Bayesian Learning Methods for Modelling Functional MRI

Adrian R. Groves

Image Analysis Group

Oxford Centre for Functional MRI of the Brain

Department of Clinical Neurology

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

December 2009

Abstract

Bayesian learning methods are the basis of many powerful analysis techniques in neuroimaging, permitting probabilistic inference on hierarchical, generative models of data. This thesis primarily develops Bayesian analysis techniques for magnetic resonance imaging (MRI), which is a noninvasive neuroimaging tool for probing function, perfusion, and structure in the human brain.

The first part of this work fits nonlinear biophysical models to multimodal functional MRI data within a variational Bayes framework. Simultaneously-acquired multimodal data contains mixtures of different signals and therefore may have common noise sources, and a method for automatically modelling this correlation is developed. A Gaussian process prior is also used to allow spatial regularization while simultaneously applying informative priors on model parameters, restricting biophysically-interpretable parameters to reasonable values.

The second part introduces a novel data fusion framework for multivariate data analysis which finds a joint decomposition of data across several modalities using a shared loading matrix. Each modality has its own generative model, including separate spatial maps, noise models and sparsity priors. This flexible approach can perform supervised learning by using target variables as a modality. By inferring the data decomposition and multivariate decoding simultaneously, the decoding targets indirectly influence the component shapes and help to preserve useful components.

The same framework is used for unsupervised learning by placing independent component analysis (ICA) priors on the spatial maps. Linked ICA is a novel approach developed to jointly decompose multimodal data, and is applied to combined structural and diffusion images across groups of subjects. This allows some of the benefits of tensor ICA and spatially-concatenated ICA to be combined, and allows model comparison between different configurations. This joint decomposition framework is particularly flexible because of its separate generative models for each modality and could potentially improve modelling of functional MRI, magnetoencephalography, and other functional neuroimaging modalities.

Acknowledgements

This thesis would not have been possible without help, support, and data from many people. I would particularly like to thank:

Mark Woolrich for outstanding supervision, for always making time for me and steering me back onto the right track whenever I was a little too interested in those fascinating-but-useless details;

Steve Smith for advice and discussions on how this all fits into the bigger picture and for fostering a collaborative and supportive atmosphere at FMRIB that has helped to make this degree a joy;

Michael Chappell for continuing to develop our nonlinear modelling software and turning it into something that's actually useful;

Laurence Hunt, Dan Gallichan, Steve Smith, and probably many others at FMRIB for the data sets that formed the foundation of this work;

Stephen Payne for being an objective and valued sounding-board;

Mark Jenkinson for giving me the opportunity to teach the FMRIB Graduate Course and help with integrals and research funding;

Saad Jbabdi and Tim Behrens for brainstorming, moral support, and endless entertainment;

Christian Beckmann for ICA expertise and German precision;

The organisers and tutors of the Life Sciences Interface Doctoral Training Centre for preparing me for the real work that was to come, and particularly the bioinformaticians who instilled in me a lasting enthusiasm for Bayesian statistics;

My classmates from the DTC who stuck together through the years and regularly remind me that there's more to academia than just neuroimaging;

The University of British Columbia Engineering Physics program for preparing me for anything and opening my eyes to how much more there is yet to learn;

Mom and Dad for all their love and support, for encouraging curiosity in all things, and for their continuing guidance;

My darling Kathleen for being a constant source of inspiration and reassurance, and for her endless patience with me; it took a little time, but you calmed me down.

The Clarendon Fund, the Overseas Research Student Award Scheme and the Natural Sciences and Engineering Research Council of Canada for funding my studies at Oxford; and

The Guarantors of Brain, the Organisation for Human Brain Mapping and the Sir Hugh Cairns Memorial fund for generous travel grants.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bayesian neuroimaging has advanced rapidly in the past decade and now provides a principled and enormously varied set of tools for modelling all aspects of brain function, structure, and connectivity. The field has now progressed well beyond simply localizing the regions involved in performing particular tasks and much of the recent interest is now on studying how information is encoded in distributed patterns of brain activity. Functional magnetic resonance imaging (FMRI) provides rich enough data to support a variety of supervised and unsupervised learning techniques which find these patterns automatically, while sophisticated models attempt to obtain a more directed understanding of the underlying brain processes by merging the neuroimaging data with prior information about the types of structured activity that would be particularly interesting to examine.

This thesis examines and develops several techniques for analysing FMRI and applies them not just to functional data but also to the analysis of perfusion, structural and diffusion MRI data. Many of the challenges faced in this work are related directly to analysing multimodal MRI data in a principled and accurate way.

Bayesian learning techniques are a perfect match for performing inference on these complex generative models of the data, because they can adaptively determine which parts of the model are useful for fitting the data and eliminate those that are not needed. This makes it easy to assemble highly-structured models from simple parts

without having to focus on the details of inference or find heuristics for avoiding overfitting. As well as automatically pruning models to an appropriate level of complexity, Bayesian inference provides built-in mechanisms for comparing models to each other, allowing direct hypothesis testing.

1.1 Learning Methods for Neuroimaging

The different types of learning methods that are described below are all valuable in the context of FMRI analysis. However, as in this thesis, they can also be directly applied to a variety of other types of neuroimaging data.

1.1.1 Locating and Quantifying Activity

In *univariate* models, each voxel's time-series is fit to a parameterized signal model and the goal is to make an inference on the relationship between the timecourses of individual voxels and the signal parameters. It is most often used for mapping responses to controlled experimental stimuli, where a general linear model (GLM) regresses each voxel against the expected time-courses of the ideal responses and locates regions that show significant activation. This produces maps of stimulus-related activation levels, and associated statistical significance measures, which are used to locate the areas of the brain that are activated by each experimental condition.

There are many areas of neuroimaging that benefit from using nonlinear modelling of the time-series, especially in terms of fitting biophysical models to make quantitative measurements. This is important for understanding the physiological changes underlying the FMRI contrasts (Hoge et al., 1999), making accurate measurements of blood perfusion profiles (Buxton et al., 1998; Woolrich et al., 2006), and measuring microstructural properties such as axon diameter distributions (Assaf et al., 2008). As well as improving our physiological understanding, these techniques are valuable for clinical research on the mechanisms of disease progression. These rely

on nonlinear generative signal models, and robust inference techniques are required to ensure the model is fit as reliably and accurately as possible from the data. A Bayesian framework makes it possible to improve the results by modelling additional properties of the data, such as modality-specific noise modelling, effective use of biophysical prior information and adaptive spatial regularization.

1.1.2 Multivariate Pattern Analysis

Multivariate decoding models use the entire data set at once to learn the spatial *patterns* in FMRI that reliably correspond to a particular target variable in each timepoint or trial. These approaches benefit from the combined explanatory power across the whole image and are much more flexible because the patterns can be spatially-distributed and can take into account interactions across different voxels. This has widespread applications in experimental psychology research by determining how mental representations map onto patterns of neural activity (Norman et al., 2006). It is especially powerful for decoding mental states when the corresponding spatial patterns are distributed and the overlapping (Haxby et al., 2001) or encoded in detailed spatial patterns (Haynes and Rees, 2005). It can also be used to test specific hypotheses about neural representations, for example of image expression in the visual cortex (Kay et al., 2008). The intrinsic problem of using multivariate methods for supervised learning is that there are typically more inputs (FMRI voxels) than outputs (target values), so overfitting can be a serious problem and regularization becomes essential. Feature selection is an essential part of these machine learning methods which makes hierarchical models very important in order to promote sparse or otherwise highly-regularized explanations of the data. A key part of this is the selection of an appropriate basis set, and these can be compared automatically in a Bayesian framework (Friston et al., 2008a).

1.1.3 Exploratory Analysis

Unsupervised learning techniques automatically find multivariate patterns buried within the data; examples of these techniques are principal component analysis (PCA) and independent component analysis (ICA). These exploratory methods are useful for finding artefacts in FMRI data, and can often recover spatial patterns and timecourses for the stimulus-related activation without being informed of the stimulus timings. These techniques have also provided a robust way to analysing resting state networks (RSNs), which are slow oscillations in resting brain activity that have been shown to be disrupted in certain disease states. By looking for an information-driven decomposition of the data, exploratory techniques can find structure that could not have been modelled explicitly.

1.2 Magnetic Resonance Imaging (MRI)

MRI is a powerful, non-invasive imaging system that is the preferred method for examining many aspects of brain structure and function. Much of the success of MRI in neuroimaging can be attributed to its flexibility. Using different pulse sequences it can interrogate the brain tissue to measure its magnetic and physical properties in a variety of different ways. Among the contrasts that can be imaged there are several that are indirectly related to brain activity; these make it possible to map, quantify, and explore the patterns of activity in patients and normal subjects in a safe, non-invasive way.

MR scanners use a very strong magnetic field (typically 3 Tesla for neuroimaging) which causes the water molecule protons to resonate at a particular (field-dependent) frequency when excited by a tuned radio frequency (RF) pulse. Magnetic field gradients are used to encode the signal spatially, making it possible to read out an image. It is a particularly flexible tool because it can produce a wide variety of imaging contrasts by the use of programmable sequences of RF pulses and gradients to prepare

the magnetization of tissue and blood in various ways. Some of these preparations will persist for several seconds and using an appropriate sequence of preparations can sensitise the signal to different magnetic properties as well as microscopic and macroscopic physical movement.

Functional magnetic resonance imaging (FMRI) is the application of MRI techniques to mapping brain activity, by imaging haemodynamic changes that are indirect, delayed measures of neurological activity (Jezzard et al., 2001). There are two main effects that are used in this thesis for detecting brain activity; the first is the blood oxygen level dependent (BOLD) contrast which is based on changing *magnetic* properties of the blood related to oxygenation and the second arterial spin labelling (ASL) is affected by blood *perfusion* into the imaging slice. Both of these are imaged using rapid pulse sequences that acquire an entire slice at once and reacquire the same slice repeatedly with a repeat time (TR) of around 3 seconds. Thus a typical 10-minute scan will have around 200 timepoints for each voxel in the image, which can be analysed to find patterns of brain activity.

1.2.1 Blood Oxygen Level Dependent (BOLD) Contrast

The BOLD contrast is the workhorse of FMRI and provides good contrast for most stimulus paradigms.

An overview of the main BOLD response is given in figure 1.1 and more detail is provided by Hoge and Pike (2001). Neural activity increases oxygen consumption, but also triggers a large increase increase in flow (up to 50–70% above baseline). This flow increase is a much larger effect and within a few seconds this results in a drop in the deoxyhaemoglobin (dHb) concentrations in the venules. The dHb in the blood creates inhomogeneities in the local magnetic field which in turn cause rapid signal decay with decay constant T_2^* (typically $20 - 40$ ms). Since there is less dHb the signal decays more slowly, so the net result is that increased activity leads to a

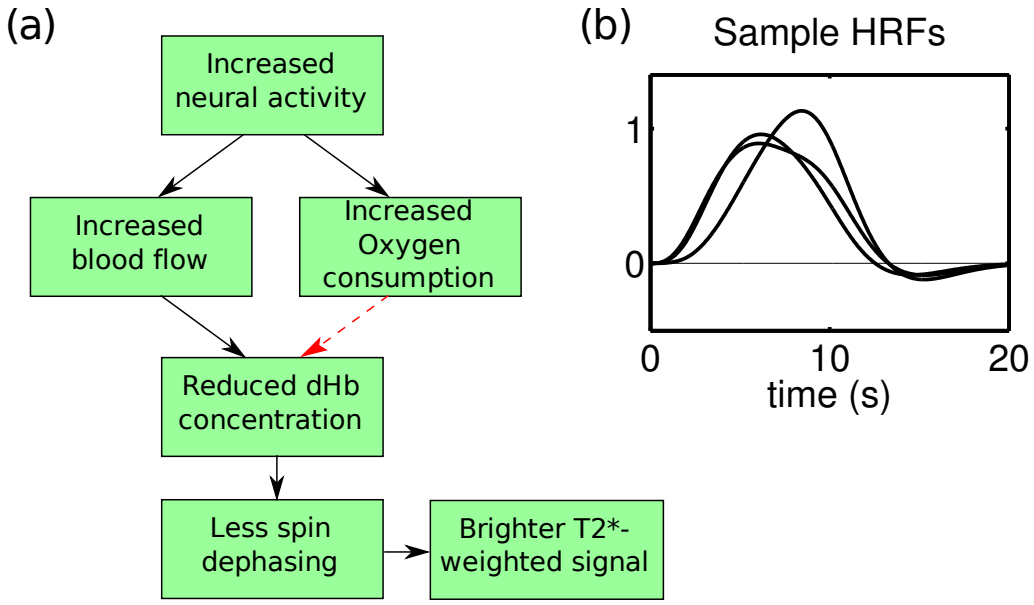


Figure 1.1: The blood oxygen level dependent (BOLD) contrast. (a) The observed effect results from the brain's physiological response to increasing neural activity. This takes several seconds and includes both positive and negative components. (b) The overall linear response in the T_2^*-weighted image is called the haemodynamic response function (HRF) and can very in delay, duration, and shape.

brighter T_2^*-weighted image (on the order of a 1% increase in signal).

Generally the response is assumed to be linear and time-invariant, i.e. fully described by an impulse response function. This haemodynamic response function (HRF) is the combination of all of these effects that links a short burst of neural activity to the BOLD response. The shape is generally determined empirically, and often modelled as a canonical HRF with some minor variability around it, but in normal grey matter the basic shape is quite consistent. This – not the TR – limits the effective temporal resolution of FMRI.

The BOLD effect provides a strong contrast, but it is difficult to convert this back into quantitative physical units due to the indirectness of the effect. There is also a great deal of low-frequency noise that can make BOLD unsuitable for very slow paradigms. These problems mean that in some situations a different FMRI contrast is needed.

1.2.2 Arterial Spin Labelling Imaging

Arterial Spin Labelling (ASL) is an MRI technique for measuring blood flow in a directly quantifiable way (Wong et al., 1998). Figure 1.2 shows the basic concept: a magnetization preparation pulse is applied to a slab of tissue below the area of interest, often in the neck, in order to "tag" the blood as it flows through that region. This creates a virtual contrast agent *in situ*, which then perfuses throughout the brain tissue. When an image is acquired in the slice of interest (typically 0.5-1.5 sec later), some of the signal returned will be from this tagged blood. The sequence alternates between tag and non-tag ("control") TRs and the only systematic difference between the tag and control images will be due to the differential magnetization of inflowing blood. This difference signal can be easily converted into a quantitative Cerebral Blood Flow (CBF) and has an obvious baseline: when there is zero flow, the tagging does not affect the image brightness so the tag/control difference is zero.

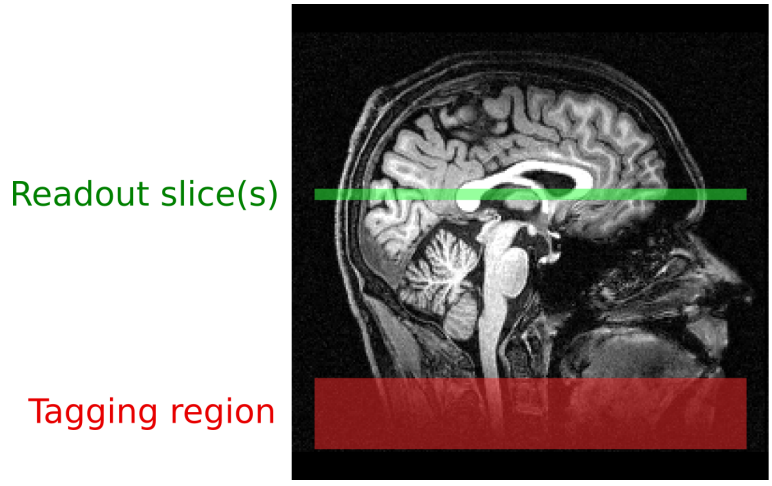


Figure 1.2: Arterial Spin Labelling (ASL) measures cerebral blood perfusion by magnetically inverting the tagging region to create a bolus of tagged blood. When an image is acquired some time later in the imaging slice, this will contain some of the tagged blood; this is the tag image. Repeating the process without applying the tagging pulse produces a control image. The difference between these images is directly proportional to the cerebral blood flow (CBF). Varying the delay between tagging and image acquisition makes it track the bolus's arrival and estimate the flow delay Δt from the tagging region to the readout slice.

