black/white flickering checkerboard (24 degrees horizontal, 18 degrees vertical) on a grey background. The checkers reversed black/white at 8 Hz. The activation pattern ($a(t)$) used to determine on- and offset of this stimulus was the same as was used to generate the synthetic data.

Fifty significantly activated (as determined by SPM2 analysis [2] [8]) voxels in visual cortex were selected, and the mean of these was used as the BOLD signal.

The results are shown in figure 2. The resulting shape of the neural model is similar to the one found for the synthetic data, and is significantly different from square. The BOLD reconstruction (see (3.1)) on test data is satisfactory, validating the model and method. We found posterior mean values of 3.11 ± 0.76 for κ and 0.87 ± 0.19 (± one standard deviation) for τ_i; again, not close to a square pulse.

Figure 3 shows the normal approximations to the posterior histograms. These indicate that although there is significant variation, the mean of all the posterior means obtained from the leave-one-out cross-validation is identifiable. We expect with longer sampling runs to bring down this variability - the important point is that this is a tool that gives guidance on the reliability of the estimated distribution.



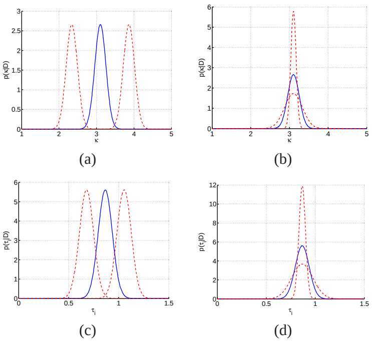
(a) (b)

Fig. 2. (a): One pulse of the estimated neural activity. The posterior mean reconstruction is shown together with the reconstruction corresponding to ± one standard deviation of the distribution of the posterior mean. (b): Prediction on real test data epochs (first 3 epochs).

5. CONCLUSION

The method detailed here can be used obtain to obtain approximate posterior distributions of model parameters together with estimates on the reliability of these approximations. For the

[2] Software available from http://www.fil.ion.ucl.ac.uk/spm/



(a)

(b)

(c)

(d)

Fig. 3. (a): Normal approximations of posterior parameter histograms. (a) and (b) show the variation of the mean and variance respectively, for the posterior distribution of κ. (c) and (d) show the same for τ_i (real data).

parameters of the non-linear neural activity model, we found posterior mean values of 3.11 ± 0.76 for κ and 0.87 ± 0.19 ($\pm$ one standard deviation) for τ_i. Both of these are on the edge of the ranges given in [3], although not statistically significantly so. There is little other information available on statistical estimation of these parameters.

We find that both κ and τ are identifiable for real BOLD data, and that κ/τ_i for our data is significantly greater than zero. Thus the square model of neural activity which is widely used in BOLD analysis is not supported by our findings.

The present model may be inverted to produce estimates of neural activity as indicated in the work of Riera et al. [4]. In [4] a regularized radial basis function set is used, with parameters estimated using a likelihood based approach which leads to rather smooth activation estimates. Using our Bayesian sampling approach from an augmented posterior distribution including parameters of the neural activity time course (such as stimulus onset times etc.) may be a way to let data determine the level of regularization, hence, potentially lead to more crisp estimates of non-trivial neural activation sequences. This would be of particular interest in more complex activation designs involving different stimulus activation conditions within epochs. In the present model we have focused on the local hemodynamics in average data from a region. The BOLD hemodynamics is non-local and it is an important future task to produce a spatio-temporal hemodynamic model, which could also lead to improved spatial resolution.

6. REFERENCES

[1] R. B. Buxton, E. C. Wong, and L. R. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: the balloon model," *MRM*, vol. 39, pp. 855–864, June 1998.

[2] 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.

[3] R. B. Buxton, K. Uludag, D. J. Dubowitz, and T. T. Liu, "Modeling the hemodynamic response to brain activation," *NeuroImage*, vol. 23, pp. 220–33, 2004.

[4] J.J. Riera, J. Watanabe, K. Iwata, N. Miura, 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.

[5] 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.

[6] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*, New York: Dover, 1972.

[7] D.J.C. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003.

[8] K.J. Friston, *Introduction: Experimental design and statistical parametric mapping*, Academic Press, second edition, 2003.

Appendix B

Publication: Neural Computation I

This appendix contains the article *Bayesian model comparison in non-linear BOLD fMRI hemodynamics*, submitted to Neural Computation, July 2006. Author list: Daniel J. Jacobsen, Kristoffer Hougaard Madsen and Lars Kai Hansen. Own contribution approximately 90%.

Bayesian model comparison in non-linear BOLD fMRI hemodynamics

Daniel J. Jacobsen, Lars Kai Hansen
and Kristoffer Hougaard Madsen[1]
Intelligent Signal Processing
Informatics and Mathematical Modelling
Technical University of Denmark
[1]Also with the Danish Research Center for Magnetic Resonance
Copenhagen University Hospital, Hvidovre, Denmark
dj,lkh,khm@imm.dtu.dk

December 1, 2006

Abstract

Non-linear hemodynamic models express the BOLD signal as a nonlinear, parametric functional of the temporal sequence of local neural activity. Several models have been proposed both for the neural activity and the hemodynamics. We compare two such combined models: the 'original' balloon model with a square-pulse neural model [5], and an extended balloon model with a more sophisticated neural model [3]. We learn the parameters of both models using a Bayesian approach, where the distribution of the parameters conditioned on the data is estimated using Markov chain Monte Carlo techniques. Using a split-half resampling procedure [14], we compare the generalization abilities of the models as well as their reproducibility, both for synthetic and real data, recorded from two different visual stimulation paradigms. The results show that the simple model is the best one for these data.

1 Introduction

Neuroimaging has made major contributions to our understanding of the relation between behavior and distribution of brain information processing. The richest neuroimaging modality is fMRI based on the BOLD (Blood Oxygenation Level Dependent) effect, involving the so-called hemodynamic response which is rather sluggish and non-linear. With the long term goal of increasing the spatio-temporal resolution of BOLD fMRI, we are interested in models linking subject behavior, neural activity, the hemodynamic response, and fMRI BOLD observations, see e.g. [4, 5, 3, 13].

Non-linear hemodynamic models express the BOLD signal as a nonlinear, parametric functional of the temporal sequence of local neural activity. Increased neural activity increases local cerebral blood flow (CBF) and metabolic rate of oxygen consumption ($CMRO_2$), and these affect the level of deoxyhemoglobin and the blood volume (CBV), giving rise to the BOLD signal.

The physiological relationship between neural activity and the BOLD signal is unclear, and several variants for the components of these models have been proposed. Two such combined models - which we simply label 'A' and 'B' - are investigated here.

The purpose of this work is to compare these models in a probabilistic manner. This is important as it links model design - based on physiological and mathematical assumptions - to the actual data (the 'real world'). This furthers progression towards better models and thus increased knowledge about the brain, in particular the BOLD fMRI modality.

2 Model A - the static hemodynamic model

By 'static hemodynamic model', we refer to a model combining the simplest model of neural activity with the 'original' hemodynamic model developed in [3] and [5].

The hemodynamic part of this model consists of a set of ordinary differential equations (ODE's) modelling the evolution in time of four basic physiological variables: The blood volume $v(t)$, blood inflow $f(t)$, amount of de-oxyhemoglobine $q(t)$ and a so-called 'flow inducing signal' $s(t)$, collected in the state vector, $\mathbf{x}(t) = [v(t)\ q(t)\ f(t)\ s(t)]^T$. The flow inducing signal is driven by an neural activation function $u(t)$ - a time function describing local neural activity. This in turn drives changes in the other state variables through the ODE's:

$$\frac{\partial \mathbf{x}}{\partial t} = f(\mathbf{x}(t), u(t)) \tag{1}$$

The measured BOLD signal y_n is then a non-linear function of 'snapshots' of the continuous states, with additive white Gaussian noise w_n; subscript indices are used for these variables to emphasize their discrete nature.

$$y_n = g(\mathbf{x}(t_n)) + w_n \tag{2}$$

The BOLD signal is measured with a sampling interval denoted TR. The model has seven parameters: σ_w^2, the variance of w_n, and six physiological parameters, combined in $\theta_A = [\alpha\ \epsilon\ \tau_0\ \tau_s\ \tau_f\ E_0\ \sigma_w^2]^T$.

We assume that the states evolve from an initial known resting state $\mathbf{x}_0 = [1\ 1\ 1\ 0]^T$ - volume, deoxyhemoglobin and flow at resting levels, stimulus at zero. This is reasonable if a suitable resting period precedes stimulation.

The specific differential equations are

$$\frac{\partial v(t)}{\partial t} = \frac{1}{\tau_0}\left(f(t) - f_{out}(t)\right) \tag{3}$$

$$\frac{\partial q(t)}{\partial t} = \frac{1}{\tau_0}\left[f(t)\frac{1-(1-E_0)^{1/f(t)}}{E_0} - v(t)^{(1-\alpha)/\alpha} q_t\right] \tag{4}$$

$$\frac{\partial s(t)}{\partial t} = \epsilon u(t) - s(t)/\tau_s - (f(t)-1)/\tau_f \tag{5}$$

$$\frac{\partial f(t)}{\partial t} = s(t) \tag{6}$$

The blood outflow $f_{out}(t)$ follows

$$f_{out}(t) = v(t)^{1/\alpha} \tag{7}$$

hence the 'static' label (see model B below).

The BOLD observation model is

$$\begin{aligned} g(\mathbf{x}(t)) = V_0[&(k_1 + k_2)(1 - q(t)) \\ &- (k_2 + k_3(1 - v(t)))] \end{aligned} \tag{8}$$

with a set of empirical constants taking values $V_0 = 0.02$, the rest depending on scanner type and settings [2].

The measurements are spatially sampled in volume elements (voxels) and divided temporally in quasi-independent baseline-activation stimulus 'epochs'.

The neural activity in this model is identical with the stimulus given to the subject, a train of square pulses. This assumption is common in BOLD fMRI analysis (e.g. [5]). The stimulus is set to 1.0 during 'on' periods and 0.0 during 'off' (no stimulus):

$$u(t) = a(t) = \begin{cases} 1.0 & \text{stimulus on at time t} \\ 0.0 & \text{stimulus off at time t} \end{cases} \tag{9}$$

3 Model B - a viscoelastic hemodynamic model with adaptation in neural activity.

Buxton et al. [3] have recently introduced an alternative dynamical model representing neural activity, positing

$$\begin{aligned} u(t) &= a(t) - I(t) \\ \frac{dI}{dt} &= \frac{\kappa u(t) - I(t)}{\tau_u}, \end{aligned} \tag{10}$$

where $a(t)$ is the square wave stimulus function and $I(t)$ is an inhibitory feedback signal. κ is a gain factor for the inhibition signal, and τ_u is a time constant that determines how quickly the neural activity is inhibited. This leads to an adaptive neural response to sustained stimuli. Note that the square pulse model obtains as a special case of this non-linear model as $\kappa/\tau_u \to 0$.

Further, it was proposed in [3] that the relation between outflow and volume in model A (7) is based on steady-state conditions and should be modified for dynamic conditions. The proposed relation is

$$f_{out}(v(t)) = v(t)^{1/\alpha} + \tau \frac{\delta v(t)}{\delta t} \tag{11}$$

which means that the 'balloon' will transiently resist changes (for example due to visco-elastic effects, hence the model label) and only after some time (controlled by τ) conform to the steady-state relationship (7) (this requires adding $f_{out}(t)$ as a fifth state variable with its own ODE, details omitted). Also, τ is proposed to be different during inflation and deflation:

$$\tau = \begin{cases} \tau_+ & f(t) \geq f_{out}(t) \\ \tau_- & f(t) < f_{out}(t) \end{cases} \tag{12}$$

This model is somewhat more complex, with 4 additional parameters (κ, τ_u, τ_+ and τ_-), as well as an extra ODE dimension. The initial resting state is extended to $\mathbf{x}_0 = [1\ 1\ 1\ 0\ 1]^T$ (blood outflow at resting level).