As before, neural activity increases CBF, leading to a measurable increase in the tag-control difference. ASL also has a similar haemodynamic response to BOLD, although it may be slightly more rapid than the BOLD response. Importantly, since the tag-control difference provides an absolute measurement, it does not suffer from low-frequency drift. However, the SNR for detecting functional activity is only about $1/10^{\text{th}}$ of the strength of the BOLD contrast for a given stimulus.

Despite this low signal, ASL can be the only practical way to detect neural differences in very slow paradigms, such as those with periods of oscillation longer than a minute (Aguirre et al., 1997). It is particularly useful in conjunction with BOLD measurements in quantitative FMRI (Hoge et al., 1999), which aims to compare the different physiological changes and potentially detect differences in brain metabolism due to disease. ASL is also clinically important for stroke imaging because it can detect abnormal blood perfusion and assess the risk to blood flow to different areas of the brain. In this case the initial arrival time of the bolus, Δt, is particularly important; it can be measured by building up a complete perfusion curve by acquiring images for range of delays (or inversion times, TI).

1.2.3 Structural and Diffusion Data

As well as mapping brain activity, MRI is a very powerful method for examining brain structure and assessing physical properties that reflect microstructure. These static images can be compared across different subjects in a group to find systematic differences, such as disease-related degeneration or learning-related changes.

One of the most important predictors of neurodegenerative disease is changes in the amount and thickness of the grey matter (GM). High-resolution structural images are segmented into tissue types and using nonlinear registration (and smoothing) to map them into the same space. This Voxel Based Morphometry (VBM) technique means that these prepared spatial maps can be analysed for systematic changes

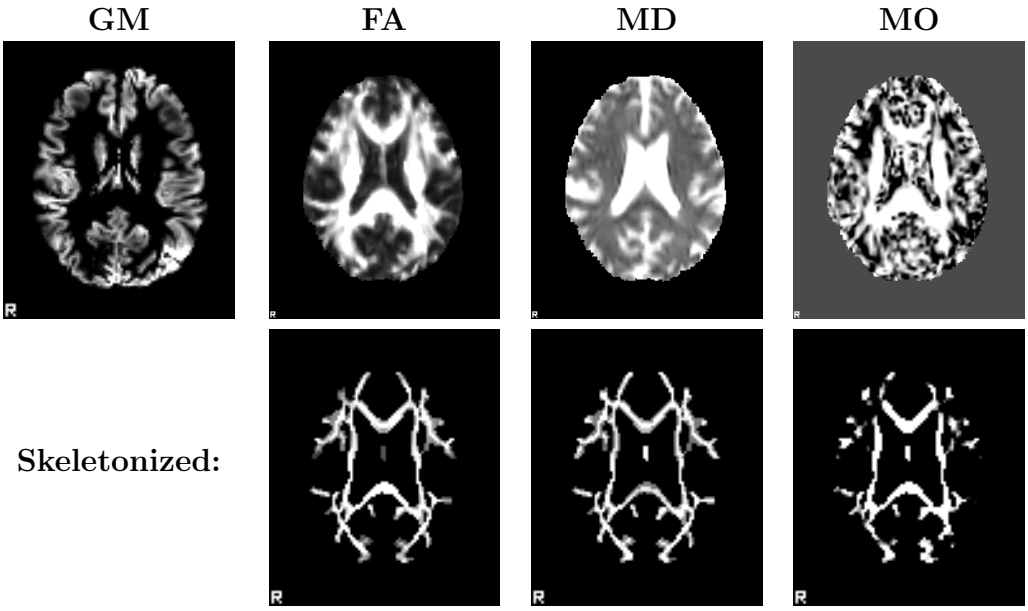


Figure 1.3: Examples of structural and diffusion images of a single subject, registered to standard space. GM shows the grey matter partial volume map, from segmentations of a structural MRI scan. The next three measures are derived from diffusion tensor imaging: Fractional Anisotropy (FA), Mean Diffusivity (MD), and tensor mode (MO). Since these measures are used for assessing the integrity of white matter tracts, they are projected onto a common white matter skeleton using TBSS to reduce the effect of misregistration. This produces the skeletonized images in the second row.

between groups (Ashburner and Friston, 2000).

Diffusion tensor imaging (DTI) is a complementary technique that indirectly images the integrity of the white matter tracts that form the long-distance connections in the brain (Le Bihan et al., 2001). By applying a series of gradients in a selected direction, the image can be sensitised to cause signal loss in the presence of random diffusion along that direction. By acquiring 6 or more directions, it is possible to fit this to a diffusion tensor model, which estimates the directions of diffusion and describes the tensor shape using three rotation-invariant quantities: Fractional Anisotropy (FA), Mean Diffusivity (MD) and tensor mode (MO). These quantities can be mapped and used in VBM analyses directly, but problems arise because because small misregistrations can can lead to misleading changes in apparent integrity of the white matter tracts. A robust solution to this is tract-based spatial statistics (TBSS)

which "skeletonises" the tracts by only considering the most-isotropic voxels in the middle of each tract (Smith et al., 2006). This skeleton forms a two-dimensional surface in three-dimensional space. The other measures are registered using the FA as a guide. Figure 1.3 shows the resulting maps for a single subject across all four of these imaging contrasts.

1.3 Summary of the Remaining Chapters

After a brief introduction to Bayesian techniques this thesis develops and evaluates methods for univariate nonlinear modelling, then develops a new data fusion model for integrated decomposition and decoding and modifies this model to perform simultaneous ICA across multi-modality structural MRI data.

Chapter 2 introduces Bayesian modelling and the variational Bayes (VB) approximation used throughout this thesis.

Chapter 3 develops a Bayesian univariate analysis technique for fitting nonlinear models, using a local linearization to allow approximate VB inference on the nonlinear signal model. This is applied to the analysis of dual-echo ASL data, which is an FMRI pulse sequence that acquires both the BOLD and ASL contrasts simultaneously in two mixed streams of data. This shows the benefits of modelling both streams of data jointly and assesses the importance of modelling between-stream noise correlation in this context.

Chapter 4 presents a novel Gaussian process prior which allows adaptive spatial regularization to be performed while still incorporating informative biophysical priors on parameter values. This is implemented as part of the VB nonlinear modelling framework, using a mixture of VB and evidence optimization (EO) for inference to deal with some strong correlations across voxels. Two models that have useful informative priors are assessed with this new framework: (i) a BOLD FMRI analysis model that permits only reasonable HRF shape variations using constrained basis

sets and (ii) a perfusion model for resting multi-inversion ASL data with constraints on reasonable arrival times.

Chapter 5 develops a novel approach to supervised learning that integrates a decomposition of the neuroimaging data into components and a linear decoding of these component weights into several behavioural variables (decoding targets) simultaneously. This is a data fusion model, treating both the neuroimaging data and the behavioural variables as two different kids of data and building separate generative models for each of them. Allowing the decoding model to refine the decomposition can estimate more accurate data components and help to give them a direct interpretation. This approach also allows several behavioural variables to be decoded simultaneously.

Chapter 6 applies this data fusion framework to an unsupervised learning problem: performing independent component analysis (ICA) on data sets that span several different modalities of neuroimaging data. By using separate Bayesian ICA models for each modality, different noise levels, different mixture models, different numbers of components and different tensorial structure can be inferred for each modality.

Chapter 7 draws some overall conclusions to summarize this work and suggests directions for future development of both the univariate and multivariate threads of this research.

Chapter 2

Bayesian methods

Bayesian inference offers several major advantages over classical statistical methods. Many of the classical preprocessing steps, such as autoregressive noise correction, spatial smoothing or principal component analysis, can be made a full part of a Bayesian model and inferred simultaneously with the parameters of interest. All parameters are expressed as probability distributions, so every point estimate comes with an uncertainty attached; this makes it possible to build detailed hierarchical models from modules that interact with each other probabilistically. Models can also be adaptive to automatically find an optimal trade-off between accuracy and complexity, avoiding much of the tuning that can be required with classical methods.

2.1 Bayesian Modelling

The goal of Bayesian analysis is to infer what one should believe about the value of the model parameters, $\boldsymbol{\Theta}$, given the data, $\mathcal{D}$, subject to an assumed model, $\mathcal{M}$. This knowledge is embodied in the "posterior probability distribution" given by $\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D},\mathcal{M})$. A single application of Bayes' Rule (Bayes and Price, 1763) makes it possible to turn these conditional probabilities around to express the posterior as

$$\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D},\mathcal{M}) \;=\; \frac{\mathrm{P}(\mathcal{D}|\boldsymbol{\Theta},\mathcal{M})\,\mathrm{P}(\boldsymbol{\Theta}|\mathcal{M})}{\mathrm{P}(\mathcal{D}|\mathcal{M})}. \tag{2.1}$$

so the *posterior distribution* on $\boldsymbol{\Theta}$ (what one should believe about the parameters, given the data and model) is the product of the likelihood $\mathrm{P}(\mathcal{D}|\boldsymbol{\Theta}, \mathcal{M})$ and the prior $\mathrm{P}(\boldsymbol{\Theta}|\mathcal{M})$, divided by the model evidence $\mathrm{P}(\mathcal{D}|\mathcal{M})$. From now on, the model $\mathcal{M}$ will be implicit in all of these expressions.

Since the evidence does not depend on $\boldsymbol{\Theta}$ it can be treated as a normalizing constant, so

$$\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D}) \quad \propto \quad \mathrm{P}(\mathcal{D}|\boldsymbol{\Theta})\,\mathrm{P}(\boldsymbol{\Theta}) \tag{2.2}$$

By this transformation, the difficult problem of predicting a parameter estimate from the data is turned into the simpler problem of predicting the data set given the parameters. A generative model is usually simpler to develop than finding an estimator of the parameters directly.

2.1.1 Generative Models

The generative model $\mathrm{P}(\mathcal{D}|\boldsymbol{\Theta})$ provides the central part of the whole Bayesian model by describing what the data should look like for a given set of parameter values. This is usually structured as a deterministic predictive model with an additive noise model.

Noise is a very important part of real data and in a Bayesian framework is modelled explicitly. The noise parameters are included as part of $\boldsymbol{\Theta}$ and must also be inferred from the data. An accurate noise model is essential to ensure that the inferences are reliable. This problem is also present in classical techniques where white Gaussian noise is generally assumed and preprocessing techniques are used to prewhiten the noise (Woolrich et al., 2001). The Bayesian approach exposes this directly and permits flexible modelling, including uncertainty on the noise parameters, rather than relying on empirical corrections to adequately correct for correlations in the the noise.

2.1.2 Role of Priors

The prior distribution $P(\boldsymbol{\Theta})$ is the other essential part of the model definition. They are always present; even a simple maximum likelihood method implicitly assumes a flat prior (all $\boldsymbol{\Theta}$ are equally probable *a priori*).

Often these are set to be uninformative, to ensure that the model is data-driven. For the purposes of this work, an uninformative distribution is any weak prior that is designed so that its contribution to the posterior distribution is dwarfed by the contribution of the data.

The simplest kind of informative prior is one that expresses a real *a priori* beliefs about the value of a parameter. This is particularly useful when the parameter is only weakly constrained by the data and has a direct meaning outside of the model where the typical range of values for this is known *a priori* from external measurements. For example, parameters that are biologically interpretable may have informative priors based on biophysical plausibility. Rather than assuming a constant for unknown parameters, it is preferable to use an explicit model parameter with a reasonable prior since this automatically includes this uncertainty in the inference of the parameters of interest.

Hierarchical priors provide a way to build more structure into the priors so that they regularize the data adaptively. The amount of regularization is derived automatically from the data itself, and the uncertainty on the hyperparameters themselves is modelled. Adaptive regularization in the model is more principled and more flexible than preprocessing steps that regularize, e.g. smooth the data by a pre-specified amount.

Typically, some of the parameters in a hierarchical model are *hyperparameters* that do not directly affect the generative model of the data, but only affect the priors on the other parameters. For example, the parameters can be divided into $\boldsymbol{\Theta} = \{\mathbf{X}, \boldsymbol{\alpha}\}$, where $\mathbf{X}$ are the parameters that directly affect the generative model and $\boldsymbol{\alpha}$ are the

hyperparameters. It is then possible to rewrite the generative model

$$\mathrm{P}(\mathcal{D}|\boldsymbol{\Theta}) \;=\; \mathrm{P}(\mathcal{D}|\mathbf{X}) \tag{2.3}$$

and split the prior into

$$\mathrm{P}(\boldsymbol{\Theta}) \;=\; \mathrm{P}(\mathbf{X}|\boldsymbol{\alpha})\,\mathrm{P}(\boldsymbol{\alpha})\,. \tag{2.4}$$

Using a hierarchical prior, a single hyperparameter can regularize the values of a large set of parameters and help to avoid the problem of overfitting (Penny et al., 2005). With enough structure, it can even be sensible to infer on a hierarchical model with more parameters than data points (Gelman et al., 2004, Ch. 5). For example, automatic relevance determination (ARD) provides a way for a single hyperparameter to set a prior on the scale of another set of variables (such as in Bishop (1999), for example). This is crucial in many applications because it allows parts of a model to be completely eliminated if they are not useful for explaining the data (by forcing the scale to zero).

2.1.3 Model Evidence

For the purposes of inferring the posterior distribution, the evidence is simply a normalizing constant because it does not depend on $\boldsymbol{\Theta}$. If desired, it can be calculated by integrating over the whole unnormalized posterior distribution:

$$\mathrm{P}(\mathcal{D}|\mathcal{M}) \;=\; \int \mathrm{P}(\mathcal{D}|\boldsymbol{\Theta},\mathcal{M})\,\mathrm{P}(\boldsymbol{\Theta}|\mathcal{M})\;\mathrm{d}\boldsymbol{\Theta}. \tag{2.5}$$

Note that the model $\mathcal{M}$ has been made explicit again.

The model evidence is primarily useful for model comparison, because it provides a quantitative way to assess which of several models is most likely to have generated the observed data. For example, if $\mathrm{P}(\mathcal{D}|\mathcal{M}_1) \gg \mathrm{P}(\mathcal{D}|\mathcal{M}_2)$, then $\mathcal{M}_1$ can be considered a better model for the observed data. More properly, if there are m models number 1 to M, then

$$\mathrm{P}(\mathcal{M}_m|\mathcal{D}) \;\propto\; \mathrm{P}(\mathcal{D}|\mathcal{M}_m)\,\mathrm{P}(\mathcal{M}_m) \tag{2.6}$$

where all models are implicitly assumed to have equal prior probability ($\mathrm{P}(\mathcal{M}_m) = 1/M$) and exactly one of the models is assumed to be correct ($\sum_m \mathrm{P}(\mathcal{M}_m|\mathcal{D}) = 1$). The posterior on the parameters, over all models considered, is then a weighted average over the models' posterior distributions:

$$\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D}) \;=\; \sum_{m=1}^{M} \mathrm{P}(\mathcal{M}_m|\mathcal{D})\,\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D},\mathcal{M}_m)\,. \tag{2.7}$$

This is referred to as model averaging (Trujillo-Barreto et al., 2004).

2.1.4 Inference

Although the full posterior distribution is given up to a constant by the equation 2.2, this is not as useful as it may appear. Finding the normalizing constant can only be found by integrating over the whole distribution, but just as importantly, this functional form does not provide the needed marginal distributions of individual parameters of interest. For example, the mean value of a model parameter $x \in \boldsymbol{\Theta}$ is given by $\int x\ \mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})\,\mathrm{d}\boldsymbol{\Theta}$, which is also expressed as the expectation $\langle x \rangle_{\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})}$.

For trivial problems the normalizing constant and marginal moments can be found by integration, but for most problems of interest these integrals are intractable. They are generally found either by numerical integration or by finding simpler analytic approximations to the posterior. The next sections will consider the use of both.

2.2 Variational Bayesian Inference

Variational methods are the main inference method used throughout this thesis. They keep all of the parameters probabilistic, and instead find an approximation to the posterior that is restricted to follow a specified *factorization* over the model parameters. The variational Bayes (VB) formalism provides a tractable way to perform approximate Bayesian inference on graphical models while keeping most of the benefits of fully-Bayesian inference (Attias, 2000).

VB infers an *approximate posterior* distribution, denoted by $\mathrm{P}'(\boldsymbol{\Theta})$. VB places restrictions on the form of this approximate posterior, and then performs a functional optimization so that it is the best possible approximation to the true posterior $\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})$ subject to those restrictions.

Given any probability distribution function $\mathrm{P}'(\boldsymbol{\Theta})$, a lower bound on the log model evidence is found by splitting it into two parts:

$$\begin{aligned}
\log\mathrm{P}(\mathcal{D}) &= \int \mathrm{P}'(\boldsymbol{\Theta}) \log\mathrm{P}(\mathcal{D}) \, \mathrm{d}\boldsymbol{\Theta} && (2.8)\\
&= \int \mathrm{P}'(\boldsymbol{\Theta}) \log\left(\frac{\mathrm{P}(\boldsymbol{\Theta},\mathcal{D})}{\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})}\right) \mathrm{d}\boldsymbol{\Theta} && (2.9)\\
&= \int \mathrm{P}'(\boldsymbol{\Theta}) \log\left(\frac{\mathrm{P}(\boldsymbol{\Theta},\mathcal{D})\,\mathrm{P}'(\boldsymbol{\Theta})}{\mathrm{P}'(\boldsymbol{\Theta})\,\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})}\right) \mathrm{d}\boldsymbol{\Theta} && (2.10)\\
&= F + KL && (2.11)
\end{aligned}$$

where F is the negative variational free energy given by the function of functions:

$$F = F(\mathrm{P},\mathrm{P}') = \int \mathrm{P}'(\boldsymbol{\Theta}) \log\left(\frac{\mathrm{P}(\boldsymbol{\Theta},\mathcal{D})}{\mathrm{P}'(\boldsymbol{\Theta})}\right) \mathrm{d}\boldsymbol{\Theta} \quad (2.12)$$

and KL is the KL divergence or relative entropy (MacKay, 2003, Ch. 33.4) between the arbitrary probability distribution $\mathrm{P}'(\boldsymbol{\Theta})$ and the true posterior distribution $\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})$:

$$KL = KL(\mathrm{P},\mathrm{P}') = \int \mathrm{P}'(\boldsymbol{\Theta}) \log\left(\frac{\mathrm{P}'(\boldsymbol{\Theta})}{\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})}\right) \mathrm{d}\boldsymbol{\Theta} \quad (2.13)$$

The KL divergence between two distributions is always non-negative and approaches zero as the distributions become identical; as a result, the VB approach seeks to minimize KL. Since the model evidence $\mathrm{P}(\mathcal{D})$ is constant for a given data set, this is equivalent to maximizing the free energy F. Since F is a lower bound on the model evidence $\mathrm{P}(\mathcal{D})$, it can be used in place of the evidence and provides a direct way to compare models.

For now, $\mathrm{P}'(\boldsymbol{\Theta})$ is just an arbitrary probability density. However, by maximizing F subject to constraints on the form, the KL divergence between $\mathrm{P}'(\boldsymbol{\Theta})$ and $\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})$

is minimized, making the approximate posterior as similar as possible to the true posterior. This functional maximization could be performed directly but this is very computationally expensive. By using distributions in the exponential family, the free energy F becomes a function of only the *sufficient statistics* of $\mathrm{P}'(\boldsymbol{\Theta})$. By suitably restricting the form of $\mathrm{P}'(\boldsymbol{\Theta})$, the approximate posterior becomes conjugate with the priors and the optimization becomes very computationally efficient. The next sections describe the restrictions on $\mathrm{P}'(\boldsymbol{\Theta})$ and show what steps are taken in order to maximize F.

2.2.1 Mean Field Approximation

To make the inference tractable, the mean field approximation is often used. This restricts the approximate posterior distribution to a form that is factorized as

$$\mathrm{P}'(\boldsymbol{\Theta}) \;=\; \prod_i \mathrm{P}'_{\boldsymbol{\Theta}_i}(\boldsymbol{\Theta}_i) \tag{2.14}$$

where the model parameters $\boldsymbol{\Theta}$ have been split into several groups, $\boldsymbol{\Theta}_i$, each one having its own approximate posterior distribution $\mathrm{P}'_{\boldsymbol{\Theta}_i}(\boldsymbol{\Theta}_i)$. From now on the subscript on P' will be omitted for clarity, i.e. $\mathrm{P}'(\boldsymbol{\Theta}_i)$.

There is some flexibility in the grouping of model parameters. Each factor should be sufficiently simple that it can be integrated to obtain a tractable analytic solution for the updates in the next section. Additional factorization beyond this will reduce the accuracy of the approximation but may make the updates much faster to calculate, for example by avoiding large matrix inversions.

One essential feature of this procedure is that the forms fall out of the free-form optimization automatically, rather than assuming a particular (e.g. Gaussian) distribution for all components. In order to assume that the factors have tractable forms, the priors also need to be picked appropriately to ensure that the updates are tractable (i.e. they are conjugate to the likelihood).

Crucially, this factorization means that expectations involving several factors can be analytically simplified to expectations over individual factors. For example, if model parameters $\mathbf{A}$ and $\mathbf{B}$ have separate corresponding factors $\mathrm{P}'(\mathbf{A})$ and $\mathrm{P}'(\mathbf{B})$, the expectation

$$\begin{aligned}\int \mathbf{AB}\ \mathrm{P}'(\boldsymbol{\Theta})\,\mathrm{d}\boldsymbol{\Theta} &= \int \mathbf{AB}\ \mathrm{P}'(\mathbf{B})\,\mathrm{P}'(\mathbf{A})\,\mathrm{d}\mathbf{B}\ \mathrm{d}\mathbf{A} \\ &= \int \mathbf{A}\left(\int \mathbf{B}\ \mathrm{P}'(\mathbf{B})\,\mathrm{d}\mathbf{B}\right)\mathrm{P}'(\mathbf{A})\,\mathrm{d}\mathbf{A} \\ &= \left(\int \mathbf{B}\ \mathrm{P}'(\mathbf{B})\,\mathrm{d}\mathbf{B}\right)\left(\int \mathbf{A}\ \mathrm{P}'(\mathbf{A})\,\mathrm{d}\mathbf{A}\right) \qquad (2.15)\end{aligned}$$

or in expectation form

$$\begin{aligned}\langle \mathbf{AB}\rangle_{\mathrm{P}'(\boldsymbol{\Theta})} &= \langle \mathbf{AB}\rangle_{\mathrm{P}'(\mathbf{A})\mathrm{P}'(\mathbf{B})} \\ &= \left\langle \mathbf{A}\,\langle \mathbf{B}\rangle_{\mathrm{P}'(\mathbf{B})}\right\rangle_{\mathrm{P}'(\mathbf{A})} \\ &= \langle \mathbf{A}\rangle_{\mathrm{P}'(\mathbf{A})}\,\langle \mathbf{B}\rangle_{\mathrm{P}'(\mathbf{B})} \qquad (2.16)\end{aligned}$$

which would not be possible without the assumption that $\mathrm{P}'(\mathbf{A},\mathbf{B}) = \mathrm{P}'(\mathbf{A})\,\mathrm{P}'(\mathbf{B})$. Since the posterior factors $\mathrm{P}'(\mathbf{A})$ and $\mathrm{P}'(\mathbf{B})$ have algebraic forms, these individual expectations are usually trivial.

2.2.2 Updates

The VB optimization procedure proceeds by a functional optimization of each factor in turn, finding the distribution $\mathrm{P}'(\boldsymbol{\Theta}_i)$ that maximizes F while holding all of the other factors fixed.

The free energy can be expressed again as

$$\begin{aligned}F =& \langle \log \mathrm{P}(\mathcal{D}|\boldsymbol{\Theta}) + \log \mathrm{P}(\boldsymbol{\Theta}) - \log \mathrm{P}'(\boldsymbol{\Theta})\rangle_{\mathrm{P}'(\boldsymbol{\Theta})} \\ =& \langle \log \mathrm{P}(\mathcal{D}|\boldsymbol{\Theta})\rangle_{\prod_i \mathrm{P}'(\boldsymbol{\Theta}_i)} + \sum_i \langle \log \mathrm{P}(\boldsymbol{\Theta}_i)\rangle_{\mathrm{P}'(\boldsymbol{\Theta}_i)} - \sum_i \langle \log \mathrm{P}'(\boldsymbol{\Theta}_i)\rangle_{\mathrm{P}'(\boldsymbol{\Theta}_i)}\,. \qquad (2.17)\end{aligned}$$

As a functional of a single factor, this is can be written as

$$F(\mathrm{P}'(\boldsymbol{\Theta}_i)) = \langle \log \mathrm{P}(\mathcal{D}|\boldsymbol{\Theta})\rangle_{\prod_{j\neq i} \mathrm{P}'(\boldsymbol{\Theta}_j)} + \log \mathrm{P}(\boldsymbol{\Theta}_i) - \log \mathrm{P}'(\boldsymbol{\Theta}_i) + \kappa \qquad (2.18)$$

where κ is a constant that does not involve $\mathbf{\Theta}_i$. Optimization of $F(\mathrm{P}'(\mathbf{\Theta}_i))$ is achieved by ensuring that its derivative is zero with respect to any small change in $\mathbf{\Theta}_i$. The free-form optimization ensures that $F(\mathrm{P}'(\mathbf{\Theta}_i))$ is constant regardless of the value of $\mathbf{\Theta}_i$, by matching the terms involving $\mathbf{\Theta}_i$ in the update equation

$$\log \mathrm{P}'(\mathbf{\Theta}_i) \quad = \quad \langle \log \mathrm{P}(\mathcal{D}|\mathbf{\Theta}) \rangle_{\prod_{j \neq i} \mathrm{P}'(\mathbf{\Theta}_j)} + \log \mathrm{P}(\mathbf{\Theta}_i) \tag{2.19}$$

therefore the likelihood and the prior determine the form of of the factorized posterior. Note that this places a restriction on the form of the priors that can be used; they must be conjugate to the likelihood function. While the parametric form of the posterior distribution is fixed by the likelihood function and the prior, the values will depend on the data and the other posterior factors. Therefore the update procedure actually involves iteratively updating the *parameters* of the posterior based on algebraic expressions.

By iterating over all of the updates, F will increase, bounded above by the true model evidence $\mathrm{P}(\mathcal{D})$. This means that convergence to at least a local maximum is guaranteed.

From now on, the subscripts on the expectations will be omitted and should be clear from the context.

2.2.3 Calculating the Free Energy

The free energy F is given in equation 2.12. Expressed in expectation notation, this is

$$F \quad = \quad \langle \log \mathrm{P}(\mathcal{D}|\mathbf{\Theta}) \rangle + \langle \log \mathrm{P}(\mathbf{\Theta}) \rangle - \langle \log \mathrm{P}'(\mathbf{\Theta}) \rangle \tag{2.20}$$

where all of these expectations are over $\mathrm{P}'(\mathbf{\Theta})$, i.e. all posterior factors. The first term is recognizable as the expected log likelihood, and assesses the accuracy of the fit to the data. The last two can be combined to form the KL divergence between

the prior and the factorized posterior:

$$\begin{aligned}\langle \log \mathrm{P}(\boldsymbol{\Theta})\rangle - \langle \log \mathrm{P}'(\boldsymbol{\Theta})\rangle &= -\sum_i \left[\left\langle \log \mathrm{P}'(\boldsymbol{\Theta}_i)\right\rangle - \left\langle \log \mathrm{P}(\boldsymbol{\Theta}_i)\right\rangle\right] \\ &= -\sum_i \mathrm{KL}(\mathrm{P}'(\boldsymbol{\Theta}_i)\,||\mathrm{P}(\boldsymbol{\Theta}_i))\,. \end{aligned} \tag{2.21}$$

This acts as a penalty for complexity on each parameter. Interestingly, there is no penalty for including an additional variable if it is not affected by the data, i.e. if its posterior distribution is identical to its prior.

2.3 Other Inference Approaches

There are several other ways to find the moments and normalization constant of equation 2.2. Three other commonly-used approaches are the Laplace approximation, evidence optimization and Markov chain Monte Carlo methods.

The **Laplace approximation** is an alternative approach for approximate Bayesian inference on continuous variables. In this approach, a numerical optimization method is used to find a mode $\boldsymbol{\Theta_0}$ of the posterior distribution, where the gradient vanishes

$$\nabla\, \mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})|_{\boldsymbol{\Theta}=\boldsymbol{\Theta_0}} = \mathbf{0} \tag{2.22}$$

and then the log-posterior is approximated by the second-order Taylor expansion

$$\log \mathrm{P}(\boldsymbol{\Theta}|\mathcal{D}) \approx \log \mathrm{P}(\boldsymbol{\Theta}_0|\mathcal{D}) - \frac{1}{2}(\boldsymbol{\Theta}-\boldsymbol{\Theta_0})^{\mathrm{T}}\mathbf{A}(\boldsymbol{\Theta}-\boldsymbol{\Theta_0}) \tag{2.23}$$

with $\mathbf{A} = \nabla\nabla \mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})|_{\boldsymbol{\Theta}=\boldsymbol{\Theta_0}}$ denoting the Hessian matrix. Renormalizing this yields a Gaussian approximation to the true posterior distribution

$$\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D}) \approx \mathrm{N}\big(\boldsymbol{\Theta}|\boldsymbol{\Theta_0}, \mathbf{A}^{-1}\big)\,. \tag{2.24}$$

The result is an approximation has the correct mode and local curvature, but may fail to capture the global properties of the distribution (Bishop, 2006, sec.4.4). This is

in contrast to VB, which produces a global approximation by matching the sufficient statistics of the approximate posterior (e.g. the mean and variance) to that of the true (marginal) posterior. Unlike VB, this does not provide a rigorous lower bound on the evidence, so model comparison is not valid. However, the simplicity of the Laplace approximation makes it easy to apply in situations where the necessary integrals for VB are intractible, since it depends only on the posterior having a defined second derivative rather than being able to integrate over the marginal posterior.

Evidence optimization (EO) provides a way to perform Bayesian inference on a model that *would* be analytic if only its hyperparameters were fixed. Integrating out everything except for the hyperparameters yields an expression for the model evidence conditioned on the hyperparameters:

$$\begin{aligned} \mathrm{P}(\boldsymbol{\alpha}|\mathcal{D}) &\propto \mathrm{P}(\mathcal{D}|\boldsymbol{\alpha})\,\mathrm{P}(\boldsymbol{\alpha}) && (2.25) \\ \mathrm{P}(\mathcal{D}|\boldsymbol{\alpha}) &= \int \mathrm{P}(\mathcal{D}|\mathbf{X},\boldsymbol{\alpha})\,\mathrm{P}(\mathbf{X}|\boldsymbol{\alpha})\,\mathrm{d}\mathbf{X} && (2.26) \end{aligned}$$

If the integral produces an analytic function of $\boldsymbol{\alpha}$, this can be maximized to obtain a maximum *a posteriori* estimate of the hyperparameters (Rasmussen and Williams, 2006, Ch. 5). Conditioned on that point estimate of $\boldsymbol{\alpha}$, the posterior distribution of the parameters $\mathbf{X}$ can also be obtained analytically. This is particularly useful for inference on Gaussian process priors because the functional form of the hyperparameters can be very complex but once these are fixed, the model is just a single multivariate normal distribution across all of the data points (Rasmussen and Williams, 2006). In practice, some of the regular parameters in the likelihood (e.g. noise variances) may also need to have optimized point estimates, increasing the dimensionality of the optimization space slightly.

Rather than trying to approximate the posterior analytically, **Markov chain Monte Carlo** (MCMC) techniques work by *sampling* from the true posterior (Bishop, 2006, Ch. 11). It is easy to estimate any required moments directly from the

samples, and the accuracy depends on the number of *independent* samples collected. MCMC approaches are designed to provide efficient integration in high-dimensional space, because they spend most of their computational time in regions of the highest posterior probability.

Since the unnormalized posterior distribution (eqn. 2.2) can be evaluated at any point, it is possible to construct a random walk through parameter space such that the amount of time spent at each point is proportional to the posterior probability at that point. Collecting samples from this random walk will eventually reconstruct the true posterior distribution in any level of detail required. Unfortunately, the number of steps required to achieve a satisfactory approximation can be very long because adjacent samples are highly correlated to one another; it may take a great many steps to get the equivalent of one independent sample. Furthermore, it can be difficult to predict how long this "mixing time" is, although there are empirical measures to estimate this. To avoid the transient effects of the initialization, a "burn-in" period, often the first 25-50% of the samples, is discarded.