4 Model comparison

The purpose of comparing models can be phrased in terms of two crucial questions:

- *which model provides the best generalization ability?*
- *which model provides the highest reproducibility?*

Together, these two questions form a sound basis for comparing the models.

4.1 Generalization

Generalization ability is well known as fundamental goal of learning (see e.g. [10]). A model should be able to learn based on one data set ('training'), and generalize to another (test) data set. In the predictive learning framework, this means that the distribution of the parameters of a model learned from one data set - the so-called posterior distribution, $p(\theta|D)$ - should be able to 'explain' an independent test set D^* according to

$$p(D^*|D, M) = \int p(\theta|D, M)p(D^*|\theta, M)d\theta. \tag{13}$$

where M is the model. To compare models, we then compare this predictive density measured at one or more test data sets.

The corresponding generalization for a given test and training data set is defined as the logarithm of the predictive distribution,

$$G(D^*, D, M) \triangleq \log p(D^*|D, M).$$

The mean predictive generalization over test sets is

$$G(D, M) \triangleq \langle G(D^*, D, M)\rangle_{p(D^*)} = \int [\log p(D^*|D, M)]p(D^*)dD^*.$$

Apart from an additive constant (the entropy of the true distribution), the mean generization is equal to minus the Kullback-Leibler distance between the 'true' distribution of the data, $p(D^*)$, and the model distribution, $p(D^*|M)$. To obtain the overall generalization we furthermore average over training sets

$$G(M) \triangleq \langle G(D,M) \rangle_{p(D)} = \left\langle \langle G(D^*, D, M) \rangle_{p(D^*)} \right\rangle_{p(D)} =$$
$$= \int \left[\int \log p(D^*|D, M) p(D^*) dD^* \right] p(D) dD$$

In applications we can not compute the integrals, hence, we estimate the model generalization using independent test- and training data sets by split-half resampling, see e.g. [14],

$$G(M) \approx \frac{1}{K} \sum_{i=1}^{K} \log p(D_i^*|D_i, M) = \frac{1}{K} \sum_{i=1}^{K} G(D_i^*, D_i, M). \tag{14}$$

With K quasi-independent epochs available, each split of the data leaves uses $K/2$ each for the test and training sets. Each of the resulting estimates is an unbiased estimate of the model generalization. The mean over all splits is a convex combination thus also an unbiased estimate, but with a reduced variance.

4.2 Reproducibility

Reproducibility concerns the sensitivity of what is learned to the particular data set used for training. This is highly relevant in model comparison, because a model that generalizes well with parameters that vary greatly depending on the particular training data set might be less attractive than a model with a slightly lower ability to generalize, but that produces more robust posterior parameter distributions. This is particularly so in the case of 'physiological' models where the parameters carry physiological meaning. The weight one assigns to generalization and reproducibility is therefore context dependent. For instance, if de-noising of the BOLD signal is the objective, generalization will be important; if one wants to understand the mechanisms underlying the BOLD signal, reproducibility could be more important.

In the current setting, it is the posterior distribution, $p(\theta|D)$ that is in question, and a measure is needed of the similarity of posteriors obtained conditioned on different data sets. We use a Kullback-Leibler measure to estimate reproducibility. Parallel with the estimation of generalization (14), the KL distance is measured between the two posterior parameter distributions of each split half, and the reproducibility is then the average over all splits

$$R(M) = -\frac{1}{K} \sum_{i=1}^{K} \int p(\theta|D_i^1, M) \log \frac{p(\theta|D_i^1, M)}{p(\theta|D_i^2, M)} d\theta \tag{15}$$

where D_i^1 and D_i^2 are the two data sets in split i. Note that we measure reproducibility as the *negative* of the KL distance.

Since this depends non-trivially on the dimensionality and shape of the distributions, we use the mean KL distance over dimensions, i.e. calculating the KL distance for each dimension and taking the mean.

5 MCMC approximation

The generalization (13) and reproducibility (15) measures both depend on an integral over the posterior. These integrals can not be solved analytically, due to the non-linearities involved. Instead, they are approximated using the Markov chain Monte Carlo principle. This is based on the law of large numbers ensuring that if samples z_i from some distribution $p(z)$ can be generated, then the approximation

$$\int f(z)p(z)dz \approx \frac{1}{K}\sum_{i=1}^{K} f(z_i) \tag{16}$$

will be increasingly accurate with K (compare with (13) and (15)) [1].

5.1 Metropolis-Hastings and parallel tempering

In order to generate samples from the $p(\theta|D)$, we employ the MCMC techniques of Metropolis-Hastings sampling and parallel tempering. Several excellent sources are available that describe these methods (see e.g. [11], [8]), and only an overview will be given here.

Metropolis-Hastings sampling works by starting at an arbitrary state,
$\theta(n = 0) = \theta_0$, and then iteratively proposing small changes through a *proposal distribution*, $p(\theta'|\theta(n))$. We use a Gaussian centered on the previous state:

$$p(\theta'|\theta(n)) = \mathcal{N}(\theta(n); \Sigma)$$

The proposal is accepted as the next discrete sample according to the ratio:

$$r = \frac{p(\theta'|D)}{p(\theta(n)|D)} \tag{17}$$

If $r \geq 1$ the proposal is accepted, and the next sample generated is $\theta(n+1) = \theta'$. If $r < 1$, the proposal is accepted with probability r. If the proposal is not accepted, the next sample is simply $\theta(n+1) = \theta(n)$.

The Metropolis-Hastings method produces samples from the true posterior distribution in the limit of large number of samples (under certain conditions, see [11]). Depending on the true target distribution, 'mixing', i.e. the ability of the algorithm to generate samples representative of the whole distribution may be slow. The technique of parallel tempering may be employed as a quite straightforward enhancement. It works by sampling from several distributions $p_i(\theta|D)$ in parallel, each using a more or less 'flattened' version of the likelihood:

$$p_i(\theta|D) \triangleq \frac{p(D|\theta)^{\beta_i} p(\theta)}{\int p(D|\theta)^{\beta_i} p(\theta) d\theta}, \quad i = 1 \ldots C \tag{18}$$

where $\beta_i \leqq 1$ are so-called *inverse temperatures* and C is the number of 'chains'. At certain intervals, a proposal is made to swap the states of two of these chains (using an acceptance ratio somewhat similar to (17), see [8]), generally leading to better mixing in the same amount of computer time. For our data, 6 chains from $\beta_1 = 1.0$ to $\beta_6 = 0.04$ was found to give good results.

[1] The KL calculation further requires a density estimation method; we use the kernel method described in [12]

A good proposal distribution is a major determinant for success of the algorithm. We have therefore implemented an automated procedure for finding a good proposal. Each sampling run is started with an arbitrary normal distribution of dimension $\dim \theta$, and several short 'scout' sampling runs are performed. After each of these, the covariance of the generated samples is used as the new proposal covariance, scaled to give an acceptance rate between 0.25 and 0.5. After a set number of these initial iterations, the main sampling run is performed keeping the proposal distribution constant (a condition for convergence; the samples of the initial runs are not used further). This procedure greatly increases mixing.

We employ a set of heuristics to obtain indications that the final sampling estimate of the posterior distribution does indeed cover the true distribution. For each data set, several runs are performed. If the mixing is insufficient, the resulting distributions can be non-overlapping, whereas for good mixing, they should be highly overlapping. For synthetically generated data, we can also verify that the distribution is sampled around the major peaks by comparing the samples of σ_w^2 with the known noise variance. We also confirm that the true parameter values are contained within the sampling distribution. We then assume that sampling with similar settings will yield a good approximation of the posterior distribution for real data, although we generate more samples for real data.

5.2 The likelihood

Bayes' rule allows us to rewrite the posterior in terms of the likelihood, $p(D|\theta)$, the prior, $p(\theta)$ and a normalizing factor, $p(D)$:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

In order to evaluate the terms in (17), we therefore need to evaluate the likelihood and the prior (note that the normalizing factor cancels out in (17)).

For given hemodynamic parameters and neural activity, the likelihood of an epoch is straightforward to set up. First, the hidden states will evolve deterministically according to (1) driven by the given neural activity $u(t)$. We use a variable step-size 4th/5th-order embedded Runge-Kutta method to solve these [1], with the starting condition $\mathbf{x}(t = 0) = \mathbf{x}_0$, the initial (relaxed) state (all values are relative to resting state). This gives a sequence of states, $\mathbf{x}_{1:N} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, corresponding to the sampling times $\{t_1, t_1 + TR, \ldots, t_1 + N \cdot TR\}$, where t_1 is the starting time of the epoch. The mean BOLD signal is given by

$$\overline{y}_n = g(\mathbf{x}_n; \theta) \tag{19}$$

with the observed output given by (2).

As the residuals are assumed normally i.i.d., $y_n \sim \mathcal{N}(\overline{y}_n, \sigma_w^2)$, the likelihood becomes

$$p(D|\theta) \triangleq p(y_{1:N}|\theta) = \prod_{n=0}^{N} p(y_n|\theta) = \prod_{n=0}^{N} \mathcal{N}_{y_n}(\overline{y}_n, \sigma_w^2). \tag{20}$$

Using MCMC, it is also possible to give empirical confidence intervals for the predicted mean (19). This simply requires calculating $\overline{y}_n$ for each sample $\theta(n)$, $n = 1 \ldots N$, sorting these values and picking the values with the proper index. For a $1 - \alpha$ confidence interval, the lower limit index is $N\alpha/2$ and the upper is $N(1 - \alpha/2)$ (rounding off).

5.3 Prior distributions

There are many approaches to choosing prior distributions. Generally it is important that the priors are as non-informative as possible, and yet they should reflect any prior beliefs we hold on the parameters. In the present case we actually have available prior physiological knowledge, and so the priors are built thereupon. We assume that the prior factorizes into a product of univariate priors, as there is little or no reason to believe - a priori - that the parameters are correlated.

The prior for the observation noise is simply set to a constant for positive values, $p(\sigma_w^2) = 0$ for $\sigma_w^2 \leq 0$. In practice it is consistently found for synthetic data that the observation noise is accurately estimated.

For all the other parameters, we use the family of scaled Beta distributions, as these are well suited to design appropriately flat distributions with upper and lower limits. The scaled beta distribution has three parameters s, u_1 and u_2 that control its range, mode and shape:

$$p(\theta|s, u_1, u_2) = \frac{1}{Z(s, u_1, u_2)}(s\theta)^{u_1-1}(1 - s\theta)^{u_2-1}$$

with

$$Z(u_1, u_2) = s\frac{\Gamma(u_1)\Gamma(u_2)}{\Gamma(u_1 + u_2)}$$

The distributions used here are based on the prior knowledge on each of the parameters as described in [5] and [3], see figure 1.

6 Experimental evaluation

Comparisons of the two models is done both for synthetically generated data and for two different real BOLD fMRI data sets.

For the real data, some preprocessing was done to remove artifacts (scanner and physiological). 9 resampling splits were used for the synthetic data, and 20 for the real data sets.