Although MCMC techniques are generally much slower than VB, they do eventually approach the true distribution given enough computation time, while VB hits a peak (given by the KL divergence) due to the mean field approximation. For models that are tractable using VB, MCMC is still useful to for checking whether these assumptions change the posterior distributions in any significant way.

Chapter 3

Nonlinear Fitting with Correlated Noise Models

3.1 Introduction

The General Linear Model (GLM) is a powerful and widely-used method for model-based inference on BOLD FMRI[1]. Given a set of model-derived timecourses known as Explanatory Variables (EVs), the GLM finds the linear combination of these that most closely approximates each voxel's time-course. It also provides a way to assess the statistical significance to individual EVs as well as contrasts and combinations of them. In BOLD analysis, the emphasis is usually on the significance of the activation (expressed as Z-statistics) rather than actual parameter estimates. The GLM can also be applied to ASL data, but the fact that the pulse sequence is not exactly the same at all timepoints means that certain preprocessing is required, either by pre-subtracting adjacent tag/control volumes or by adding regressors to model the tag/control properties of the scan (Mumford et al., 2006).

However, a linear model of the data is not an appropriate approximation for all applications. Nonlinear generative models are increasingly necessary for dealing with complex pulse sequences and for multimodal imaging (with more than one type of FMRI contrast). New multimodal pulse sequences, such as those explored in

[1]This is not to be confused with the more flexible General*ized* Linear Model (Nelder and Wedderburn, 1972), which is what the abbreviation GLM usually means in statistics.

this chapter and the next, can contain several signals mixed together in a nonlinear fashion.

In addition, physiological models of brain haemodynamics are normally non-linear, for example, the balloon model developed by Buxton et al. (1998). There is an increasing interest using these techniques for quantitative FMRI, which promises more repeatable activity estimates by relating observed signals back to physiologically-meaningful changes in oxygen metabolism and flow (Hoge et al., 1999). These non-linear models are often approximated by a set of linear regressors and the parameters of interest are calculated post-hoc from the GLM parameter estimates.

This chapter investigates a particular pulse sequence, "dual-echo ASL", which acquires both ASL-optimized and BOLD-optimized images at almost the same time. These two sequences of contrast-optimized images can be thought of as two parallel streams. Inaccuracies in the linearized version of the model can also cause the signals embedded in the two streams to become somewhat contaminated by each other and by other effects. Using a full nonlinear model has previously been shown to correct for some of this; see figure 3.1. Furthermore, the linear models in this case can only be applied to one stream of the data at a time, while the nonlinear models fit the whole data set at once. Multimodal, simultaneously-acquired data contains mixtures of different signals and therefore may have common noise sources. If this noise correlation is not allowed in the Bayesian model, then uncertainty estimates can be biased.

The approach from Chappell et al. (2009) is applied to infer on this model using variational Bayes and a local linear approximation. The effect of this approximation on the results is evaluated on real data against an existing MCMC implementation. This chapter also looks at the application of this approach to the analysis of the dual-echo ASL pulse sequence, and assesses the differences between this approach and existing GLM-based approaches. The biases that result from not modelling the

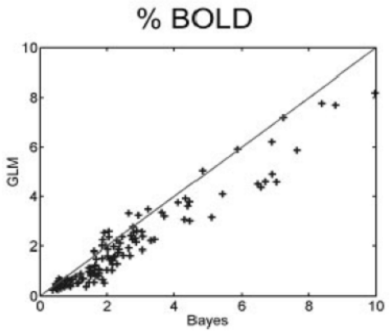


Figure 3.1: An example of the differences between nonlinear and linear estimates of the BOLD effect, reproduced from Woolrich et al. (2006) with permission. The "Bayes" method (x-axis) uses a nonlinear model to infer from the whole data set, while the GLM approach uses a linear model and estimates BOLD using only the long-echo-time (TE_2) data. In the nonlinear Bayesian approach the estimates of BOLD are more accurate and about 20% higher than the GLM. This is because the GLM approach is inadequately modelling effects due to blood flow and volume changes, which are reducing the apparent BOLD signal.

correlated noise are also examined.

3.2 Bayesian Inference on Nonlinear Signal Models

This chapter presents a voxelwise time-series analysis method to estimate the unknown "signal parameters" $\mathbf{w} \in \mathscr{R}^{K\times 1}$ in each voxel, where the number of signal parameters is given by K. Using the details of the pulse sequence and applying models of the MR physics and tissue physiology, a generative model $\mathbf{g} : \mathscr{R}^{K\times 1} \rightarrow \mathscr{R}^{T\times 1}$ produces a prediction of what the (noise-free) time-series (of length T) should be, given any set of signal parameters. The measured signal $\mathbf{y} \in \mathscr{R}^{T\times 1}$ in that voxel is then expressed as

$$\mathbf{y} \;=\; \mathbf{g}(\mathbf{w}) + \mathbf{e}, \tag{3.1}$$

where $\mathbf{e}$ is an additive noise term. This can be thought of as an extension to the Bayesian GLM framework of Penny et al. (2003) to fit general nonlinear signal models and applies it to a particular biophysical model of the dual-echo ASL pulse sequence.

Inferring on the resulting nonlinear Bayesian model is relatively straightforward when using sampling methods such as MCMC, but fast approximate techniques such as Variational Bayes rely on integrating over the likelihood function to obtain a set of algebraic update equations. These integrals are only possible when the likelihood has a conjugate form, which generally requires that $\mathbf{g}$ be linear. A local linear approximation $\overline{\mathbf{g}}$ is used instead of $\mathbf{g}$.

Penny et al. (2003) have developed a framework for Bayesian inference on FMRI time series using linear models. The EVs are placed in the columns of $\mathbf{X}$ and the signal is modelled as:

$$\mathbf{y} \;=\; \mathbf{X}\mathbf{w} + \mathbf{e} \tag{3.2}$$

with an autoregressive noise model on $\mathbf{e}$. For nonlinear functions, analytic expressions generally do not exist, so a linear approximation is used: $\overline{\mathbf{g}}(\mathbf{w}) \approx \mathbf{g}(\mathbf{w})$. This is the first-order Taylor expansion of the nonlinear function about a point $\mathbf{w} = \mathbf{z}$:

$$\overline{\mathbf{g}}(\mathbf{w})\big|_{\mathbf{z}} \;=\; \mathbf{g}(\mathbf{z}) + \mathbf{J}(\mathbf{w} - \mathbf{z}) \tag{3.3}$$

where $\mathbf{J} = \left.\frac{\partial \mathbf{g}(\mathbf{w})}{d\mathbf{w}}\right|_{\mathbf{z}}$ is the $T \times K$ Jacobian matrix. This can be calculated either analytically, but this would require additional derivations (and potentially introduce errors) every time a new nonlinear model $\mathbf{g}$ is implemented. To make the software as flexible and easy to extend as possible, the Jacobian is calculated numerically from $\mathbf{g}$ itself, using a step size of 10^{-5} of the absolute parameter value, or 10^{-10}, whichever is larger. After every iteration, the Taylor expansion is re-centred so that $\mathbf{z}$ coincides with the mean of the estimated $\mathbf{w}$ distribution. This ensures that the approximation is most accurate where the probability density is highest; this local approximation becomes more accurate as the confidence in the posterior estimate of $\mathbf{w}$ increases.

This approach to using VB on nonlinear models has been successfully used before for many applications, including EEG/MEG analysis (Kiebel et al., 2008) and general dynamical systems (Friston et al., 2008b). A similar approach is to use the full

nonlinear model and a Laplace approximation to the factorized posterior (Friston et al., 2007). One caveat of these methods is that changing the linearization effectively changes the problem with every iteration, so (as with most nonlinear optimization problems) convergence is not strictly guaranteed. In practice, however, sensible choices of parameterization for $\mathbf{g}$ and a reasonable initial guess mean that convergence can be quite reliable.

3.2.1 Noise Models

To complete the likelihood function, a model of the noise vector $\mathbf{e}$ is needed. (Penny et al., 2003) derived the VB updates for a general autoregressive (AR) noise model in FMRI. In this chapter the first-order autoregressive noise model is used for simplicity; this is sufficient in most voxels in standard FMRI data (Woolrich et al., 2001). Applying this AR(1) noise model to each of the streams separately yields

$$e_t^{(i)} = a_i e_{t-1}^{(i)} + \epsilon_t^{(i)}, \qquad \epsilon_t^{(i)} \sim \mathrm{N}\big(0, \phi_i^{-1}\big) \tag{3.4}$$

where $e_t^{(i)}$ is the additive noise in volume t from echo time i, and $\{a_i\}$ and $\{\phi_i\}$ are additional (non-signal) parameters to be inferred from the data. This assumes that the noise processes on the two echo times are independent.

This assumption does not cause problems if each stream is analysed separately, but if the noise processes are not independent this can cause problems for simultaneous modelling of the whole data set. The approach used in this chapter (where there are two streams) is to introduce adaptive correlation terms into the existing AR noise model by adding two additional "crossover" parameters, a_{12} and a_{21}:

$$\begin{aligned} e_t^{(1)} &= a_1 e_{t-1}^{(1)} + a_{12} e_t^{(2)} + \epsilon_t^{(1)}, \qquad \epsilon_t^{(1)} \sim \mathrm{N}\big(0, \phi_1^{-1}\big) \\ e_t^{(2)} &= a_2 e_{t-1}^{(2)} + a_{21} e_t^{(1)} + \epsilon_t^{(2)}, \qquad \epsilon_t^{(2)} \sim \mathrm{N}\big(0, \phi_2^{-1}\big) \end{aligned} \tag{3.5}$$

or in matrix form

$$\mathbf{e}_t = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} \mathbf{e}_{t-1} + \begin{bmatrix} 0 & a_{12} \\ a_{21} & 0 \end{bmatrix} \mathbf{e}_t + \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \phi_1^{-1} & 0 \\ 0 & \phi_2^{-1} \end{bmatrix}\right). \tag{3.6}$$

where $\boldsymbol{\epsilon}_t = \begin{bmatrix} \varepsilon_t^{(1)} \\ \varepsilon_t^{(2)} \end{bmatrix}$ and $\mathbf{e}_t = \begin{bmatrix} e_t^{(1)} \\ e_t^{(2)} \end{bmatrix}$. Notice that there is no time delay in the crossover terms because a large portion of the shared noise is expected to be instantaneous. Although this looks ill-defined, it is easy to find the likelihood since this only involves solving for $\boldsymbol{\epsilon}_t$ in terms of $\mathbf{e}_t$, thus providing a reasonably flexible correlated white noise model. The reason for this is equation 3.6 can be rearranged as:

$$\begin{bmatrix} 1 & -a_{12} \\ -a_{21} & 1 \end{bmatrix} \mathbf{e}_t = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} \mathbf{e}_{t-1} + \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \phi_1^{-1} & 0 \\ 0 & \phi_2^{-1} \end{bmatrix}\right) \tag{3.7}$$

$$\mathbf{e}_t = \frac{\begin{bmatrix} a_1 & a_2 a_{12} \\ a_1 a_{21} & a_2 \end{bmatrix}}{1 - a_{12}a_{21}} \mathbf{e}_{t-1} + \frac{\begin{bmatrix} 1 & a_{12} \\ a_{21} & 1 \end{bmatrix}}{1 - a_{12}a_{21}} \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \phi_1^{-1} & 0 \\ 0 & \phi_2^{-1} \end{bmatrix}\right) \tag{3.8}$$

so that some covariance between the noise sources can be modelled by inferring appropriate values for a_{12} and a_{21} (although the scale of ϕ_1 and ϕ_2 will also have to change due to the $1 - a_{12}a_{21}$ denominator). The AR(1) model is intact and is inferred by a_1 and a_2, with similar scale adjustments required. All of these parameters can still be inferred automatically by the model without any significant changes to the VB updates, although the meaning of these noise parameters is not as interpretable as before. The same $\mathrm{N}(0, 10^4)$ prior is used for these additional parameters and, as before, the values for which the noise model becomes unstable (in this case, $a_{12}a_{21} \approx 1$) are automatically avoided because they provide a poor model of real noise, in the same way that $|a_i| > 1$ are avoided in a normal AR(1) noise model.

A cleaner noise model would be to have a correlated white noise source with independent AR(1) noise evolution for each stream:

$$e_t^{(i)} = a_i e_{t-1}^{(i)} + \epsilon_t^{(i)}, \qquad \text{for all } i \tag{3.9}$$

$$\boldsymbol{\epsilon}_t \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Phi}^{-1}\right). \tag{3.10}$$

where $\boldsymbol{\Phi}^{-1}$ is any positive-definite matrix. Among other things, this would generalize better because it has half as many correlation parameters than the previous model.

This is not implemented here because of the changes it would have required to the VB update equations, including a Wishart posterior distribution for $\mathbf{\Phi}$ (however this type of correlated noise is implemented in later work, section 5.2.3).

3.2.2 VB Implementation

For VB inference, the posterior is approximated by the factorized form:

$$\mathrm{P}(\mathbf{w}, a, \phi|\mathbf{y}) \approx \mathrm{P}'(\mathbf{w})\,\mathrm{P}'(\mathbf{a})\,\mathrm{P}'(\boldsymbol{\phi}) \tag{3.11}$$

resulting in update equations that converge to give probabilistic estimates of all of these (signal and non-signal) parameters. For linear models, VB update equations for $\mathbf{w}$, ϕ and a are derived by Penny et al. (2003). The likelihood requires conjugate priors, so $\mathrm{P}(\mathbf{w})$ and $\mathrm{P}(a)$ are taken to be multivariate normal distributions and $\mathrm{P}(\phi)$ is Gamma-distributed. The posterior distributions on these parameters have the same form as these priors.

Note that the AR(1) parameter is only a stable generative model for $-1 < a < 1$ (and for FMRI, $0 \leq a < 1$ is expected). The VB framework requires a Gaussian prior on a so there is no way to restrict it to a given range. In fact, to avoid biasing the value of a, an essentially uninformative $\mathrm{N}(0, 10^4)$ prior is used. In practice this is not a problem because a is well-determined by the data and values of a outside the valid range do not reflect the observed (stable) data. An uninformative prior is also used on the noise precision: $\mathrm{P}(\phi_i) = \mathrm{Ga}(10^6, 10^{-6})$.

The VB updates are nearly identical to those derived in (Penny et al., 2003), with the only differences being that the linearization-dependent Jacobian matrix $\mathbf{J}$ replaces the fixed regressors $\mathbf{X}$, and that separate noise precisions ϕ_i are inferred for timepoints corresponding to each echo time. In particular the noise correlation terms are easily modelled in the standard "embedding matrix" AR framework used by Penny *et al.*, by simply appending them to that matrix as if they were higher-order AR components.

3.3 Application to Dual-echo ASL Data

This nonlinear VB framework is used to analyse FMRI data from a dual-echo ASL pulse sequence. The biophysical model described in the next section was previously used to analyse these scans in an MCMC framework, producing some important differences in parameter estimates compared to linear modelling techniques (Woolrich et al., 2006). The long calculation time (hours per slice) however meant that it was somewhat inconvenient to use it as a practical analysis tool.

3.3.1 Dual-echo ASL Pulse Sequence

Most FMRI relies on the blood oxygen level dependent (BOLD) contrast. As blood oxygen levels change with nearby activity, there is a measurable change in R_2^* (spin dephasing rate) in the area. By allowing the spins to dephase for some time TE (the echo time) before acquiring each image, this effect can be seen as a brightening of the image in active areas. Generally a fairly long TE is used to maximize the BOLD contrast.

The standard ASL sequence involves "tagging" a thick slab of tissue in the neck with an inversion pulse and allowing the blood from there to flow into the slice of interest before acquiring the image (see section 1.2.2). These are interleaved with "control" images from which the tagging pulse is omitted (or shifted outside the brain). The systematic differences between these two sets of images can be attributed mainly to the cerebral blood flow (CBF) into the imaging slice from the tag region. To maximize signal, a very short TE should be used. ASL still has a much lower contrast-to-noise ratio than BOLD, but is more stable over time and is more quantitative. It also measures a different physical quantity, so simultaneous measurements are useful for calibrating physiological models of brain haemodynamics (Buxton et al., 1998).

Dual-echo ASL is identical to the standard ASL sequence except that it deals with these conflicting demands on the choice of TE by collecting two images a few tens of

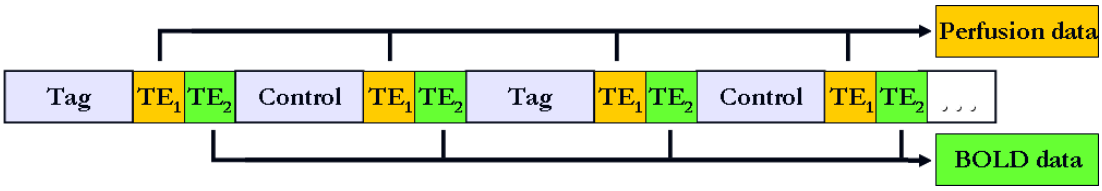


Figure 3.2: A rough sketch of the dual-echo ASL sequence. In each repeat, images are read out at two different echo times. Each repeat is either a tag or control repeat, which describes the different type of magnetization preparation that is used before the readout. The only difference between the tag and control repeats is in the magnetization of the blood that flows into the brain from the tagging region (in the neck).

milliseconds apart, where standard ASL would only collect a single, short TE. The data sets used here were collected using the QUIPSS II pulsed ASL (Luh et al., 2000) and dual spiral readouts at echo times of $TE_1 = 9.1$ ms and $TE_2 = 30$ ms. This is illustrated in figure 3.2.

The dual-echo ASL data was acquired on seven normal subjects using the dual-echo ASL pulse sequence on a single slice with a 3-second TR. A strong visual stimulus was used: an 8 Hz flashing radial checkerboard in blocks of 30 seconds off, 30 seconds on. These data sets are the ones previously analysed using an MCMC approach by Woolrich et al. (2006).

3.3.2 Dual-echo ASL Signal Model

Generative models of the signal can be constructed using pulse sequence information and a few simple assumptions about the underlying physiology.

A simple two-compartment model of the physiology was used, with no mixing and the same R_2^* in both compartments. In this model, there are three time-varying unknowns: the Cerebral Blood Flow (CBF), BOLD R_2^* decay rate, and the static magnetization. This static magnetization is simply the spins that are in the voxel at the time of tagging and remain there when the image is acquired. Since the nonlinear signal model $\mathbf{g}$ is a physical model it also contains several other constant values: the pulse sequence parameters (echo times TE_1 and TE_2, inversion times TI_1 and TI_2,

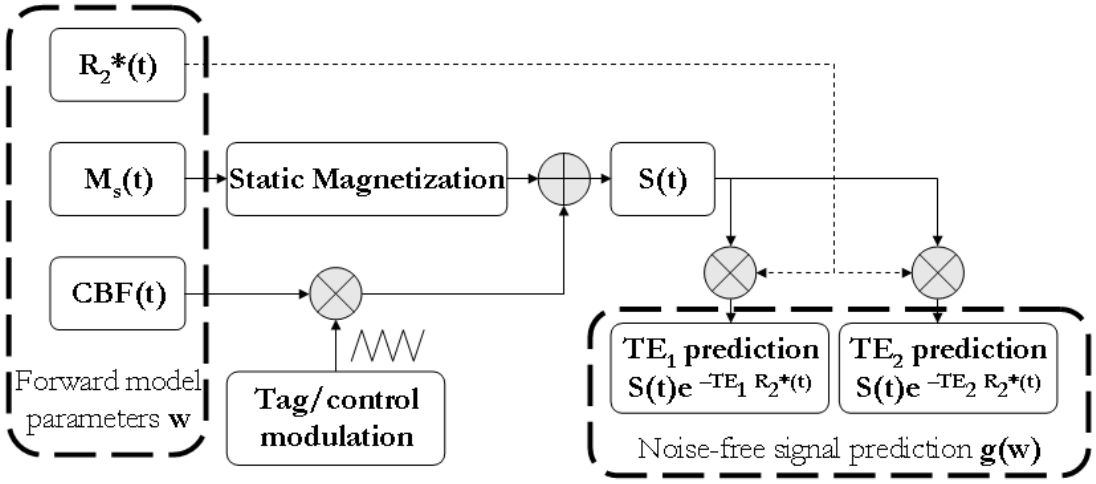


Figure 3.3: An overview of the dual-echo ASL signal model which predicts the voxel's timecourse $\mathbf{y} \approx \mathbf{g}(\mathbf{w})$ from its per-voxel parameters $\mathbf{w}$. "Static magnetization" is simply the magnetization from any material that are in the voxel at the time of tagging and remains there until the image is acquired; this appears to decrease during brain activity primarily due to an increase in the outflow of blood from a voxel due to increased CBF.

and the tag/control order) and the remaining physiological parameters (T_{1b} of blood and bolus arrival time Δt. The model is illustrated in figure 3.3, and a detailed derivation of this model can be found in Woolrich et al. (2006). Note that the single data vector $\mathbf{y}$ is made by interleaving data from the two echo times, so that the odd elements are from TE_1 and the even elements are from TE_2.

The assumption of independent noise for each echo time is suitable for GLM analysis because the TE_1 stream (ASL-optimized signal) and the TE_2 stream (BOLD-optimized signal) are analysed completely separately. In each of these streams the signal is predominantly from one contrast, but in practice there will be part of the other contrast contaminating the signal. The information to reverse this is in the data, and the contrasts are extracted more accurately by using streams together in a single nonlinear model (Woolrich et al., 2006).

The two images are acquired so close together that many types of noise could easily affect both images in a pair, especially because they both have the same

dependency on the magnetization $S(t)$ present at the time of the same excitation pulse (see figure 3.3 and equation A.6). Since ASL-related signal differences also occur before excitation, the relevant information from the two readings is probably nearly identical. Clearly, noise correlation is a feature of the data that needs to be modelled.

The basic characteristics of this sequence (TE_1, TE_2, tag/control order, etc.) are used in the model as constants, and some other biophysical parameters (e.g. the T_1 of blood is fixed at 1.66 sec, and the ASL inversion efficiency α is assumed to be 1).

The model from Woolrich et al. (2006) is summarized in appendix A.

3.3.3 Choice of Parameterization

The model has so far been defined in terms of three time-varying quantities. These are actually parameterized as a baseline and a stimulus-related activation, so that $CBF(t)$ and $M(t)$ are given by the vectors

$$\mathbf{CBF}(Q_0, Q_1) = Q_0 + Q_1\mathbf{X_Q} \tag{3.12}$$

$$\mathbf{M_s}(M_0, M_1) = M_0 - M_1\mathbf{X_M} \tag{3.13}$$

where the $\mathbf{X}$ vectors are standard regressors generated by convolving the stimulus timings with an HRF. Only a single stimulus is used for the data in this chapter, but this could easily be extended to several stimuli. The regressor shapes are derived from the experimental design, and are treated as constants. It would be possible to use a canonical HRF which would mean $\mathbf{X_Q} = \mathbf{X_M} = \mathbf{X_R}$, but for greater accuracy the three HRFs used here are derived separately from an ROI analysis of the same data sets by Woolrich et al. (2006). Note that brain activation tends to increase BOLD and CBF signal but *decreases* static magnetization, so a minus sign is used in equation 3.13 so that positive M_1 denotes activation.

The expression for $R_2^*(t)$ is defined somewhat differently from the other time-varying parameters:

$$\mathbf{R_2^*}(R_0, R_1) \;=\; -\frac{1}{TE_2}\log\left(R_1\mathbf{X_R} + \mathrm{e}^{-TE_2 R_0}\right). \tag{3.14}$$

This is because the nonlinearity of the signal model is particularly severe in $R_2^*(t)$ because this parameter is exponentiated to calculate $\mathbf{g}(\mathbf{w})$. The total magnetization $S(t)$ is modulated by the signal decay function $\mathrm{e}^{-TE_i R_2^*(t)}$. The parameterization of equation 3.14 is more linear with respect to the data because (from equation A.6) it expands to

$$\left(\mathbf{g}(\mathbf{w})\right)_{i,t} \;=\; \mathbf{S}(Q_0, Q_1, M_0, M_1)\left(R_1\mathbf{X_R} + \mathrm{e}^{-TE_2 R_0}\right)^{TE_i/TE_2} \tag{3.15}$$

So that R_1 is perfectly linear with respect to the predicted signal when $i = 2$, which is the stream where the BOLD signal is strongest.

The parameters for the signal model $\mathbf{g}$ are therefore $\mathbf{w} = [Q_0, Q_1, M_0, M_1, R_0, R_1]^{\mathrm{T}}$. An uninformative $\mathrm{N}(0, 10^{12})$ prior is used for each element. All voxels were initialized to a reasonable set of initial values: $Q_0 = 200, M_0 = 15000, R_0 = 25, Q_1 = M_1 = R_1 = 0$.

3.4 Evaluation of VB Inference

On the dual-echo ASL model, the model converged reliably and very rapidly (within ten iterations). However this stability did depend on the parameterization. A simpler parameterization was also evaluated, which replaced equation 3.14 with $\mathbf{R_2^*}(R_0, R_1) = R_0 + R_1\mathbf{X_R}$. Under this parameterization, around 10% of voxels ended up making huge jumps through the parameter space and showed no signs of converging. This type of convergence problem is possible because VB is not aware of the underlying nonlinearity of the forward model. Every time the algorithm relinearizes about a new point, this essentially means that VB is working on a new linear problem. As a result of this, the convergence guarantees from linear VB no longer apply.

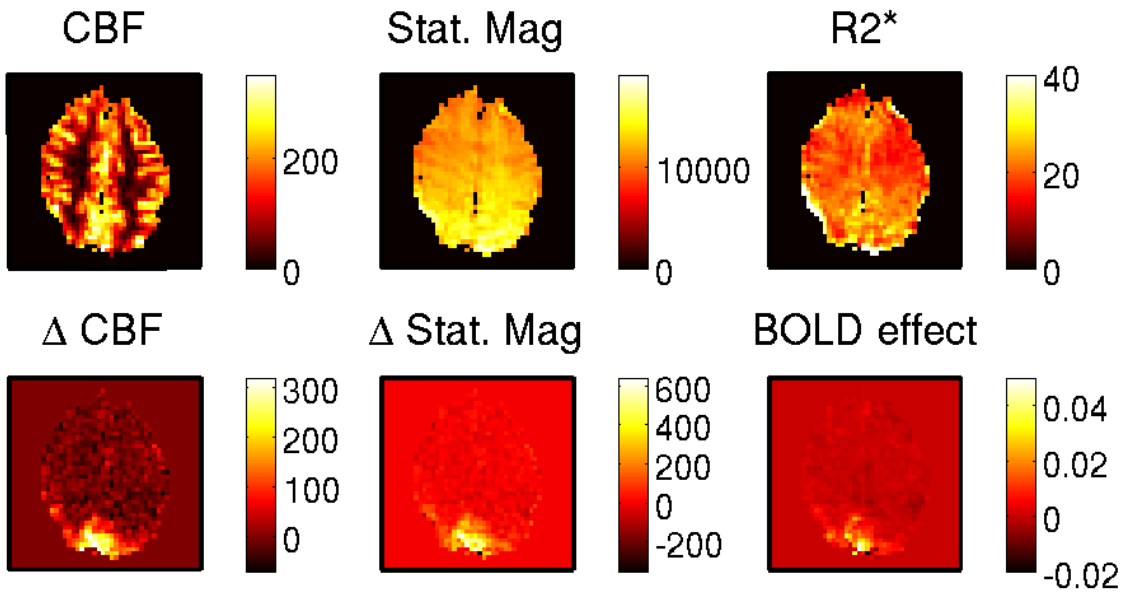


Figure 3.4: Maps of the main signal parameters (mean estimates) produced by VB inference on the dual-echo ASL data of one subject. Top row: baseline values. Bottom row: absolute CBF change, absolute static magnetization change, fractional BOLD effect change. All are in in the same (arbitrary) scanner units, except R_2^* (in 1/sec) and the BOLD effect (a fraction of TE_2 signal change from baseline).

For example, an update on $\mathrm{P}'(\mathbf{w})$ can jump directly from reasonable parameter values to implausibly large parameter values, if these jumps produce a good fit in the old linearization ($\mathbf{y} \approx \overline{\mathbf{g}}(\mathbf{w}_{new})|_{\mathbf{z}=\mathbf{w}_{old}}$). Depending on the nonlinearities on the model, these may become very poor fits after relinearization at the new location ($\mathbf{y} \not\approx \overline{\mathbf{g}}(\mathbf{w}_{new})|_{\mathbf{z}=\mathbf{w}_{new}} = \mathbf{g}(\mathbf{w}_{new})$). However if the nonlinearities are small enough, i.e. the Jacobian $\mathbf{J}$ and offset $\overline{\mathbf{g}}(0)$ do not change too much from one update to the next, then convergence can still occur in most cases.

Using the robust parameterization, the resulting mean parameter maps for a single subject are shown in figure 3.4. These show the baseline and activation-related change in each of the three time-varying quantities. Figure 3.5 shows the same activation information thresholded to show only significant activations ($Z > 3$). Note in particular the strong activation in the visual cortex (rear of the brain) visible in all three Z-statistic images (obtained from the posterior mean divided by standard

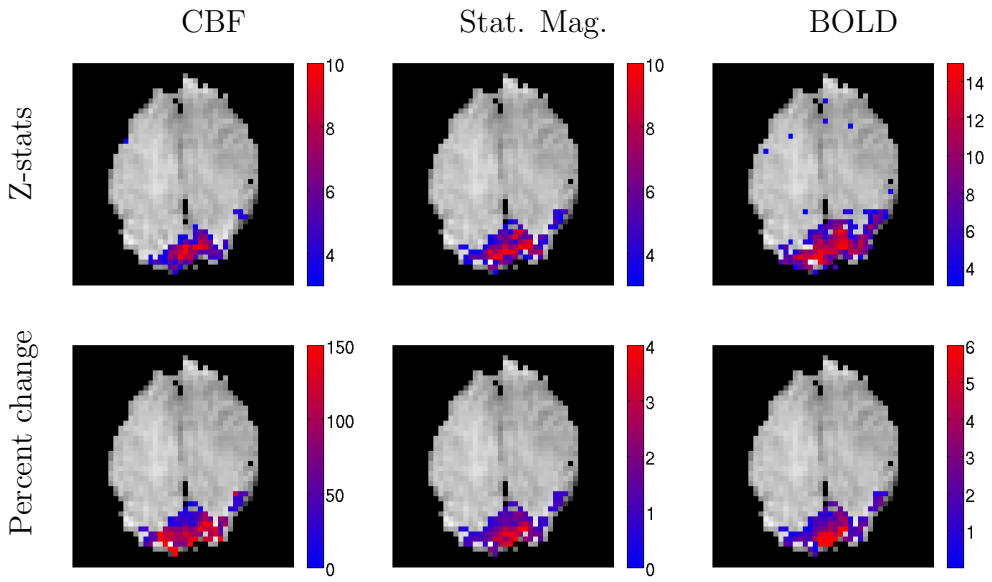


Figure 3.5: The threshold Z-stat maps ($Z > 3$) and thresholded percent change maps ($Z > 3$) from the three activation-related signal parameters, in a single subject.

deviation).

3.4.1 Evaluation of VB Results Against MCMC

To evaluate the validity of the VB approximations, the data sets were analysed using exactly the same Bayesian model but with Markov chain Monte Carlo (MCMC) inference. MCMC is a sampling method that does not make any approximations, but suffers from long (and hard to monitor) convergence times (MacKay, 2003). A 10-hour MCMC run on this dataset produced clean histograms and consistent results between runs, so this is used as the gold standard.