6.1 Synthetic data

The synthetic data was obtained by simulating the hemodynamic model with parameters set to $\theta = [0.40.52.02.52.50.4\ \sigma_w^2]^T$, with σ_w^2 set to produce a desired SNR (signal-to-noise ratio, measured as the ratio of the de-noised BOLD and observation noise signals) close to 5.0 dB, similar to real recording conditions. When generating data from model 'B', κ and τ_u were set to 2.0 and 1.0 respectively, and τ_+ and τ_- were both set to 15.0.

The model is initialized in $\mathbf{x}_0$, the states are evolved using a Runge-Kutta solver, and observations are made, adding Gaussian white noise with the prescribed variance. Each epoch contains 138 samples with sampling interval $TR = 0.725s$, exactly as for data set 2 (below). The stimulus signal is highly random, with many short stimuli that are often very close in time.

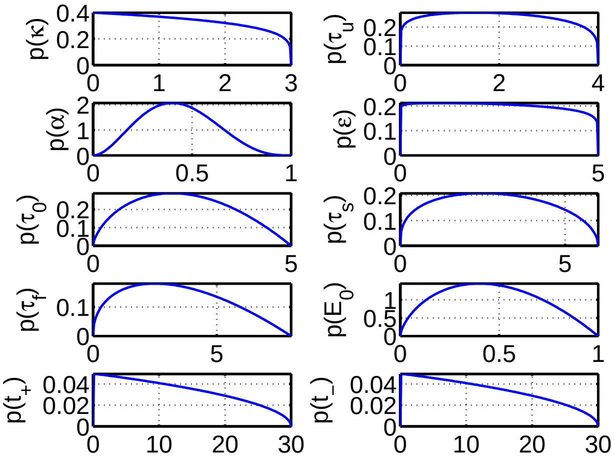


Figure 1: Prior distributions for the model parameters.

To justify our assumption that the BOLD signal is independent between epochs, the stimulus for each epoch is set to zero for the last 30 seconds. This is helpful for two further reasons. Preprocessing (e.g. removing low-frequency noise), is aided in that such artifacts can be more accurately estimated using these 'resting' portions of data. Finally, it allows us to assume a known, resting, physiological state ($\mathbf{x}_0$) at the start of each epoch. The stimulus is designed to evoke non-linear behavior in the model; this is achieved by inter-stimulus intervals (ISI) and stimulus durations (SD) being sampled from a suitable gamma distribution.

The models where compared for both synthetic data sets, i.e. those generated by model 'A' and by model 'B', and the results are shown in figure 2. When data are generated by the simpler model 'A', then - as might be expected - both the generalization ability and reproducibility are seen to be higher for the simple model, although reproducibility is not significantly higher (figure 2A). But when data are generated by model 'B', the situation is more complex: model 'B' is able to generalize significantly better, but model 'A' is more reproducible. This means that either model could be chosen as the 'best' one, depending on the intended use of the model.

6.2 Data Set 1

Data Set 1 was acquired by Dr. Egill Rostrup at Hvidovre Hospital, Denmark, on a 1.5 T Magnetom Vision scanner. The scanning sequence was a 2D gradient echo EPI (T2* weighted) with 66-ms TE, FA=50, FOV=230 mm, TR=330ms. Single slice data was obtained in a para-axial orientation parallel to the calcarine sulcus. The visual stimulus consisted of a rest period of 20s of darkness

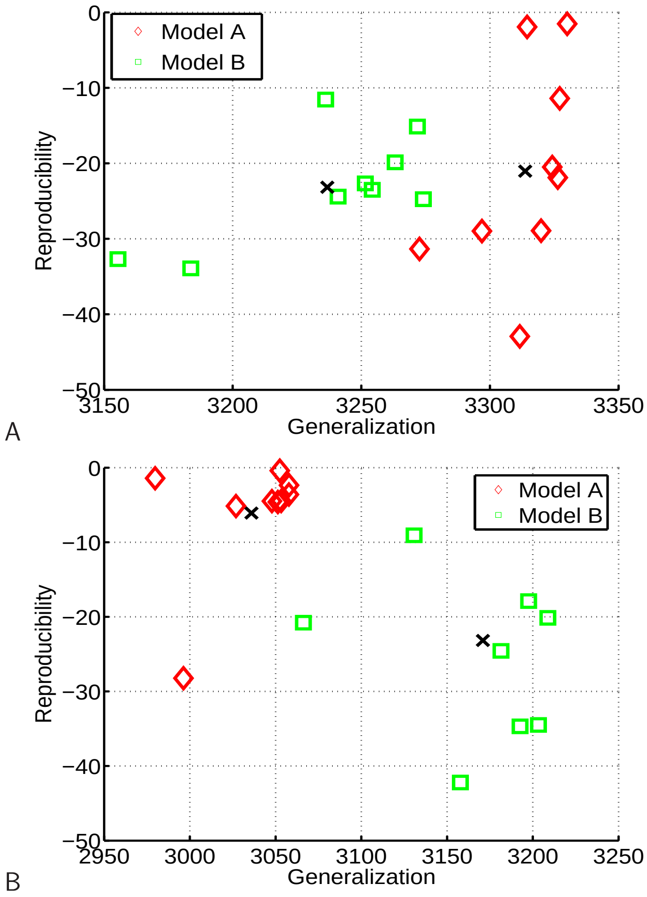


Figure 2: Generalization and reproducibility for synthetic data; crosses mark the mean. A: Data generated by model ’A’. B: Data generated by model ’B’.

(using a light fixation dot), followed by 10s of full-field checker board reversing at 8 Hz, and ending by 20s of darkness. Ten separate runs were completed, a total of 1000s recorded at each voxel. The data were preprocessed according to [7]. A ROI of 42 (7 by 6) significantly activated (as determined by SPM2 analysis[2] [6])) voxels from the visual cortex were selected and the mean of the signals was used (see figure 3A).

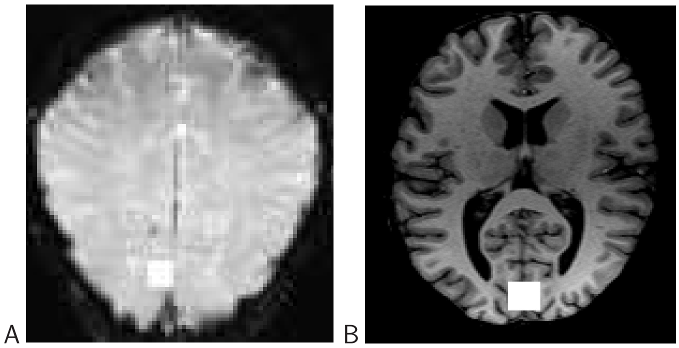


Figure 3: Regions of Interest, marked with white squares. A: Data set 1; T2* weighted image slice parallel to the calcarine sulcus. B: Data set 2; MPRAGE (Magnetization Prepared Rapid Gradient Echo) horizontal slice.

Figure 4 shows the posterior mean prediction together with 95% confidence interval on the predicted mean for this data set, for the first 2 epochs. Model 'B' clearly has more trouble explaining this data (although the first epoch is actually the worst case).

6.3 Data Set 2

The data was acquired at Hvidovre Hospital, Denmark, using a 3T scanner (Magnetom Trio, Siemens). We obtained 1382 GRE EPI volumes each consisting of twelve 3mm slices oriented along the calcarine sulcus. Additional parameters where TR=725 ms, TE=30 ms FOV=192 mm, 64x64 acquisition matrix, FA = 82. The stimulus consisted in a circular black/white flickering checkerboard (24 degrees horizontal, 18 degrees vertical) on a grey background. The checkers reversed black/white at 8 Hz. The activation pattern ($a(t)$) used to determine on- and offset of this stimulus was the same as was used to generate the synthetic data. A ROI of 75 (25 from each of 3 slices) significantly activated (as determined by SPM2 analysis, contiguous voxels in the visual cortex were selected, and the mean of these was used as the BOLD signal (see figure 3B).

Figure 5 shows the posterior mean prediction together with 95% confidence interval on the predicted mean for this data set (first epoch). Here it is difficult for the naked eye to spot a difference in predictive ability of the models.

[2] Software available from http://www.fil.ion.ucl.ac.uk/spm/

The model comparison results for data set 1 and 2 are shown in figure 6. As expected from the predictions (figure 4) data set 1 demonstrates model ’A’ as superior in both generalization and reproducibility. Figure 6B shows a similar relation for data set 2, although it is less convincing.

7 Discussion

This method can be used to compare models of BOLD fMRI data in a principled manner, estimating both the ability of the models to generalize and learn robustly.

The results indicate that the simple model is better than the more complicated one for the real data used here. It generalizes somewhat better for both data sets, and is significantly more reproducible. With more data, generalization and reproducibility would increase for both models, and a different comparison could result - but for the data sets used here, model ’A’ is best. It would be interesting to see if this holds for other types of data, such as across different parts of the brain and different stimuli. Alternative models not investigated here also exist (e.g. [9] and [15]); these would be interesting candidates to compare with the present models.

These models may be inverted to produce estimates of neural activity as indicated in the work of Riera et al. [13]. In [13] a regularized radial basis function set is used, with parameters estimated using a likelihood based approach which leads to rather smooth activation estimates. Using our Bayesian sampling approach from an augmented posterior distribution including parameters of the neural activity time course (such as stimulus onset times etc.) may be a way to let data determine the level of regularization, hence, potentially lead to more crisp estimates of non-trivial neural activation sequences. This would be of particular interest in more complex activation designs involving different stimulus activation conditions within epochs.

In the present model we have focused on the local hemodynamics in average data from a region. The BOLD hemodynamics is non-local and it is an important future task to produce a spatio-temporal hemodynamic model, which could also lead to improved spatial resolution.

References

[1] M. Abramowitz and I. A. Stegun (eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables, 9th printing*, Dover, 1972.

[2] P. A. Bandettini, R. M. Birn, D. Kelley, and Z. S. Saad, *Dynamic nonlinearities in bold contrast: neuronal or hemodynamic?*, 2002, pp. 73–85.

[3] R. B. Buxton, K. Uludag, D. J. Dubowitz, and T. T. Liu, *Modeling the hemodynamic response to brain activation*, NeuroImage **23** (2004), 220–33.

[4] R. B. Buxton, E. C. Wong, and L. R. Frank, *Dynamics of blood flow and oxygenation changes during brain activation: the balloon model*, MRM **39** (1998), 855–864.

[5] K. J. Friston, A. Mechelli, R. Turner, and C.J. Price, *Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics*, NeuroImage **12** (2000), 466–477.

[6] K.J. Friston, *Introduction: Experimental design and statistical parametric mapping*, second ed., Academic Press, 2003.

[7] C. Goutte, L. K. Hansen, M. G. Liptrot, and E. Rostrup, *Feature-space clustering for fmri meta-analysis*, Human Brain Mapping **13** (2001), 165 – 183.

[8] P. Gregory, *Bayesian logical data analysis for the physical sciences: a comparative approach with mathematica support*, Cambridge University Press, 2005.

[9] Y. Kong, Y. Zheng Y, D. Johnston, J. Martindale, M. Jones, S. Billings, and J. Mayhew, *A model of the dynamic relationship between blood flow and volume changes during brain activation*, J Cereb Blood Flow Metab **24** (2004), 1382–1392.

[10] Jan Larsen and Lars Kai Hansen, *Generalization: The hidden agenda of learning*, `http://www.citeseer.ist.psu.edu/larsen97generalization.html`.

[11] D.J.C. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003.

[12] Y. Moon, B. Rajagopalan, and U. Lall, *Estimation of mutual information using kernel density estimators.*, Physical Review E **52** (1995), 2318–2321.

[13] J.J. Riera, J. Watanabe, K. Iwata, N. Miura, E. Aubert, T. Ozaki, and R. Kawashima, *A state-space model of the hemodynamic approach: nonlinear filtering of bold signals*, Neuroimage **21** (2004), 547–567.

[14] S. C. Strother, J. Anderson, and L. K. Hansen, *The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework*, NeuroImage **15** (2002), no. 4, 747–771.

[15] Y. Zheng, J. Martindale, D. Johnston D, M. Jones M, J. Berwick J, and J. Mayhew, *A model of the hemodynamic response and oxygen delivery to brain*, Neuroimage **16** (2002), 617–37.

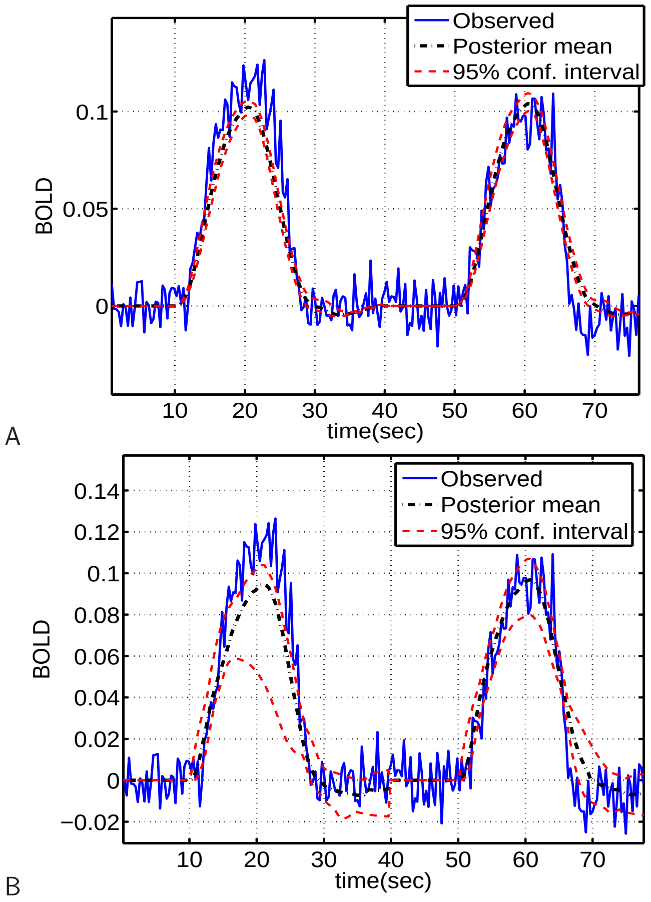


Figure 4: Prediction of data set 1. A: Model 'A'. B: Model 'B'. Note that the confidence interval is an empirical confidence interval for the mean prediction, based on the MCMC samples.

.

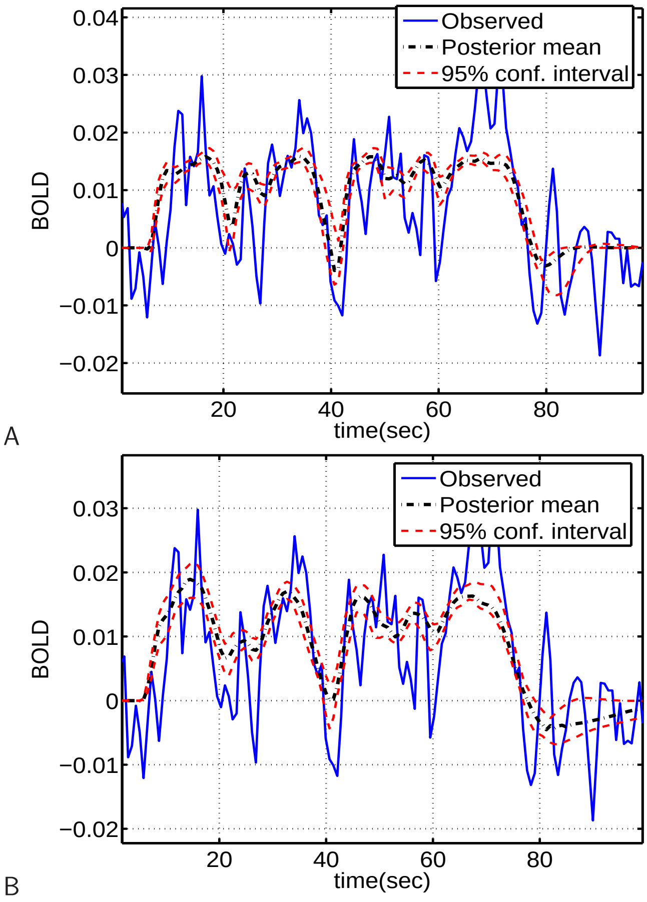


Figure 5: Prediction of data set 2. A: Model 'A'. B: Model 'B'. Note that the confidence interval is an empirical confidence interval for the mean prediction, based on the MCMC samples.

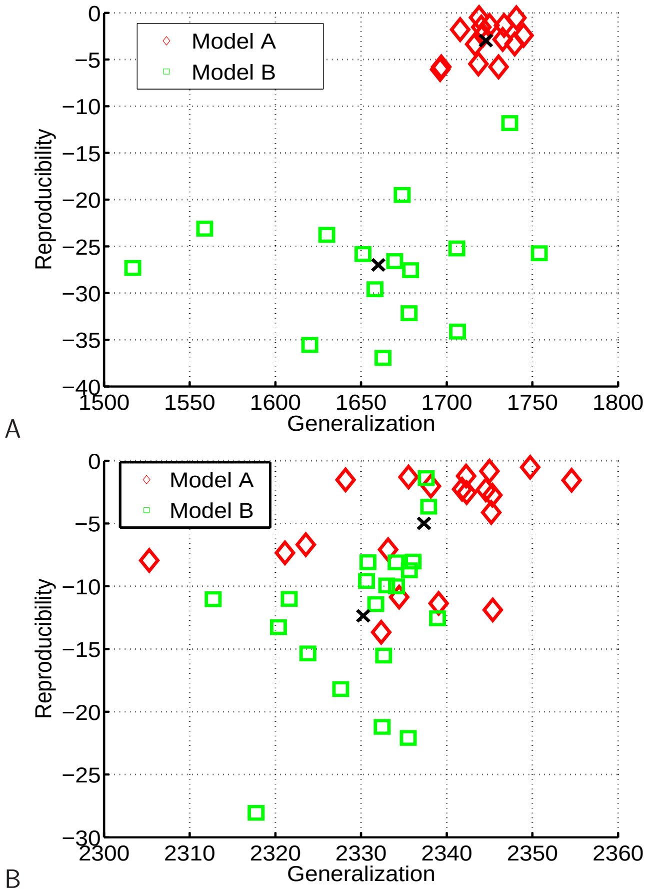


Figure 6: Generalization and reproducibility for real data; crosses mark the mean. A: Data set 1. B: Data set 2.

Appendix C

Publication: Neural Computation II

This appendix contains the article *Deterministic versus stochastic dynamics in non-linear hemodynamic BOLD fMRI - a Bayesian comparison using unscented Kalman filtering*, submitted to Neural Computation, October 2006, author: Daniel J. Jacobsen.

Deterministic versus stochastic dynamics in non-linear hemodynamic BOLD fMRI - a Bayesian comparison using unscented Kalman filtering

Daniel J. Jacobsen
Intelligent Signal Processing
Informatics and Mathematical Modelling
Technical University of Denmark
dj@imm.dtu.dk

December 1, 2006

Abstract

Non-linear hemodynamic models express the BOLD signal in terms of ordinary differential equations. For one such model - the 'balloon' model [6] - the benefit of considering the hemodynamic state variables as stochastic rather than deterministic is investigated by transforming the model into a stochastic state space system. To solve the resulting stochastic differential equations and evaluate the likelihood, an unscented Kalman filter is employed. The parameters of the deterministic model are learned in terms of their approximate posterior distributions, using Markov chain Monte Carlo techniques. For the stochastic model, maximum a posteriori parameters are estimated using simulated annealing. A split-half resampling procedure [19] is then performed that divides available data into training and test sets, allowing unbiased estimation of the generalization abilities of the models, as well as of their reproducibility. This is done both for synthetic and real data, recorded from two different visual stimulation paradigms. The results show that the stochastic state space system is a better model for the more complex data.

1 Introduction

Given the long term goal of increasing the spatio-temporal resolution of BOLD fMRI, models linking subject behavior, neural activity, the hemodynamic response, and fMRI BOLD observations are highly relevant, see e.g. [4, 6, 3, 17].

The precise physiological relationship between subject stimulus, neural activity and the BOLD signal is unclear, and several models have been proposed. Non-linear hemodynamic models express the BOLD signal in terms of ordinary differential equations for a set of hemodynamic state variables. The most widely known and used model is the 'balloon' model, originally developed by Buxton et al. [4] and expanded by Friston et al. [6]. In this model, increased neural activity increases local cerebral blood flow (CBF) and metabolic rate of oxygen consumption (CMRO$_2$), and these affect the level of deoxyhemoglobin and the blood volume (CBV), giving rise to the BOLD signal.

The noise in these nonlinear models is usually assumed to enter in at the BOLD measurement level only, and very little work has been done to investigate the possible benefit of considering the underlying physiological processes to be noisy (the author knows of only [17] and [5]).

The purpose of this work is to compare the classical balloon model to a stochastic state space formulation in a Bayesian manner, taking both reproducibility and the ability to generalize into account.

2 The balloon model

The balloon model consists of a set of ordinary differential equations (ODE's) modelling the evolution in time of four basic physiological state variables, connecting subject stimulus to neural activity. The final part is the output non-linearity describing the BOLD signal as a function of the underlying states.

The four states are blood volume $v(t)$, blood inflow $f(t)$, amount of de-oxyhemoglobine $q(t)$ and a so-called 'flow inducing signal' $s(t)$, collected in the state vector, $\mathbf{x}(t) = [v(t)\ q(t)\ f(t)\ s(t)]^T$. These 'hidden' states are not measurable.

The local neural activity $u(t)$ is considered to be identical to the subject stimulus. The flow inducing signal $s(t)$ is driven by $u(t)$. This in turn drives changes in the other state variables through the ODE's:

$$\frac{\partial \mathbf{x}}{\partial t} = f(\mathbf{x}(t), u(t); \theta) \tag{1}$$

The measured BOLD signal y_n is a non-linear function of 'snapshots' of the continuous states, with additive white Gaussian noise w_n; subscript indices are used for these variables to emphasize their discrete nature.

$$y_n = g(\mathbf{x}(t_n); \theta) + w_n \tag{2}$$

The BOLD signal is measured with a sampling interval denoted TR. The model has seven parameters: σ_w^2, the variance of w_n, and six physiological parameters, combined in $\theta = [\alpha\ \epsilon\ \tau_0\ \tau_s\ \tau_f\ E_0\ \sigma_w^2]^T$.

The states are assumed to evolve from an initial known resting state $\mathbf{x}_0 = [1\ 1\ 1\ 0]^T$ - volume, deoxyhemoglobin and flow at resting levels, stimulus at zero.