Figure 3.6 shows a comparison of the signal parameter estimates and Z-statistics for each voxel, across all seven subjects. The Z-stat images are very similar, although they do broaden somewhat for small Z values because many of these represent voxels with a large posterior variance on the parameter estimate. This means that the local linear approximation will be less accurate. In voxels with $Z > 3$, the estimates of stimulus-correlated change in CBF, BOLD, and static magnetization are also nearly

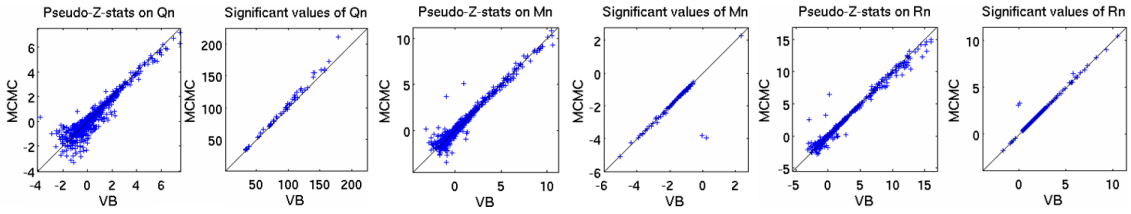


Figure 3.6: Comparisons of VB and MCMC. Each point corresponds to single voxel (across seven subjects). The Z-stat images show all voxels while the others only show values there $Z > 3$.

identical. This consistency is important for producing accurate quantitative estimates for calibrated FMRI.

The main benefit of using VB is that the inference can be performed much more quickly; these VB results took about 1 second per slice, while MCMC results took hours. This means that there is no real time penalty for Bayesian inference on a nonlinear model compared to a standard GLM. The C++ software was released as part of the FMRIB Software Library (FSL).

3.5 Evaluation of the Correlated Noise Model

Adding cross-terms makes the noise model more flexible. If these terms were not modelling noise correlation that is present in the data, then the posterior distributions on the signal parameters should remain mostly unchanged. Instead, figure 3.7 shows quite a striking change in the Z-statistics when the new noise model is used. CBF Z-stats drop considerably, while at the same time there is an increase in the BOLD Z-stats. This is possibly because the improved noise description means that CBF-like noise (where the TE_1 and TE_2 have positively correlated noise) more common in the data, while BOLD-like noise (with a negative noise correlation between the echo times) is less common. This demonstrates the importance of accurate noise modelling if Bayesian Z-statistics are to be trusted. Interestingly, the posterior means shown in the bottom half of figure 3.7 are mostly unchanged (apart from a few voxels), so the

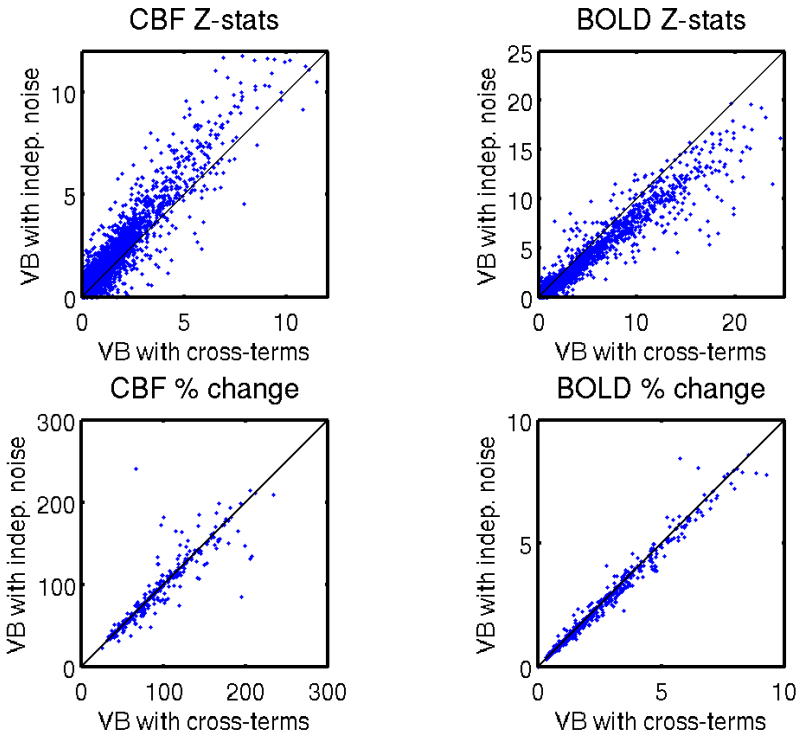


Figure 3.7: Changes in the VB results when additional noise correlation parameters are introduced: (upper) CBF percent-change Z-stats are reduced while BOLD-effect Z-stats increase. This suggests that the noise at the two echo times is correlated, and that the direction of correlation tends to look more like CBF signal than BOLD signal. Each point represents a single voxel, with the points from all seven subjects plotted on the same graph. (lower) The corresponding point estimates of the CBF %-change and BOLD effect size are unchanged in most voxels.

observed changes in Z-stats are due to increased or decreased uncertainty on these posterior estimates.

3.6 Comparisons to GLM results

It is also important to compare the new approach (using a nonlinear model and combining all the data) to the state-of-the-art *linear* ASL analysis methods (Mumford et al., 2006). These use an explicit linear model of the BOLD and ASL effects (figure 3.8). Both echo times provide some information about the BOLD and CBF changes, but the shorter echo time has higher ASL contrast and is used for CBF estimates,

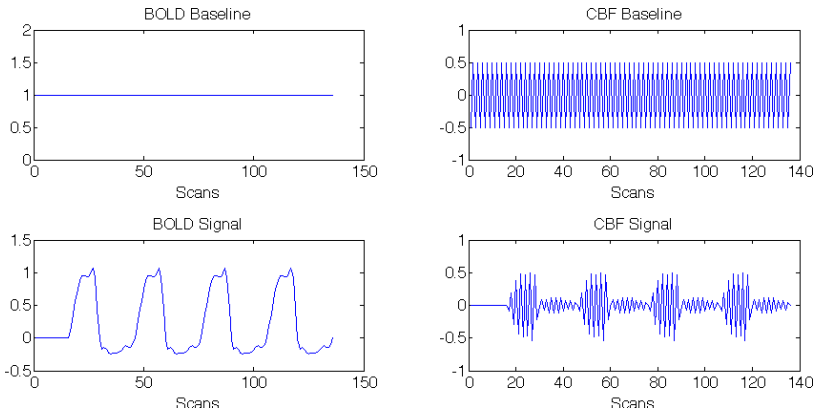


Figure 3.8: The four regressors in the linear analysis of dual-echo ASL data, as described by Mumford et al. (2006). Data from TE_1 and TE_2 are analysed separately, but all four of these regressors need to be modelled in both TE_1 and TE_2 data. The TE_2 has higher BOLD contrast so the BOLD estimates are extracted from this data set only, while the TE_1 data has higher CBF contrast so CBF is estimated from this data set only.

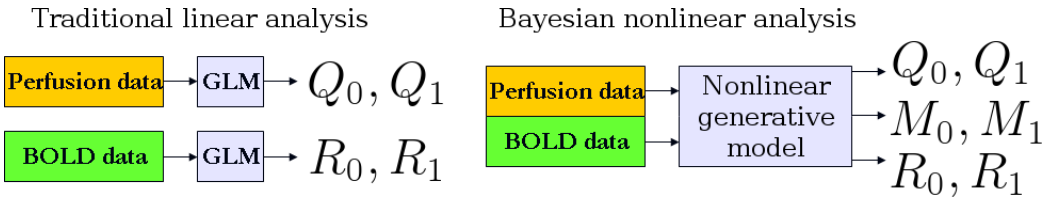


Figure 3.9: The GLM approach analyses the two echo times separately, while the nonlinear approach uses all of the data in a single model. This enables it to model the change in static magnetization, which is otherwise mixed into the BOLD effect.

while the longer TE has more T_2^* weighting and is therefore used to estimate the BOLD effect. The traditional linear inference method would use the data from the two echoes separately, with the CBF estimated from the TE_1 data and BOLD estimated from TE_2 data; see figure 3.9.

Analysing both echo times together is not possible in a standard GLM framework because the two echo times have different noise levels, require separate AR(1) modelling, and exhibit noise correlation. Furthermore, duplicate sets of regressors would be needed because there is no fixed linear scaling between the size of the BOLD effect (or CBF effect) at the two echo times.

Figure 3.10(top) reproduces one of the results of Woolrich et al. (2006): that the

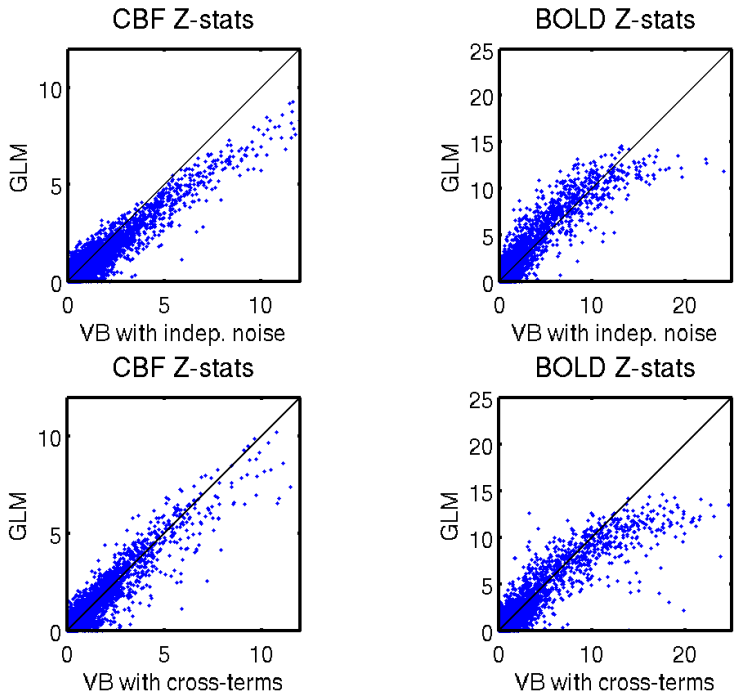


Figure 3.10: A comparison of CBF and BOLD Z-stats produced by the GLM (y axis) and the VB methods (x axis). In the top row the independent noise models are used for VB, and below the correlated noise model is used. This shows results for all 7 subjects combined; each point is a single voxel in one subject. Note that by modelling noise correlation, the apparent increase in CBF Z-statistics is reversed; this suggests that very little is gained by using data from both echo times rather than just TE_1 to estimate CBF. However, using both echo times still offers benefits in BOLD Z-statistics, especially when the noise correlation is correctly modelled; this suggests that using the data from both echo times is beneficial for estimating the rate of T_2^* decay. Note that the point estimates are virtually unchanged so these differences are due to the noise model changing the uncertainty estimates.

Bayesian approach produces higher CBF Z-stats than the GLM. This is primarily because the Bayesian model uses the entire data set to extract the ASL tag/control differences, rather than just using the short-TE data set. As would be expected, doubling the number of measurements yields more precise estimates. However, when the noise correlation is modelled in figure 3.10(bottom), this difference virtually disappears. This indicates that very little additional information about CBF is gained by using the TE_2 data, but more confidence in BOLD estimates is gained than previously realized. The point estimates are virtually unchanged.

3.7 Discussion

This chapter developed a tool for fast VB fitting of nonlinear models to simultaneously acquired data with common noise sources, and applied it to analysing data from a complex ASL-FMRI pulse sequence. This dual-echo ASL model was validated against MCMC results and it was found that the locally-linear VB approximation does not noticeably affect the estimates of signal parameters or their Z-statistics. By careful choice of the model parameterization, convergence problems were avoided.

Parameterization

Convergence problems were avoided by choosing a different parameterization of the model that is reasonably close to being linear in each parameter. This is not always an option, and further work is needed (on a variety of signal models) to find a robust and unbiased way to avoid these problems. In many cases this may be as simple as using a more robust non-linear optimization method for initialization.

A more general approach to improving robustness is to avoid drops in free energy and use more cautious updates for the signal parameters $\mathbf{w}$, based on the Levenberg-Marquardt (L-M) approach, when necessary (Chappell et al., 2009). However, drops in free energy are not always a problem in the nonlinear VB framework because the relinearization sometimes causes the free energy F to decrease in each iteration even as the VB updates are converging to a stable solution. As a result, these more robust approaches can sometimes lead to unwanted early stopping even in voxels where there are no convergence problems. However, large drops in F do generally indicate convergence problems, so both L-M and simpler heuristics were found to produce more plausible results on other ASL models. Furthermore, recent work by Friston et al. (2007) proposes a "temporal regularization" approach that may improve convergence in problematic models.

Noise correlation

The importance of accurately modelling correlated noise has been demonstrated in dual-echo ASL data. This correlated noise approach is a general approach that could be used anywhere where there may be a particularly close attachment between data from different streams. However, the correlation is likely to be particularly strong in dual-echo ASL because any noise that influences the magnetization $S(t)$ would affect both streams of data in highly-correlated ways.

The correlated noise model in equation 3.5 was motivated by ease of implementation but its parameters are not particularly interpretable. It introduces two new noise parameters (a_{12} and a_{21}) even though the main noise feature it was designed to model (correlation in the components of $\boldsymbol{\varepsilon}_t$) could be achieved using only a single extra parameter. A more interpretable noise correlation model (equation 3.10) could be implemented but there is no guarantee that it would be actually be a better model of real dual-echo ASL noise. They may in fact describe a fairly similar range of noise possibilities; it would be useful to create simulations with correlated noise and evaluate the quality of the fit using the proposed model, and vice-versa (using $a_{1,2,12,21}$ parameters derived from real data). Model comparison should also play a role in selecting which of these noise models is better, assuming the estimates of F are reliable despite the linearization.

Conclusions

The Bayesian framework automatically infers the value and uncertainty in all of the parameters, including noise parameters. The only errors in the uncertainty are due to the inaccuracy of the linear approximation to the nonlinear model (which will be minimal when the confidence in $\mathbf{w}$ is high) and some overestimation of confidence due to the VB factorization (between signal parameters and noise parameters). The comparisons made to MCMC suggest that these inaccuracies are very small.

This is more powerful than other nonlinear fitting methods (such as nonlinear least squares) even when they provide confidence estimates, because VB can automatically include uncertainty from non-signal model parameters, such as the autoregressive noise coefficients. More importantly, it can also include prior information such as informative priors and the adaptive spatial priors that will be discussed in the next chapter.

Chapter 4

Spatial Gaussian Process Priors

4.1 Introduction

Hierarchical Bayesian methods provide a flexible framework for the analysis of functional MRI and other MRI time-series data (Penny et al., 2003; Woolrich et al., 2006). In particular, the use of priors on signal parameters provides a principled approach to incorporating prior physical information into statistical inference (Friston, 2002). The "signal parameters" are analogous to the regression coefficients inferred by Penny et al. (2005), and the purpose of the analysis is to infer their true value at each voxel in the brain.

Two types of prior information are considered in this chapter: fixed non-spatial priors and spatial priors. Fixed non-spatial priors provide information on the plausible range of values a signal parameter could have, often based on a prior biophysical understanding of the underlying process. Since these are informative priors based on real *a priori* knowledge, it is important that they be included in the model to ensure that only sensible interpretations of the data are considered. They are not to be confused with global shrinkage priors, which provide regularization by automatically inferring the scale of each signal parameter from the data.

Another type of prior, the Laplacian spatial smoothness prior, explicitly encodes the belief that the parameter values in one voxel should not be dramatically different from those in its neighbours. It thereby provides spatial regularization which is similar

to spatial presmoothing, but it is adaptive on each signal parameter separately (Penny et al., 2005). If there is not enough information in the data to justify additional spatial detail, then a simpler, smoother parameter image is produced.

The model developed in this chapter is intended for applications in which a fixed biophysical prior provides useful information during inference, but spatial regularization is also desirable. There are many approaches in the literature for combining different types of prior within a Bayesian framework. These include many examples in neuroimaging applications:

- Model averaging, in which several independent models are evaluated and the posteriors are averaged based on global model evidence (Trujillo-Barreto et al., 2004). In the event that one model has much higher evidence than the others, it will dominate the posterior solution; otherwise the result is a weighted average of the separate results.

- Mixture models, in which the prior's covariance matrix is the weighted sum of several different types of covariance matrix, each with adaptively determined parameters (Mattout et al., 2006). For example, a signal parameter might have two sources of variance: a spatially-smooth component and an additive non-spatial component. The variance of each component would be determined from the data.