The specific differential equations are

$$\frac{\partial v(t)}{\partial t} = \frac{1}{\tau_0}\left(f(t) - f_{out}(t)\right) \tag{3}$$

$$\frac{\partial q(t)}{\partial t} = \frac{1}{\tau_0}\left[f(t)\frac{1-(1-E_0)^{1/f(t)}}{E_0} - v(t)^{(1-\alpha)/\alpha}q_t\right] \tag{4}$$

$$\frac{\partial s(t)}{\partial t} = \epsilon u(t) - s(t)/\tau_s - (f(t)-1)/\tau_f \tag{5}$$

$$\frac{\partial f(t)}{\partial t} = s(t) \tag{6}$$

The blood outflow $f_{out}(t)$ follows

$$f_{out}(t) = v(t)^{1/\alpha} \tag{7}$$

The BOLD observation model is

$$\begin{aligned} g(\mathbf{x}(t);\theta) = V_0[&(k_1+k_2)(1-q(t)) \\ &-(k_2+k_3(1-v(t)))] \end{aligned} \tag{8}$$

with a set of empirical constants ($V_0 = 0.02$, the rest depend on scanner type and settings [2]).

The measurements $Y^N = \{y_0, y_1, \ldots, y_{t_N}\}$ are spatially sampled in volume elements (voxels) and divided temporally into quasi-independent baseline-activation stimulus 'epochs'. Data sets D are defined as collections of one or more epochs.

The neural activity in this model is identical with the stimulus given to the subject, a train of square pulses. This assumption is common in BOLD fMRI analysis (e.g. [6]). The stimulus is set to 1.0 during 'on' periods and 0.0 during 'off' (no stimulus):

$$u(t) = \begin{cases} 1.0 & \text{stimulus on at time t} \\ 0.0 & \text{stimulus off at time t} \end{cases} \tag{9}$$

3 Stochastic hemodynamic variables

An alternative model obtains by considering the state variables as stochastic rather than deterministic variables, giving a set of stochastic differential equations (SDE):

$$d\mathbf{x}(t) = f(\mathbf{x}(t), u(t); \theta)dt + \mathbf{A}d\mathbf{w} \tag{10}$$

Here, $\mathbf{x}(t)$ is the hidden state vector, and $\mathbf{w}$ is a 4-dimensional Wiener process, a stochastic process where the variance of the increments, $w(t) - w(s), t > s$, equals $t - s$. It makes $\mathbf{x}(t)$ a stochastic vector variable. $\mathbf{A}$ is assumed to be a constant diagonal matrix, so the parameters of θ that apply to $\mathbf{A}$ are simply its diagonal elements. The rationale being that the sources of randomness for the different variables are generally physiologically distinct entities,

$$\mathbf{A} = \begin{pmatrix} \sigma_v^2 & 0 & 0 & 0 \\ 0 & \sigma_q^2 & 0 & 0 \\ 0 & 0 & \sigma_f^2 & 0 \\ 0 & 0 & 0 & \sigma_s^2 \end{pmatrix} \tag{11}$$

Further, the choice is made to assume that the noise variance of the hidden states is stationary in time, as is the observation noise. This formulation is not the most general one, but is based on the form of the relevant hemodynamic models and the above mentioned, informed assumptions.

The theory and applications of SDE's is a major research area, and several books have been written on the subject (e.g. [11], [13]). The present application is based on one particular approach [18], considering only the first two moments (mean and covariance) of $\mathbf{x}(t)$, leading to approximate numerical solutions.

4 Likelihoods

The likelihood of the parameters of a model is defined as

$$\mathcal{L}(\theta) \triangleq p(D|\theta) \tag{12}$$

With deterministic hidden states, the likelihood is straightforward to set up. First, the hidden states will evolve deterministically according to (1) driven by $u(t)$. Here, a variable step-size 4th/5th-order embedded Runge-Kutta method was used to solve these [1], with the starting condition $\mathbf{x}(t = 0) = \mathbf{x}_0$, the initial (relaxed) state (all values are relative to resting state). This gives a sequence of states, $\mathbf{x}_{1:N} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, corresponding to the sampling times $\{t_1, t_1 + TR, \ldots, t_1 + N \cdot TR\}$, where t_1 is the starting time of the data set. The estimated mean BOLD signal is given by

$$\overline{y}_n = g(\mathbf{x}_n; \theta) \tag{13}$$

with the observed output given by (2). As the residuals are assumed normally i.i.d., the likelihood becomes

$$p(D|\theta) \triangleq p(y_{1:N}|\theta) = \prod_{n=0}^{N} p(y_n|\theta) = \prod_{n=0}^{N} \mathcal{N}(\overline{y}_n - y_n; 0, \sigma_w^2). \tag{14}$$

The likelihood function with stochastic hidden variables is not so easily expressed; it instead factorizes as

$$\begin{aligned}\mathcal{L}(\theta) \triangleq p(Y^i|\theta) &= p(y_0)p(y_1|y_0)p(y_2|y_1, y_0)\ldots \times p(y_N|Y^{N-1}) \\ &\triangleq \mathcal{L}_0(\theta) \prod_{i=1}^{N} \mathcal{L}_i\end{aligned} \tag{15}$$

where the $\mathcal{L}_i(\theta) \triangleq p(y_i|Y^{i-1})$ and $\mathcal{L}_0(\theta) \triangleq p(y_0)$. This factorization is of course valid for any multivariate stochastic variable.

The $\mathcal{L}_i(\theta)$ terms are calculated using a continuous-discrete unscented Kalman filter (see [18]) and are given as

$$\mathcal{L}_{i+1}(\theta) = \mathcal{N}(0; y_{i+1} - E\left[g_{i+1}|Y^i\right], D(g_{i+1}|Y^i) + \mathbf{A}) \tag{16}$$

where $E\left[g_{i+1}|Y^i\right]$ is the predicted output based on all previous measurements, and $D(g_{i+1}|Y^i)$ is the variance of this prediction.

5 Approximate learning with MCMC and simulated annealing

The posterior $p(\theta|D)$ cannot be evaluated analytically with the present models, so Markov chain Monte Carlo (MCMC) and the related simulated annealing (SA) techniques are used to generate estimates.

Bayes' rule allows us to rewrite the posterior in terms of the likelihood, $p(D|\theta)$, the prior, $p(\theta)$ and a normalizing factor, $p(D)$:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

In order to evaluate the terms in (17), it is therefore necessary to evaluate the likelihood and the prior.

5.1 Prior distributions

There are many approaches to choosing prior distributions. Generally it is important that the priors are as non-informative as possible, and yet they should reflect any prior beliefs held about the parameters. In the present case there actually exists prior physiological knowledge ([6], [3]), so the priors are built thereupon. The priors are assumed to factorize into a product of univariate priors, as there is little or no reason to believe - a priori - that the parameters are correlated.

The prior for the observation noise is simply set to a constant for positive values, $p(\sigma_w^2) = c$ for $\sigma_w^2 > 0$ and $p(\sigma_w^2) = 0$ for $\sigma_w^2 \leq 0$. In practice it is consistently found for synthetic data that the observation noise is accurately estimated with this completely non-informative prior. For all the other parameters, the family of scaled Beta distributions is used as these are well suited to design appropriately flat distributions with upper and lower limits, see figure 1.

5.2 Metropolis-Hastings and parallel tempering

For the deterministic state space model, $p(\theta|D)$ is approximated by a number of samples θ_i, $i = 1..L$. In order to generate these samples the MCMC techniques of Metropolis-Hastings sampling and parallel tempering are employed. Several excellent sources are available that describe these methods (see e.g. [16], [9]), and only an overview will be given here.

Metropolis-Hastings sampling starts at an arbitrary state,
$\theta(i = 0) = \theta_0$, and iteratively proposes small changes through a *proposal distribution*, $p(\theta'|\theta(i))$. Here, a Gaussian centered on the previous state is used:

$$p(\theta')|\theta(i)) = \mathcal{N}(\theta(i), \Sigma)$$

The proposal is accepted as the next discrete sample according to the ratio:

$$r = \frac{p(\theta'|D)}{p(\theta(i)|D)} = \frac{p(D|\theta')p(\theta')}{p(D|\theta(i))p(\theta(i))} \tag{17}$$

If $r \geq 1$ the proposal is accepted, and the next sample generated is $\theta(n+1) = \theta'$. If $r < 1$, the proposal is accepted with probability r. If the proposal is not accepted, the next sample is simply $\theta(i+1) = \theta(i)$.

The Metropolis-Hastings method produces samples from the true posterior distribution in the limit of large number of samples (under certain conditions, see [16]). Depending on the properties of the posterior, 'mixing', i.e. the ability of the algorithm to generate samples representative of the whole distribution, may be slow. The technique of parallel tempering may be employed as a quite

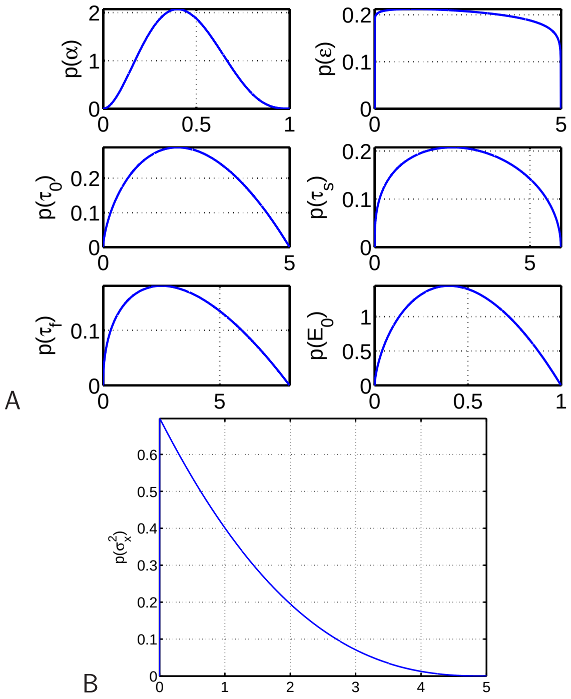


Figure 1: Prior distributions, $p(\theta)$ A: Hemodynamic parameters. B: State space noise variance parameters; the same distribution is used for σ_v^2, σ_q^2, σ_f^2, and σ_s^2.

straightforward enhancement. It works by sampling from several distributions $p_i(\theta|D)$ in parallel, each using a more or less 'flattened' version of the likelihood:

$$p_i(\theta|D) \triangleq \frac{p(D|\theta)^{\beta_i} p(\theta)}{\int p(D|\theta)^{\beta_i} p(\theta) d\theta}, \quad i = 1 \ldots C \tag{18}$$

where $\beta_i \leqq 1$ are so-called *inverse temperatures* and C is the number of 'chains'. At certain intervals, a proposal is made to swap the states of two of these chains (using an acceptance ratio somewhat similar to (17), see [9]), generally leading to better mixing in the same amount of computer time. For the present data, using 6 chains from $\beta_1 = 1.0$ to $\beta_6 = 0.04$ was found to give good results.

A good proposal distribution is a major determinant for success of the algorithm, so an automated procedure for finding a good proposal has been implemented based on iterations of short sampling runs. A set of heuristics are employed to obtain indications that the final sampling estimate of the posterior distribution does indeed cover the true distribution (overlapping sample distributions for independent runs, retrieving the true parameters with synthetic data etc.).

5.3 Simulated annealing

For the stochastic model, evaluation of the likelihood takes too long (around 50 times longer than for the deterministic model) for the MCMC sampling approach to be practically viable. Instead, a point estimate approximating the maximum a posteriori (MAP) is used,

$$p(\theta|D) \approx \delta(\theta - \theta_{MAP}) \tag{19}$$

where θ_{MAP} is defined as

$$\theta_{MAP} \triangleq \arg\max_{\theta} p(\theta|D) \tag{20}$$

The simulated annealing technique produces an approximation to θ_{MAP} and is related to both MCMC sampling and parallel tempering. The procedure is in fact identical to the Metropolis-Hastings sampling described above with the modifications that only the final sample is used as the estimate of θ_{MAP}, and the temperature (see (18)) is gradually decreased towards zero during sampling. This means that $\theta(i)$ will move relatively freely around initially, but as the temperature decreases will only move in the direction of the gradient of $p(\theta|D)$, hopefully resulting in a final value close to θ_{MAP}. For synthetic data, this generally is the case, using a starting temperature of $T \triangleq \frac{1}{\beta} = 10.0$ and slowly decreasing over a run of 4000 samples.