- Hierarchical models, in which a hierarchy of signal parameters is used, with each level having a generative model to predict the level below it, and each level introducing its own type of additive noise. This approach has been successfully used for the analysis of MEG data (Trujillo-Barreto et al., 2008; Sato et al., 2004), for example by having spatial regression coefficients, which predict the noisy spatio-temporal electrical activity, which predicts the noisy MEG data. The overall variance at the lowest level is effectively a sum of different types of

structured covariance matrix, with the weights determined by the magnitude of the "noise" added at each level.

Notably, none of these attempt to incorporate fixed biophysical prior information as part of the inference. Instead, they use adaptive spatial priors in addition to adaptive non-spatial priors as a way of modelling different types of variance in the observed signal.

The objective of this work is different; it aims to ensure that a fixed biophysical prior is respected, by conditioning the inference on the prior knowledge that the signal parameters must remain near a physically-plausible range of values. In technical terms, the marginal prior distribution in each voxel should always correspond to the fixed biophysical prior.

The proposed approach is based on Gaussian process priors (GPPs), which provide a flexible approach to spatial regularization. In particular, the GPP form proposed in this chapter was previously used in neuroimaging for analysing stimulus-response functions in neural recordings (Sahani and Linden, 2003) and for spatio-temporal analysis using functional distance metrics (Bowman, 2007). In these papers it is used to provide both spatial regularization and a global shrinkage prior. However, by fixing the global shrinkage parameter the GPP can instead be used to enforce the informative biophysical prior while also providing adaptive spatial regularization. The covariance matrix is parameterized to provide separate control over the variance (the diagonal elements) and the between-voxel correlation (due to off-diagonal elements). The mean and variance are set to match the fixed non-spatial prior, while the correlation between voxels drops exponentially with distance and is controlled by an adaptive correlation-distance parameter. This provides adaptive spatial regularization while ensuring the non-spatial prior still applies to each voxel.

This combined prior can be used as a drop-in replacement for the Laplacian spatial smoothness prior in the hierarchical FMRI model developed by Penny et al. (2005); in

fact, it is straightforward to use the combined prior on some signal parameters while using the Laplacian prior on others, within the same inference. That framework used variational Bayes (VB) for inference, which is an iterative approach that provides fully probabilistic results. A hybrid approach is developed for inference on the combined prior: VB estimates are used for inference on the noise parameters, while an empirical Bayes technique called evidence optimization (EO) is used to estimate the Gaussian process hyperparameters. EO also provides probabilistic estimates of the signal parameters, which are equivalent to the regression coefficients in a general linear model (GLM).

This technique is extended to work with non-linear modelling problems. Non-linear time-series models are beneficial in neuroimaging applications such as multi-contrast FMRI (Woolrich et al., 2006). This is particularly useful because parameters in non-linear models usually have real biophysical meanings, and as a result often have informative priors associated with them. This is achieved by incorporating the local linearization approach developed in Chappell et al. (2009) and used in the previous chapter.

This technique is evaluated using two applications in brain MRI. Both use non-linear forward models and have informative non-spatial priors on some parameters. In each application, the combined prior is compared against two alternatives: using the non-spatial information only, or using the spatial smoothness prior that ignores the informative non-spatial prior.

The first of these applications is FMRI analysis using constrained linear basis sets. The non-spatial prior allows the haemodynamic response function (HRF) to undergo reasonable variations in its shape in order to fit the data, while avoiding unrealistic HRF shapes (Woolrich et al., 2004). The second application is to multi-inversion-time ASL, where blood flow and bolus arrival time are estimated from resting-state data by fitting an intrinsically non-linear model of cerebral perfusion.

4.2 Voxelwise Models of MRI Time-series

Four-dimensional MRI data is usually analysed as a collection of V separate time series (one for each brain voxel). A generative model $\mathbf{g}()$ predicts the $T \times 1$ vector of time-series data $\mathbf{y}_v$ in voxel v given a $K \times 1$ vector of parameter values $\mathbf{w}_v$ in that voxel:

$$\mathbf{y}_v = \mathbf{g}(\mathbf{w}_v) + \mathbf{e}_v \tag{4.1}$$

where $\mathbf{e}_v$ is the additive noise in that voxel. The forward model $\mathbf{g}$ could be linear, as used in most previous work (Penny et al., 2003; Roberts and Penny, 2002), but could also be a general non-linear function (Friston, 2002; Woolrich et al., 2006).

In addition to the forward model parameters $\mathbf{w}_v$, each voxel has noise parameters. This is the standard AR(1) autoregressive noise model, which is commonly used for FMRI analysis (Woolrich et al., 2001; Penny et al., 2003):

$$e_v^{(t)} = a_v e_v^{(t-1)} + \epsilon_v^{(t)}, \qquad \epsilon_v^{(t)} \sim N(0, \phi_v^{-1}). \tag{4.2}$$

with noise precision (inverse variance) ϕ_v and AR parameter a_v. This can be condensed into a matrix form:

$$\mathbf{R}_v \mathbf{e}_v \sim N(\mathbf{0}, \phi_v^{-1}\mathbf{I}) \tag{4.3}$$

where $\mathbf{I}$ is the $(T-1) \times (T-1)$ identity matrix and the $\mathbf{R}_v$ is a T-1 $\times$ T bidiagonal matrix referred to as the "noise-whitening matrix":

$$\mathbf{R}_v = \begin{bmatrix} 1 & -a_v & & & \\ & 1 & -a_v & & \\ & & \ddots & \ddots & \\ & & & 1 & -a_v \end{bmatrix} \tag{4.4}$$

This is equivalent to the "embedded" form of AR used by Penny et al. (2003), but this matrix formulation is more convenient for these derivations. For a white noise model, simply use $\mathbf{R}_v = \mathbf{I}$, the identity matrix.

The forward model and noise model define the likelihood for voxel v:

$$P(\mathbf{y}_v|\mathbf{w}_v, \phi_v, a_v) = N(\mathbf{R}_v\mathbf{y}_v; \mathbf{R}_v\mathbf{g}(\mathbf{w}_v), \phi_v^{-1}\mathbf{I}). \tag{4.5}$$

4.3 Hierarchical Models for FMRI Time-series

Hierarchical Bayesian models provide a way to perform inferences on each voxel while maintaining some higher-level priors; these can help to pool information and provide regularization. In the hierarchical model, the voxelwise inferences are linked together by joint priors on the signal parameters, and a small number of hyperparameters are used to make this pooling adaptive.

All V voxels are inferred on simultaneously, so some new notation needs to be introduced to split the parameters in different ways. As before, $\mathbf{w}_v$ is a $K \times 1$ vector of the parameter values in voxel V. The $V \times 1$ vector $\mathbf{w}_k$ refers to the value of signal parameter k across all voxels. The "global" parameter vector, $\mathbf{w}_*$, is a $KV \times 1$ vector formed from the concatenation of all the $\mathbf{w}_k$, $k = 1 \ldots K$. Individual elements are referenced as $w_{k,v}$. These subscript notations will be used later to indicate other quantities that are per-voxel, per-parameter or global. Similarly, the voxelwise data vectors $\mathbf{y}_v$ are stacked to form $\mathbf{y}_*$, and $\mathbf{g}(\mathbf{w}_*)$ is used as shorthand for the concatenation of all $\mathbf{g}(\mathbf{w}_v)$ (i.e. it is the model prediction of $\mathbf{y}_*$). The vectors $\boldsymbol{\phi}$ and $\mathbf{a}$ are $V \times 1$ vectors of the scalars ϕ_v and a_v respectively. The $(T-1)V \times TV$ matrix $\mathbf{R}_*$ is defined as the block-diagonal concatenation of all $\mathbf{R}_v$, and $\boldsymbol{\Phi}_*$ as the block-diagonal concatenation of $\phi_v\mathbf{I}$.

Next, consider the joint prior controlling the signal parameters. To make efficient inference possible later, this is restricted to a multivariate normal distribution:

$$P(\mathbf{w}_*|\mathbf{C}_*) \quad = \quad N(\mathbf{w}_*; \boldsymbol{\mu}_*, \mathbf{C}_*) \tag{4.6}$$

Since $\boldsymbol{\mu}_k$ is a constant, it is left out of the probability notation. Furthermore, the problem is broken down by using a separate prior on each model parameter, in the

same manner as in Penny et al. (2005) and Trujillo-Barreto et al. (2008):

$$N(\mathbf{w}_*; \boldsymbol{\mu}_*, \mathbf{C}_*) \;=\; \prod_{k=1}^{K} N(\mathbf{w}_k; \boldsymbol{\mu}_k, \mathbf{C}_k) \tag{4.7}$$

so that $\boldsymbol{\mu}_*$ is the concatenation of all $\boldsymbol{\mu}_k$. This factorization also implies a block-diagonal form for $\mathbf{C}_*$. Note that this lack of correlation between different parameters in the prior does not prevent correlation between parameters from arising in the posterior distribution (due to contributions from the likelihood function). This modular approach means that the priors on different signal parameters can have different forms. The form of the covariance matrix $\mathbf{C}_k$ should be selected based on what types of prior information are expected to be relevant to parameter k. $\mathbf{C}_k$ may be completely fixed *a priori* or may be adaptive using a small number of hyperparameters.

Assuming that the noise parameters ($\boldsymbol{\phi}$ and $\mathbf{a}$) and the smoothness parameters in $\mathbf{C}_*$ are known, the full hierarchical model is given by

$$\begin{aligned} P(\mathbf{y}_*, \mathbf{w}_*, \mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a}) \;=\; & \\ & \prod_v \Big[P(\mathbf{y}_v | \mathbf{w}_v, \phi_v, a_v) P(\phi_v) P(a_v) \Big] \; \prod_k \Big[P(\mathbf{w}_k | \mathbf{C}_k) P(\mathbf{C}_k) \Big] \end{aligned} \tag{4.8}$$

4.4 Priors on Forward Model Parameters $\mathbf{w}_k$

This section considers three forms of covariance matrices $\mathbf{C}_k$: a non-spatial prior, an adaptive spatial smoothness prior, and a Gaussian process prior that combines an informative non-spatial prior with adaptive spatial regularization.

4.4.1 Fixed Non-spatial Prior

Non-spatial priors define the prior belief about the parameter values. Highly implausible parameter values should not be used, even if they provide a better fit to the voxel's data. These priors often appear as part of biophysical models, since there

may be strong prior knowledge about the reasonable range of values for a parameter if it has a physical or physiological meaning.

To permit VB and EO-based inference later, non-spatial priors must be normally distributed, of the form $N(\mu_k, \sigma_k^2)$. As in equation 4.7, prior correlation between parameters are forbidden. For the applications under consideration, a Gaussian is a reasonable approximation to the *a priori* beliefs derived from biophysical information about the signal model. These are the same priors used in previous research (Chappell et al., 2009; Woolrich et al., 2004) and in the case of the simulated data it is the true marginal prior on each voxel. For simplicity, the non-spatial prior is also assumed to be the same for all voxels (i.e. it is not location-dependent).

The joint distribution of this non-spatial prior across all voxels is a multivariate normal distribution:

$$\mathbf{w}_k \sim N(\boldsymbol{\mu}_k, \mathbf{C}_k), \tag{4.9}$$

$$\mathbf{C}_k = \sigma_k^2 \mathbf{I} \tag{4.10}$$

and $\boldsymbol{\mu}_k$ is simply a $V \times 1$ column vector filled with μ_k. These priors are easy to implement because they have no hyperparameters to infer on and no inter-voxel prior correlation.

4.4.2 Laplacian Spatial Smoothness Prior

Laplacian spatial smoothness priors provide an adaptive Bayesian alternative to presmoothing the data with a Gaussian kernel. Presmoothing is a common preprocessing method that reduces spatially-decorrelated noise at the cost of also blurring the signal of interest. This reduced noise means that models can be more reliably fit to the data contained in a single voxel, so simple voxelwise (non-spatial) inference methods can then be used. However, this simplistic approach has several drawbacks. The results are often highly sensitive to the smoothing kernel size, which

is often selected arbitrarily. In addition, the optimal smoothing level for one signal parameter image can be inappropriate for the others: higher noise usually demands higher smoothing, while the fine spatial details of strong signals will be lost if too much smoothing is used. A single kernel diameter is used to presmooth the whole data set, so this is generally a compromise value that may not be suitable for some of the parameter images.

With spatial smoothness priors the *unsmoothed* data is used directly as $\mathbf{y}_*$. Different smoothness hyperparameters are used on each parameter image, avoiding the drawbacks associated with presmoothing the data. Most importantly, the amount of smoothness is determined adaptively in the Bayesian framework, automatically trading off accuracy against complexity.

A Laplacian spatial prior on parameter k is given by Penny et al. (2005):

$$\mathbf{w}_k \sim N(\mathbf{0}, (\alpha_k \mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}) \tag{4.11}$$

where α_k is the smoothness hyperparameter. The Laplacian is defined there as $\mathbf{L} = 4\mathbf{I} - \mathbf{A}$, where $\mathbf{A}$ is the incidence matrix ($\mathbf{A}_{ij} = 1$ if voxels i and j are spatially adjacent and 0 otherwise). To avoid edge effects due to this definition's fixed boundary conditions, this is slightly modified by using the regularized unweighted graph Laplacian (Zhu et al., 2003): $\mathbf{L} = \mathbf{D} - \mathbf{A} + (10^{-3})\mathbf{I}$, where $\mathbf{D}$ is a diagonal matrix giving the degree (number of neighbours) of each voxel, i.e. $\mathbf{D} = diag(rowsum(\mathbf{A}))$. This latter definition of $\mathbf{L}$ is used throughout this chapter.

This prior is a multivariate normal across voxels, thereby fitting into the hierarchical framework (see equation 4.7). $\mathbf{C}_k$ provides correlation between the values of a parameter in nearby voxels. Importantly, even though the precision matrix $\mathbf{C}_k^{-1} = \alpha_k \mathbf{L}^{\mathrm{T}}\mathbf{L}$ is sparse, the covariance $\mathbf{C}_k$ is a full matrix and the covariance between each pair of voxels is positive (dropping towards zero for voxels that are only distantly connected). Effectively this is a global scale prior on the Laplacian (curvature) of the parameter image, with adaptive weight α_k.

The Laplacian spatial prior has a fixed-form sparse precision matrix, making it very efficient and simple to infer on. The prior also conflates magnitude and smoothness (Penny et al., 2005), and there is no obvious way to control the voxelwise marginal priors without completely changing its structure (but see Harrison et al. 2007 for a related method that adjusts smoothness using the matrix-exponential of the Laplacian). The combined prior in the next section makes this easy, allowing spatial and non-spatial prior information to be combined in a natural way.

4.4.3 Combined Spatial/Non-spatial Gaussian Process Prior

In order to use both spatial and non-spatial information for the same parameter, a single covariance matrix $\mathbf{C}_k$ is constructed which has these desired properties. The covariance matrix needs to perform two separate purposes: controlling the marginal prior variance of each element of $\mathbf{w}_k$, and adjusting the amount of correlation between the different elements of $\mathbf{w}_k$. Here, the first part needs to be held fixed while the second part is allowed to vary.

An informative non-spatial prior represents an underlying belief about reasonable values for a parameter to take on, so it is actually a statement on each voxel's marginal prior distribution: $w_{k,v} \sim N(\mu_k, \sigma_k^2)$. The marginal should not change as the level of smoothness varies. Note that the marginal distribution of a multivariate normal is simply the corresponding subset of the mean vector and covariance matrix, and so to obtain the desired marginal prior on each element $\mathbf{w}_k$ is obtained by using a multivariate normal prior:

$$P(\mathbf{w}_k) \quad = \quad N(\boldsymbol{\mu}_k, \mathbf{C}_k) \tag{4.12}$$

where all elements of $\boldsymbol{\mu}_k$ are set to μ_k, and it is necessary to ensure that the diagonal elements of the covariance matrix $\mathbf{C}_k$ match the informative prior variance σ_k^2.

To obtain this type of control, the spatial covariance matrix $\alpha_k^{-1}(\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}$ does not provide a good starting point. Since there is only a single scaling factor α_k^{-1}, it is

not possible to keep the diagonal of $\mathbf{C}_k$ constant while independently adjusting the correlation between voxels (the smoothness).