6 Model comparison

The purpose of comparing models can be phrased in terms of two crucial questions:

- *which model provides the best generalization ability?*
- *which model provides the highest reproducibility?*

Together, these two questions form a sound basis for comparing the models.

6.1 Generalization

Generalization ability is well known as fundamental goal of learning (see e.g. [14]). A model should be able to learn based on one data set ('training'), and generalize to another (test) data set. In the predictive learning framework, this means that the distribution of the parameters of a model learned from one data set - the so-called posterior distribution, $p(\theta|D)$ - should be able to 'explain' an independent test set D^* according to

$$p(D^*|D, M) = \int p(\theta|D, M)p(D^*|\theta, M)d\theta. \tag{21}$$

where M is the model. This integral can not be solved analytically due to the non-linearities involved. For the deterministic state space model, the MCMC approximation

$$\int p(\theta|D,M)p(D^*|\theta,M)d\theta \approx \frac{1}{L}\sum_{i=1}^{L} p(D^*|\theta(i),M) \tag{22}$$

can be used, since it approximately holds that $\theta(i) \sim p(\theta|D,M))$. In the stochastic model case, the MAP (Maximum A Posteriori) approximation is used to give

$$p(D^*|D,M) \approx p(D^*|\theta_{MAP},M) \tag{23}$$

This predictive density measured at one or more test data sets is then compared. The corresponding generalization for a given test and training data set is defined as the logarithm of the predictive distribution,

$$G(D^*,D,M) \triangleq \log p(D^*|D,M).$$

The mean predictive generalization over test sets is

$$G(D,M) \triangleq \langle G(D^*,D,M)\rangle_{p(D^*)} = \int [\log p(D^*|D,M)]p(D^*)dD^*.$$

Apart from an additive constant (the entropy of the true distribution), the mean generization is equal to minus the Kullback-Leibler distance between the 'true' distribution of the data, $p(D^*)$, and the model distribution, $p(D^*|M)$. To obtain the overall generalization, averaging is furthermore done over training sets

$$\begin{aligned} G(M) \triangleq \langle G(D,M)\rangle_{p(D)} &= \left\langle \langle G(D^*,D,M)\rangle_{p(D^*)} \right\rangle_{p(D)} = \\ &= \int \left[\int \log p(D^*|D,M)p(D^*)dD^* \right] p(D)dD \end{aligned}$$

In applications the integrals can not be computed, hence the model generalization is estimated using independent test- and training data sets by split-half resampling, see e.g. [19],

$$G(M) \approx \frac{1}{K}\sum_{i=1}^{K} \log p(D_i^*|D_i,M) = \frac{1}{K}\sum_{i=1}^{K} G(D_i^*,D_i,M). \tag{24}$$

With K quasi-independent epochs available, each split of the data leaves uses $K/2$ each for the test and training sets. Each of the resulting estimates is an unbiased estimate of the model generalization. The mean over all splits is a convex combination thus also an unbiased estimate, but with a reduced variance.

6.2 Reproducibility

Reproducibility concerns the sensitivity of what is learned to the particular training data. A model that generalizes well with parameters that vary greatly depending on the particular training data set might be less attractive than a model with a slightly lower ability to generalize, but that produces more robust posterior parameter distributions, particularly in the case of 'physiological' models where the parameters carry physiological meaning. The weight one assigns to generalization and reproducibility is therefore a context dependent trade off.

As an estimate of reproducibility for the present models is chosen the negative of the percentage-wise difference between the parameter estimate from each split of the training data and the mean of the two estimates,

$$R_i(M) \triangleq -\frac{\hat{\theta}_{i1} - \overline{\theta}_i}{\overline{\theta}_i},\ \overline{\theta}_i = (\hat{\theta}_{i1} + \hat{\theta}_{i2})/2 \tag{25}$$

where $R_i(M)$ is the reproducibility estimate for the current split and $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$ are the parameter estimates for the first and second half of the current split, respectively.

For the deterministic model, the mean of the approximated posterior is used as the representative point estimate of the distribution,

$$\hat{\theta} = \frac{1}{L}\sum_{i=1}^{L}\theta_i \tag{26}$$

For the stochastic state space model, the MAP estimate is of course used.

7 Results and discussion

Comparisons of the two models is done both for synthetically generated data and for two different real BOLD fMRI data sets. For the real data, some preprocessing was done to remove artifacts (scanner and physiological). 10 resampling splits were used for all data sets. To justify our assumption that the BOLD signal is independent between epochs, the stimulus for each epoch was set to zero for at least 20 seconds for all data sets.

7.1 Synthetic data

Two synthetic data sets were created, one by the deterministic state space model, the other by the stochastic model, both using the same stimulus function as the one used to generate data set 2 (below). This stimulus is designed to evoke non-linear behavior in the model, achieved by using random and rapid stimulus pulses.

The first synthetic data set was obtained by simulating the standard balloon model with parameters set to $\theta = [0.4\ 0.5\ 2.0\ 2.5\ 2.5\ 0.4\ \sigma_w^2]^T$,
with σ_w^2 set to produce a desired SNR (signal-to-noise ratio, measured as the ratio of the de-noised BOLD and observation noise signals) of around 25.0 dB, deliberately higher than for real recording conditions to ensure clear results. Each epoch contains 138 samples with sampling interval $TR = 0.725s$, exactly as for data set 2 (below).

The second synthetic data set was obtained by simulating the stochastic balloon model with the same stimulus signal and parameters as for the deterministic model (see above), and with the noise variances on the hemodynamic processes set to $\sigma_v^2 = \sigma_q^2 = \sigma_f^2 = \sigma_s^2 = 1 \cdot 10^{-2}$. The simulation of the stochastic state variables between the observation time points was done with Euler's method with very small time steps (see e.g. [11]).

The models were compared for both synthetic data sets and the results are shown in figures 3 and 4. When data are generated by the simpler deterministic model, then - as might be expected - both the generalization ability and reproducibility are seen to be higher for the simple model, see figure 3A. But when data are generated by the stochastic state space model, the situation is more complex: the true (stochastic) model is able to generalize significantly better, but the deterministic model is still more reproducible (figure 4A). This relatively poor reproducibility is probably due to the added complexity of the model, and means that either model could be chosen as the 'best' one, depending on the intended use of the model. It is possible that with higher variance in the hidden state noise, the true model would outperform the simpler model in both reproducibility and generalization.

7.2 Data set 1

Data Set 1 was acquired by Dr. Egill Rostrup at Hvidovre Hospital, Denmark, on a 1.5 T Magnetom Vision scanner. The scanning sequence was a 2D gradient echo EPI (T2* weighted) with 66-ms TE, FA=50, FOV=230 mm, TR=330ms. Single slice data was obtained in a para-axial orientation parallel to the calcarine sulcus. The visual stimulus consisted of a rest period of 20s of darkness (using a light fixation dot), followed by 10s of full-field checker board reversing at 8 Hz, and ending by 20s of darkness - a simple block design. Ten separate runs were completed, a total of 1000s recorded at each voxel. The data were preprocessed according to [8]. A ROI of 42 (7 by 6) significantly activated (as determined by SPM2 analysis[1] [7]) voxels from the visual cortex were selected and the mean of the signals was used (see figure 2A).

Figure 5 shows the results for the two models with this data set. 5A shows that the deterministic model has both a higher reproducibility and a higher generalization ability than the stochastic model. 5B shows the predicted BOLD signal for the deterministic model (mean of the approximate $p(D^*|D, M)$) together with bounds corresponding to upper and lower confidence intervals (minimum and maximum predicted means of $p(D^*|D, M)$ across data set splits). As there are 10 splits, each test epoch is predicted 10 times and thus using the second-highest and second-lowest mean predictions would be an approximate 80% confidence interval, but here the highest and lowest are used instead to get the widest possible confidence interval. 5C shows the same prediction for the stochastic model. Although the difference in predictions is difficult to see, the G-R plot reveals the deterministic model to be better for this data.

7.3 Data set 2

The data was acquired at Hvidovre Hospital, Denmark, using a 3T scanner (Magnetom Trio, Siemens). 1382 GRE EPI volumes each consisting of twelve 3mm slices oriented along the calcarine sulcus were obtained. Additional parameters where TR=725 ms, TE=30 ms FOV=192 mm, 64x64 acquisition matrix, FA = 82. The stimulus consisted in a circular black/white flickering checkerboard (24 degrees horizontal, 18 degrees vertical) on a grey background. The checkers reversed black/white at 8 Hz. The activation pattern ($a(t)$) used to determine on- and offset of this stimulus was the same as was used to generate the synthetic data. A ROI of 75 (25 from each of

[1] Software available from http://www.fil.ion.ucl.ac.uk/spm/

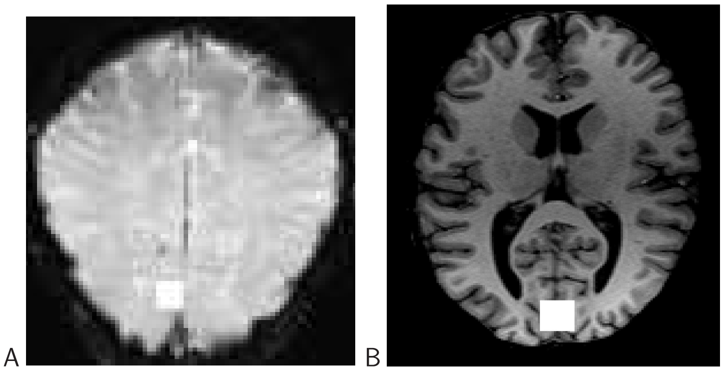


Figure 2: Regions of Interest, marked with white squares. A: Data set 1; T2* weighted image slice parallel to the calcarine sulcus. B: Data set 2; MPRAGE (Magnetization Prepared Rapid Gradient Echo) horizontal slice.

3 slices) significantly activated (again as determined by SPM2 analysis), contiguous voxels in the visual cortex were selected, and the mean of these was used as the BOLD signal (see figure 2B).

Figure 6A shows that for this more 'complex' data set, the stochastic model proves to have a higher generalization ability, although again, the simpler deterministic model is more reproducible. The ability of the stochastic model to express greater variation in the mean BOLD signal through the addition of noise in the hidden state space and thus fit to more complex BOLD signals is reflected in the comparison of 6B and 6C, but this may also explain the reduced reproducibility. Incidentally, this lower reproducibility is not only in the σ_x^2 parameters, but also applies to the hemodynamic parameters (separate analysis, not shown).

7.4 Discussion

The method outlined here can be used to compare models of BOLD fMRI data in a principled manner, estimating both the ability of the models to generalize and learn robustly. It is important to apply such comparison methods to new models to determine their relative merit for use in different contexts.

The present results indicate that the simpler deterministic model is better than the more complicated stochastic state space model for the real data based on a block stimulus design. When compared on data generated with a more complex stimulus function, the stochastic model is shown to be better able to capture the structure of the resulting BOLD signal. The price seems to be a reduced reproducibility, so the physiological interpretation of the stochastic state space model is less clear.

With more data, generalization and reproducibility would increase for both models, and a dif-

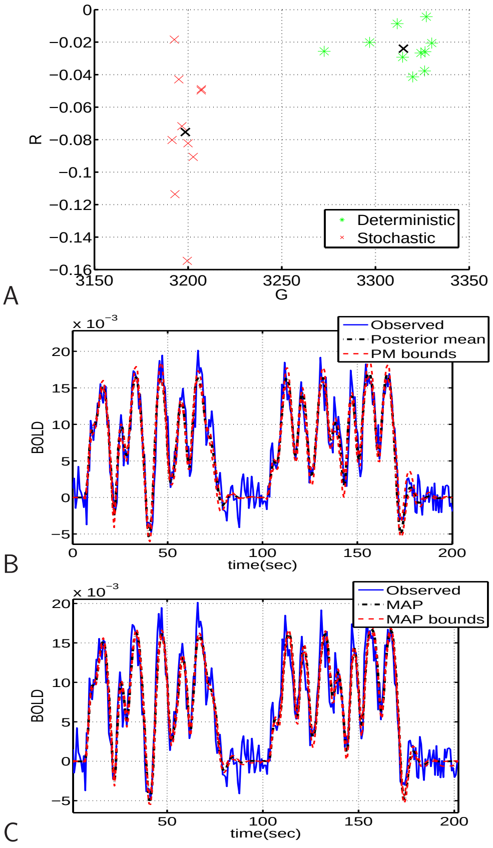


Figure 3: Results for synthetic data from the deterministic model. A: Generalization (G) and reproducibility (R). B and C: Prediction of the first two test epochs by the deterministic (B) and stochastic (C) model.

ferent comparison result could be obtained. It would be of great interest to see similar comparisons

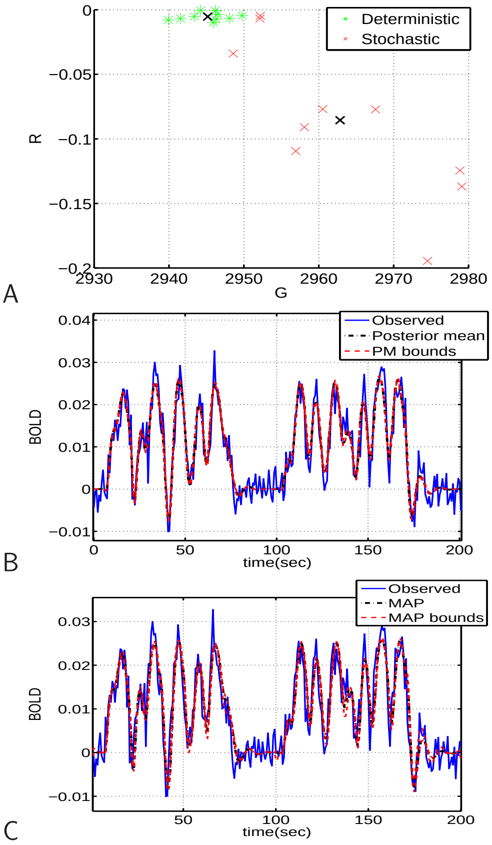


Figure 4: Results for synthetic data from the stochastic model. A: Generalization (G) and reproducibility (R). B and C: Prediction of the first two test epochs by the deterministic (B) and stochastic (C) model.

for other types of data - such as across different parts of the brain and different stimuli - and for

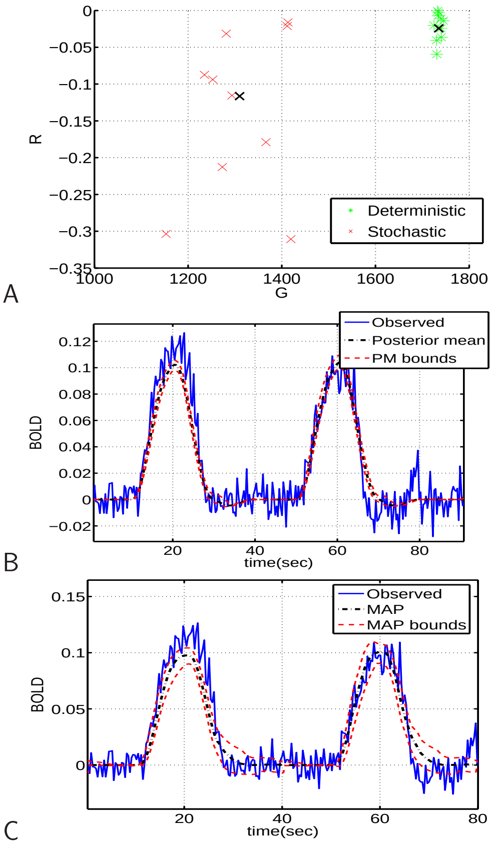


Figure 5: Results for data set 1. A: Generalization (G) and reproducibility (R). B and C: Prediction of the first two test epochs by the deterministic (B) and stochastic (C) model.

other models not investigated here (e.g. [12] and [20]).

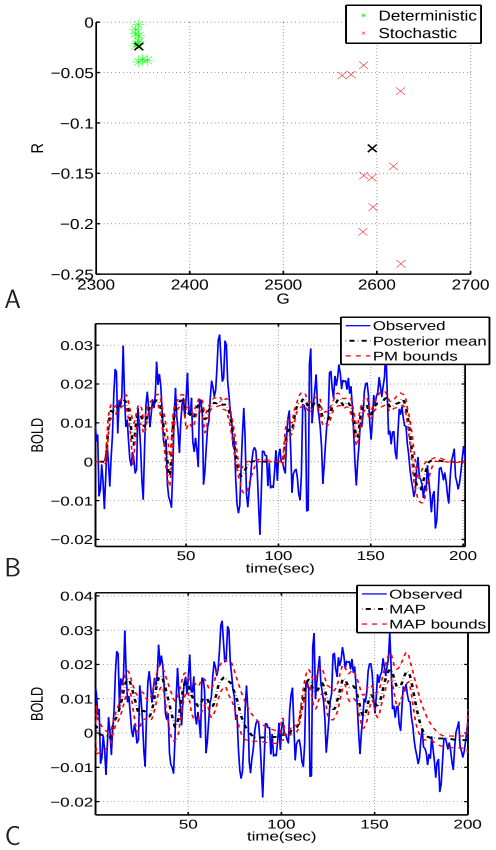


Figure 6: Results for data set 2. A: Generalization (G) and reproducibility (R). B and C: Prediction of the first two test epochs by the deterministic (B) and stochastic (C) model.

References

[1] M. Abramowitz and I. A. Stegun (eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables, 9th printing*, Dover, 1972.

[2] P. A. Bandettini, R. M. Birn, D. Kelley, and Z. S. Saad, *Dynamic nonlinearities in bold contrast: neuronal or hemodynamic?*, 2002, pp. 73–85.

[3] R. B. Buxton, K. Uludag, D. J. Dubowitz, and T. T. Liu, *Modeling the hemodynamic response to brain activation*, NeuroImage **23** (2004), 220–33.

[4] R. B. Buxton, E. C. Wong, and L. R. Frank, *Dynamics of blood flow and oxygenation changes during brain activation: the balloon model*, MRM **39** (1998), 855–864.

[5] T. Deneux and O. Faugeras, *Eeg-fmri fusion of non-triggered data using kalman filtering*, Tech. Report RR-5760, INRIA, France, 2005.

[6] K. J. Friston, A. Mechelli, R. Turner, and C.J. Price, *Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics*, NeuroImage **12** (2000), 466–477.

[7] K.J. Friston, *Introduction: Experimental design and statistical parametric mapping*, second ed., Academic Press, 2003.

[8] C. Goutte, L. K. Hansen, M. G. Liptrot, and E. Rostrup, *Feature-space clustering for fmri meta-analysis*, Human Brain Mapping **13** (2001), 165 – 183.

[9] P. Gregory, *Bayesian logical data analysis for the physical sciences: a comparative approach with mathematica support*, Cambridge University Press, 2005.

[10] S. J. Julier and J. K. Uhlmann, *Unscented filtering and nonlinear estimation*, Proceedings of the IEEE **92** (2004), 401–422.

[11] P.E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations.*, Springer, 1995.

[12] Y. Kong, Y. Zheng Y, D. Johnston, J. Martindale, M. Jones, S. Billings, and J. Mayhew, *A model of the dynamic relationship between blood flow and volume changes during brain activation*, J Cereb Blood Flow Metab **24** (2004), 1382–1392.

[13] B. ksendal.

[14] Jan Larsen and Lars Kai Hansen, *Generalization: The hidden agenda of learning*.

[15] R.S. Liptser and A.N. Shiryayev, *Statistics of random processes.*, vol. II, Springer-Verlag, 1978.

[16] D.J.C. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003.

[17] J.J. Riera, J. Watanabe, K. Iwata, N. Miura, E. Aubert, T. Ozaki, and R. Kawashima, *A state-space model of the hemodynamic approach: nonlinear filtering of bold signals*, Neuroimage **21** (2004), 547–567.

[18] H. Singer, *Continuous-discrete unscented kalman filtering*, Tech. report, FernUniversitt in Hagen, 2005.

[19] S. C. Strother, J. Anderson, and L. K. Hansen, *The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework*, NeuroImage **15** (2002), no. 4, 747–771.

[20] Y. Zheng, J. Martindale, D. Johnston D, M. Jones M, J. Berwick J, and J. Mayhew, Neuroimage **16** (2002), 617–37.

Bibliography

[1] M. Abramowitz and I. A. Stegun (eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables, 9th printing*, Dover, 1972.

[2] D. G. Albrecht, S. B. Farrar, and D. B. Hamilton, *Spatial contrast adaptation characteristics of neurones recorded in the cat's visual cortex*, Journal of Physiology **347** (1984), 713–739.

[3] A.M.Zouhir and B.Boashash, *The bootstrap and its application in signal processing*, IEEE signal processing magazine (1998).

[4] M. W. Andrews, *Learning and inference in nonlinear state-space models*, http://www.citeseer.ist.psu.edu/721245.html.

[5] O.J. Arthurs and S. Boneface, *How well do we understand the neural origins of the fmri bold signal?*, TRENDS in Neuroscience **25** (2002), 27 – 31.

[6] D. Attwell and C. Iadecola, *The neural basis of functional brain imaging signals*, Trends in Neuroscience **25** (2002), no. 12, 621–625.

[7] A. Aubert and R. Costalat, *A model of the coupling between brain electrical activity, metabolism, and hemodynamics: application to the interpretation of functional neuroimaging*, Neuroimage **17** (2002), 1162–81.

[8] P. A. Bandettini, R. M. Birn, D. Kelley, and Z. S. Saad, *Dynamic nonlinearities in bold contrast: neuronal or hemodynamic?*, 2002, pp. 73–85.

[9] Y. Behzadi and T. T. Liu, *An arteriolar compliance model of the cerebral blood flow response to neural stimulus*, NeuroImage **25** (2005), 1100–11.

[10] P. S.F. Bellgowan, Z.S. Saad, and P.A.Bandettini, *Understanding neural system dynamics through task modulation and measurement of functional mri amplitude, latency and width*, Proceedings of the National Academy of Sciences **100** (2003), no. 3, 1415–1419.

[11] J. O. Berger, *Bayesian analysis: a look at today and thoughts of tomorrow*, Journal of the American Statistical Association **95** (2000), 1269–1276.

[12] J. O. Berger and L. R. Pericchi, *Objective bayesian methods for model selection: introduction and comparison*, http://www.citeseer.ist.psu.edu/401338.html, 2001.

[13] C. M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 1995.

[14] A.B. Bonds, *Temporal dynamics of contrast gain in single cells of the cat striate cortex*, Visual neuroscience **6** (1991), 239–55.

[15] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger, *Linear systems analysis of functional magnetic resonance imaging in human v1*, The journal of neuroscience **16** (1996), 4207–4221.

[16] R. B. Buxton, K. Uludag, D. J. Dubowitz, and T. T. Liu, *Modeling the hemodynamic response to brain activation*, NeuroImage **23** (2004), 220–33.

[17] R. B. Buxton, E. C. Wong, and L. R. Frank, *Dynamics of blood flow and oxygenation changes during brain activation: the balloon model*, MRM **39** (1998), 855–864.

[18] P. Ciuciu, *Modeling the bold response in fmri*, http://www.madic.org/people/ciuciu/semin.php, 2004.

[19] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley, 1991.

[20] M. K. Cowles and B. P. Carlin, *Markov chain monte carlo convergence diagnostics: A comparative review*, Journal of the American Statistical Association **91** (1996), no. 434, 883–904.

[21] N. de Freitas, P. A. d. F. R. Højen-Sørensen, and S. J. Russell, *Variational mcmc*, UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (San Francisco, CA, USA), Morgan Kaufmann Publishers Inc., 2001, pp. 120–127.

[22] B. Efron, *Better bootstrap confidence intervals*, Journal of the American statistical association **82** (1987), no. 397.

[23] K. J. Friston, A. Mechelli, R. Turner, and C.J. Price, *Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics*, NeuroImage **12** (2000), 466–477.

[24] K.J. Friston, L L. Harrison, and W. Penny, *Dynamic causal modelling*, NeuroImage **19** (2003).

[25] A. Gelb (ed.), *Applied optimal estimation*, MIT Press, 1974.

[26] Z. Ghahramani and M.J. Beal, *Graphical models and variational methods*, MIT Press, 2000.

[27] Z. Ghahramani and G. E. Hinton, *Parameter estimation for linear dynamical systems*, Tech. Report CRG-TR-96-2, Department of Computer Science, University of Toronto, February 1996.

[28] Z. Ghahramani and S. Roweis, *Learning nonlinear dynamical systems using an em algorithm*, Advances in Neural Information Processing Systems, vol. 11, MIT Press, 1999, pp. 431–437.

[29] G. H. Glover, *Deconvolution of impulse response in event-related bold fmri*, NeuroImage **9** (1999), 416–429.

[30] C. Goutte, L. K. Hansen, M. G. Liptrot, and E. Rostrup, *Feature-space clustering for fmri meta-analysis*, Human Brain Mapping **13** (2001), 165 – 183.

[31] C. Goutte, F. Å. Nielsen, and L. K. Hansen, *Modeling the haemodynamic response in fmri using smooth fir filters*, IEEE transactions on medical imaging **19** (2000), no. 12.

[32] P. Gregory, *Bayesian logical data analysis for the physical sciences: a comparative approach with mathematica support*, Cambridge University Press, 2005.

[33] W. Greiner, *Quantum mechanics: An introduction (3rd edition)*, Springer-Verlag, 1994.

[34] L. Hansen, J. Larsen, and T. Fog, *Early stop criterion from the bootstrap ensemble*, http://www.citeseer.ist.psu.edu/hansen97early.html, april 1997.

[35] L. K. Hansen, *Bayesian averaging is well-temperated*, Advances in Neural Information Processing Systems (S. Solla et al., ed.), MIT Press, 1999.

[36] D. J. Heeger and D. Ress, *What does fmri tell us about neuronal activity=*, Nature **3** (2002), 142–151.

[37] B. Horwitz, K. J. Friston, and J. G. Taylor, *Neural modeling and functional brain imaging: an overview*, Neural networks **13** (2000), 829–846.

[38] J. P. Huber, P. M. Drysdale, and P. A. Robinson, *A physiologically derived spatiotemporal model for fmri hemodynamic responses.*, Abstract and poster at human brain mapping (HBM) (2006).

[39] S. Huettel, A. W. Song, and G. McCarthy, *Functional magnetic resonance imaging*, Sinauer Associates, 2004.

[40] S. Julier and J. Uhlmann, *A new extension of the kalman filter to nonlinear systems*, http://www.citeseer.nj.nec.com/julier97new.html, 1997.

[41] S. J. Julier and J. K. Uhlmann, *Unscented filtering and nonlinear estimation*, Proceedings of the IEEE **92** (2004), 401–422.

[42] R. E. Kaas, B. P. Carlin, A. G., and R. M. Neal, *Markov chain monte carlo in practice: A roundtable discussion*, citeseer.ist.psu.edu/article/kaas97markov.html.

[43] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Transactions of the ASME–Journal of Basic Engineering **82** (1960), no. Series D, 35–45.

[44] R. Kass and A. Raftery, *Bayes factors*, Journal of the American Statistical Association **90** (1995), 773–795.

[45] P.E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations.*, Springer, 1995.

[46] Y. Kong, Y. Zheng Y, D. Johnston, J. Martindale, M. Jones, S. Billings, and J. Mayhew, *A model of the dynamic relationship between blood flow and volume changes during brain activation*, J Cereb Blood Flow Metab **24** (2004), 1382–1392.

[47] B. Øksendal, *Stochastic differential equations. an introduction with applications (5th edition).*, Springer Verlag, 1998.

[48] J. Larsen and L. K. Hansen, *Generalization: The hidden agenda of learning*, http://www.citeseer.ist.psu.edu/larsen97generalization.html.

[49] M. Lauritzen and L. Gold, *Brain function and neurophysiological correlates of signals used in functional neuroimaging*, The journal of neuroscience **23** (2003), 3972–3980.

[50] R.S. Liptser and A.N. Shiryayev, *Statistics of random processes.*, vol. II, Springer-Verlag, 1978.

[51] N. K. Logothetis, *The neural bais of the blood-oxygenation-level-dependent functional magnetic resonance imaging signal*, Phil. Trans. R. Soc. Lond. (2002), no. 357, 1003–1037.

[52] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, *Neurophysiological investigation of the basis of the fmri signal*, Nature **412** (2001), 150–157.

[53] N.K. Logothetis and B. A. Wandell, *Interpreting the bold signal*, Annual Review Physiology **66** (2004), 735–769.

[54] T. E. Lund, K. H. Madsen, K. Sidaros, W. L. Luo, and T. E. Nichols, *Non-white noise in fmri: does modelling have an impact?*, Neuroimage **29(1)** (2006), 54–66.

[55] D.J.C. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003.

[56] J. Martindale, J. Berwick, C. Martin, Y. Kong, Y. Zheng, and J. EW Mayhew, *Long duration stimuli and nonlinearities in the neural-haemodynamic coupling*, Journal of cerebral blood flow and metabolism **25** (2005), 651–661.

[57] Y. Moon, B. Rajagopalan, and U. Lall, *Estimation of mutual information using kernel density estimators.*, Physical Review E **52** (1995), 2318–2321.

[58] R. M. Neal, *Probabilistic inference using Markov chain Monte Carlo methods*, Tech. Report CRG-TR-93-1, University of Toronto, 1993.

[59] ______, *Suppressing random walks in markov chain monte carlo using ordered overrelaxation*, (1999), 205–228.

[60] ______, *Annealed importance sampling*, Statistics and Computing **11** (2001), no. 2, 125–139.

[61] 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 **21** (2004), 144–153.

[62] S. Ogawa, R. S. Menon, D. W. Tank, S. G. Kim, H. Merkle, J. M. Ellerman, 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 **64** (1993).

[63] S. Ogawa, D. W. Tank, R. Menon, J. M. Ellermann, S. G. Kim, H. Merkle, and K. Ugurbil, *Intrinsic signal changes accompanying sensory stimulation: Functional brainmapping with magnetic resonance imaging*, Proc. Natl. Acad. Sci. (USA) **89** (1992), 5951–5955.

[64] J. Pfeuffer, J. C. McCullough, P.-F. Vand de Moortele, K. Ugurbil, and X. Hu, *Spatial dependence of the nonlinear bold response at short stimulus duration*, NeuroImage **18** (2003), 990–1000.

[65] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipies in c++ (2nd ed.)*, Cambridge University Press, 2002.

[66] C. Rasmussen and Z. Ghahramani, *Occam's razor*, Advances in Neural Information Processing, vol. 13, 2000.

[67] R. Redner and H. Walker, *Mixture densities, maximum likelihood and the em algorithm*, SIAM Review **26** (1984), no. 2, 195–239.

[68] J.J. Riera, J. Watanabe, K. Iwata, N. Miura, E. Aubert, T. Ozaki, and R. Kawashima, *A state-space model of the hemodynamic approach: nonlinear filtering of bold signals*, Neuroimage **21** (2004), 547–567.

[69] B. D. Ripley, *Pattern recognition and neural networks*, Cambridge university press, 1996.

[70] P. A. Robinson, P.M. Drysdale, H. Van der Merwe, E. Kyriakou, M.K. Rigozzi, B. Germanoska, and C.J. Rennie, *Bold responses to stimuli: dependence on frequency, stimulus form, amplitude, and repetition rate*, NeuroImage **31** (2006), 585–599.

[71] S. Roweis and Z. Ghahramani, *An em algorithm for identification of nonlinear dynamical systems*, Kalman Filtering and Neural Networks, S. Haykin, ed. (to appear), 2000.

[72] I. Shoji, *Approximation of continuous time stochastic processes by a local linearization method*, Math. Comput. **67** (1998), no. 221, 287–298.

[73] H. Singer, *Continuous-discrete unscented kalman filtering*, Tech. report, FernUniversität in Hagen, 2005.

[74] K. E. Stephan, L. M. Harrison, W. D. Penny, and K. J. Friston, *Biophysical models of fmri responses*, Curr. Opin. Neurobiol. **14** (2004), no. 5, 629–635.

[75] L. Stephen, D. Rottenberg, S. Strother, J. Anderson, S. Muley, J. Ashe, S. Frutiger, K. Rehm, L. K. Hansen, E. Yacoub, and X. Hu, *The evaluation of preprocessing choices in single-subject BOLD fmri using NPAIRS performance metrics*, NeuroImage **18** (2003), no. 1, 10–27.

[76] S. C. Strother, J. Anderson, and L. K. Hansen, *The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework*, NeuroImage **15** (2002), no. 4, 747–771.

[77] H. Valpola and J. Karhunen, *An unsupervised ensemble learning method for nonlinear dynamic state-space models*, Neural Comput. **14** (2002), no. 11, 2647–2692.

[78] R. van der Merwe and E. Wan, *The square-root unscented kalman filter for state and parameter-estimation*, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2001.

[79] A. Vehtari and J. Lampinen, *Bayesian model assessment and comparison using cross-validation predictive densities*, Neural Computation **10** (2002), 735–769.

[80] M. Welling, *The kalman filter (class notes)*, http://www.cs.toronto.edu/ welling/classnotes/classnotes.html.

[81] G. Winkler, *Image analysis, random fields and markov chain monte carlo methods.*, Springer, 2003.

[82] M. W. Woolrich, M. Jenkinson, J. M. Brady, and S. M. Smith, *Fully bayesian spatio-temporal modelin of fmri data*, IEEE transactions on medical imaging **23** (2004), no. 2.

[83] Y. Zheng, J. Martindale, D. Johnston D, M. Jones M, J. Berwick J, and J. Mayhew, *A model of the hemodynamic response and oxygen delivery to brain*, Neuroimage **16** (2002), 617–37.