Instead, the Gaussian process approach is used to directly construct the covariance matrix. A Gaussian process is defined over all points in space, but only evaluated it at a finite number of points (the voxel locations $\mathbf{x}_v$, $v \in 1 \dots V$). The distance matrix $\mathbf{\Delta}$ is defined so that each element is the Euclidean distance between voxels i and j: $\mathbf{\Delta}_{ij} = |\mathbf{x}_i - \mathbf{x}_j|$. In the applications presented in this chapter, this produced better results than the more commonly-used squared distance, $\mathbf{\Delta}_{ij} = |\mathbf{x}_i - \mathbf{x}_j|^2$, possibly because the latter induces very high posterior correlation between adjacent voxels (see discussion in section 4.10). In principle it would also be possible to use other distance measures, for example a functional distance metric used in a similar prior by Bowman (2007).

The correlation decays exponentially with distance, with a characteristic length scale controlled by an adaptive δ_k for each parameter image. Using the non-spatial prior variance σ_k^2 leads to the positive-definite covariance matrix:

$$\mathbf{C}_k \quad = \quad \sigma_k^2 \exp(-\mathbf{\Delta}/\delta_k). \tag{4.13}$$

where $\exp(\cdot)$ is the elementwise (not matrix) exponential. This is just the product of the "constant" and "exponential" covariance functions (Rasmussen and Williams, 2006, pp. 94–95). What is novel is that this formulation is used to express an informative biophysical prior by fixing the constant term σ_k^2 based on prior information rather than including it as a parameter to infer from the data.

In particular, the diagonal values of $\mathbf{C}_k$ are exactly σ_k^2 because $\mathbf{\Delta}_{ij} = 0$ for $i = j$. This satisfies the biophysical prior by providing the desired, fixed marginal distribution. Note that this prior is defined in terms of covariance, so the precision matrix $\mathbf{C}_k^{-1}$ does not have a simple, sparse form; it is found by inverting $\mathbf{C}_k$ numerically. This quickly becomes a significant computational cost, even when there are only a few hundred voxels.

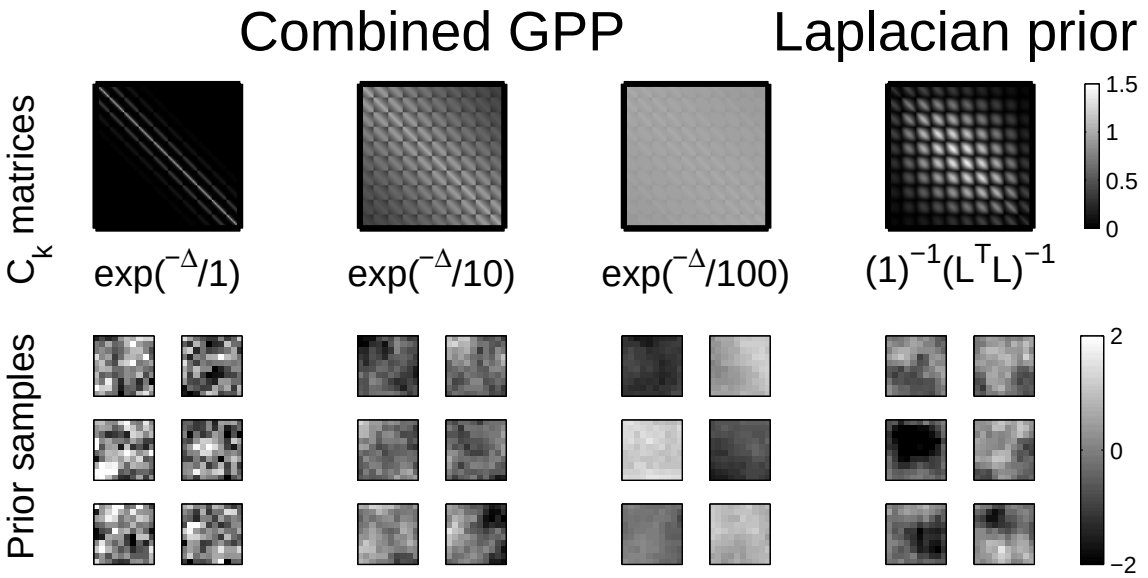


Figure 4.1: Top row: the spatial covariance matrices $\mathbf{C}_k$ corresponding to the combined GPP prior (with $\sigma_k^2 = 1$ and $\delta_k = 1$, 10, and 100) and the spatial smoothness prior (with $\alpha_k^{-1} = 1$). The equations for these matrices are printed below each one. There are 100 voxels, arranged in a 10×10 grid. Remaining plots: Six independent samples of the $\mathbf{w}_k$ were drawn from each of the spatial prior distributions, $\mathbf{w}_k \sim N(\mathbf{0}, \mathbf{C}_k)$. Note that the GPP keeps the voxel values mostly in the range of –2 to +2, across a wide range of smoothnesses; this is because each voxel has a fixed marginal prior of $N(0, 1)$. In contrast, adjusting the weight factor of the spatial smoothness prior would simply rescale these samples in proportion to $\alpha_k^{-1/2}$, so smoothness and marginal variance cannot be controlled separately.

4.4.4 Prior Samples and Limiting Cases

The combined GPP is parameterized by a correlation length δ_k, which gives it noticeably different behaviour from the spatial smoothness prior with its weight parameter α_k. Figure 4.1 shows the behaviour of the GPP at three correlation lengths ($\delta_k = 1$, 10, and 100 voxels) as well as the spatial smoothness prior (with fixed weight $\alpha_k = 1$). For each of these the covariance matrix is plotted in the top row, with some samples from the prior shown below, using a 10×10 grid of voxels as the spatial structure and a $N(0, 1)$ informative prior in the GPP. Note that all of the samples use the same colour range, from –2 to +2.

In the left column $\delta_k = 1$, so the off-diagonal elements decay quickly and the

samples look fairly similar to unstructured $N(0,1)$ white noise. In the limit of $\delta_k \to 0$, the non-spatial prior dominates and the prior approaches $P(\mathbf{w}_k) \sim N(\boldsymbol{\mu}_k, \sigma_k^2\mathbf{I})$, which is the fixed non-spatial prior (equations 4.9–4.10). The other extreme would be $\delta \to \infty$, where the correlation between voxels approaches unity. This implies that $\mathbf{w}_k$ must be the same in all voxels; so $\mathbf{w}_k$ effectively becomes a global parameter, and the same informative $N(\mu_k, \sigma_k^2)$ prior applies to this value. Approximating this limit, column three of the figure shows $\delta = 100$; all elements of the covariance matrix are close to 1. The samples show nearly-identical values in all voxels, but this global value is sampled from the same $N(0,1)$ prior.

A more typical value of $\delta_k = 10$ is shown in column two, which shows clear spatial structure. This is visually quite similar to the samples from the spatial smoothness prior in column four.

Notice that across all of these values of δ_k, the non-spatial prior is observed; across many samples, each voxel is individually $N(0,1)$ distributed. This level of smoothness is comparable to samples from the spatial smoothness prior (column four). Note, however, that it is only possible to adjust the scale of the spatial smoothness prior (proportional to the weight α_k^{-1}), so the variance of that prior always varies in proportion to its smoothness.

4.5 Inference on the Hierarchical Model

The previous section described the Bayesian hierarchical model, including a general forward model $\mathbf{g}$, an AR(1) noise model, and several ways of expressing spatial and non-spatial prior information in this framework. In this section, the approaches used for inference are described. The most important output is the posterior distribution on the forward-model parameters $\mathbf{w}_*$. These are normally broken down into separate parameter images ($\mathbf{w}_k$) for viewing the results. In addition, estimates of the noise in each voxel (precision ϕ_v and AR parameter a_v) and the spatial smoothness

hyperparameters (α_k and δ_k) are also obtained, but these noise-parameter images and smoothness scalars are usually not meaningful in isolation.

For computational efficiency it is desirable to use established approximations such as variational Bayes (Attias, 2000). The objective is to maximize the "free energy", which is a lower bound on the model evidence (MacKay, 2003). The most commonly used method is mean-field VB, in which the posterior distribution is broken into several factors. With a suitable factorisation and conjugate priors, the posterior distribution has a known form. The factorized posterior distribution is found by using an expectation maximization algorithm with a set of analytic update equations.

In the case of this Gaussian process prior, the closed algebraic solution for the factorized distribution of δ_k is not known (even on linear models). One approach is to maximize the free energy numerically to obtain a point estimate (or alternatively, a quadratic approximation) to δ_k. However, it was found that under certain circumstances this technique resulted in extremely slow convergence on the combined prior, essentially becoming locked in on a δ_k value that was highly sensitive to initial conditions. This occurred when the prior dominated each voxel's conditional distribution. A derivation of this technique and a discussion of the problems encountered are found in appendix B.1.

Evidence optimization is another inference method that is commonly used for inference on Gaussian processes (Rasmussen and Williams, 2006). EO works by expressing the model evidence in terms of the hyperparameters (δ_k) and integrating out the intermediate level of the hierarchy (in this case, the signal parameters $\mathbf{w}_*$). Numerically maximizing the evidence expression yields a point estimates of the δ_k's, and also provides the posterior distribution of $\mathbf{w}_*$.

EO is most commonly run as a single optimization; once the maximum evidence is found, the results of the inference are all known. However, the EO integrals are not tractable for general non-linear models, so they use a local linear approximation

which needs to be updated periodically. This means that an iterative approach is needed, where the forward model is relinearized based on the latest EO estimates, and EO is re-run on the new linear model.

In practice a hybrid approach is used for inference: the VB updates are still used to obtain estimates of the voxelwise noise parameters and of any non-GPP hyperparameters (i.e. the α_k's for signal parameters that use the Laplacian prior instead of the GPP prior). Since both EO and VB are iterative methods for maximizing the free energy, these approaches are compatible; for more details see section 4.10. These VB updates are interleaved between the EO steps and the relinearization of the signal model, and the entire procedure is iterated until convergence.

4.5.1 Estimating δ_k using Evidence Optimization

Evidence optimization is one of the preferred methods for inferring on Gaussian process priors (Rasmussen and Williams, 2006), and is used to infer on the parameter δ_k. EO integrates out some of the intermediate parameter distributions (in this case, $\mathbf{w}_*$) so that the hyperparameters (δ_k) can be optimized directly from the data ($\mathbf{y}_*$). It is particularly useful for obtaining a point estimate of a hyperparameter when it cannot be factored out of the covariance matrix.

EO can be used to infer on all top-level parameters, and could be used to obtain point estimates for the noise parameters at the same time. However this means that there would be $K+2V$ parameters to optimize over rather than just K. Instead a hybrid approach is used, where VB update equations are still used to infer the voxelwise noise parameters, ϕ_v and a_v. The (conditional) evidence expression is obtained by using these estimated noise parameters and integrating out the unknown $\mathbf{w}_*$. The smoothness parameters δ_k are found by maximizing the evidence expression. The entire process is iterative, with EO steps interleaved between VB updates on the

noise and re-linearizations of the model $\overline{\mathbf{g}}$.

The conditional evidence, $\mathscr{E}(\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a}) = P(\mathbf{y}_*|\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a})$, is found by integrating out $\mathbf{w}_*$ from the model (equation 4.8):

$$\mathscr{E}(\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a}) = \int P(\mathbf{y}_*|\mathbf{w}_*, \boldsymbol{\phi}, \mathbf{a}) P(\mathbf{w}_*|\mathbf{C}_*)\, d\mathbf{w}_* \tag{4.14}$$

A detailed derivation is given in appendix B.2. The quantity of interest is its derivative with respect to the hyperparameter, which is given by

$$\frac{\partial \log \mathscr{E}}{\partial \delta_k} = \frac{1}{2}\mathrm{Tr}\left[(\mathbf{C}_* - \boldsymbol{\Sigma}_* - (\mathbf{m}_* - \boldsymbol{\mu}_*)(\mathbf{m}_* - \boldsymbol{\mu}_*)^{\mathrm{T}})\frac{\partial(\mathbf{C}_*^{-1})}{\partial \delta_k}\right]. \tag{4.15}$$

where $\boldsymbol{\Sigma}_*$ and mean $\mathbf{m}_*$ are functions of $\mathbf{C}_*$ and are defined as follows:

$$\boldsymbol{\Sigma}_*^{-1} = \mathbf{J}_*^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*\mathbf{J}_* + \mathbf{C}_*^{-1} \tag{4.16}$$

$$\boldsymbol{\Sigma}_*^{-1}(\mathbf{m}_* - \boldsymbol{\mu}_*) = \mathbf{J}_*^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*(\mathbf{y}_* - \overline{\mathbf{g}}(\boldsymbol{\mu}_*)). \tag{4.17}$$

These are useful later because $N(\mathbf{m}_*, \boldsymbol{\Sigma}_*)$ is the posterior distribution of the signal parameters $\mathbf{w}_*$. Finding a descending zero of $\partial \log \mathscr{E}/\partial \delta_k$ yields an evidence-maximizing estimate of δ_k. Since this hyperparameter estimate is collecting information from all voxels, it seems reasonable to assume that the true distribution of δ_k is quite narrow so that using a point estimate provides a good approximation.

This is iterated several times, optimizing the evidence by adjusting each hyperparameter δ_k separately, then re-estimating the noise parameters (and α_k) with the VB updates, and finally re-centering the linearization $\overline{\mathbf{g}}$ on the estimated mean of $\mathbf{w}_*$. In particular, the relinearization step must be kept outside of the EO calculation because $\mathscr{E}(\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a})$ are not directly comparable across linearizations; for more details see the discussion (section 4.10).

4.5.2 Inferring Forward Model Parameter Distributions from EO Results

In constructing the EO equations, the unknown forward model parameters $\mathbf{w}_*$ were integrated out. Now that $\mathbf{C}_*^{-1}$ is known, the right-hand-side of equations 4.16 and

4.17 can be evaluated to obtain $\mathbf{m}_*$ and $\mathbf{\Sigma}_*$. To preserve inter-parameter correlations these are calculated without factorizing over parameters (using the full $KV \times KV$ matrices). The posterior distribution on the forward model parameters is given by $\mathbf{w}_* \sim N(\mathbf{m}_*, \mathbf{\Sigma}_*)$. Note that this is possible even if the covariance matrices $\mathbf{C}_k$ are of different forms, for example if the Laplacian spatial smoothness prior (equation 4.11) is used on some parameters.

4.5.3 Inference on Noise Parameters using VB

For the VB steps in this section, the elements relevant to voxel v are taken from the mean vector $\mathbf{m}_*$ to obtain $\mathbf{m}_v$, and the appropriate rows and columns from the precision matrix $\mathbf{\Sigma}_*^{-1}$ form $\mathbf{\Sigma}_v^{-1}$. This yields the voxelwise distributions $\mathbf{w}_v \sim N(\mathbf{m}_v, \mathbf{\Sigma}_v)$ and the standard VB updates can be applied to obtain probabilistic estimates for ϕ_v and a_v. The update equations for these factorized distributions are described elsewhere (Penny et al., 2003). For most of the applications in this paper a simple white noise model is used by fixing all $a_v = 0$ in the AR noise model and using $\mathbf{R}_v = \mathbf{I}$. The update equations in this case are straightforward. The factorized posterior distribution on ϕ_v is given by $q(\phi_v)$:

$$
\begin{aligned}
q(\phi_v) &= Ga(\phi_v; b_v, c_v) && (4.18)\\
\frac{1}{b_v} &= \frac{1}{2}\left(\mathbf{y}_v - \overline{\mathbf{g}}(\mathbf{m}_v)\right)^{\mathrm{T}}\left(\mathbf{y}_v - \overline{\mathbf{g}}(\mathbf{m}_v)\right) + \frac{1}{2}\mathrm{Tr}\left[\mathbf{\Sigma}_k \mathbf{J}_v^{\mathrm{T}} \mathbf{J}_v\right] + \frac{1}{b_0} && (4.19)\\
c_v &= \frac{T}{2} + c_0 && (4.20)
\end{aligned}
$$

with a non-informative prior $\phi_v \sim Ga(b_0, c_0)$, with $b_0 = 10^6$ and $c_0 = 10^{-6}$. The expected value of ϕ_v is simply $b_v c_v$. The $T \times T$ matrices $\phi_v \mathbf{I}$ can be assembled into the global noise precision matrix $\mathbf{\Phi}_*$ for use in the EO calculations. Using the expectation value as a point estimate is equivalent in this case to integrating through the full distributions on ϕ_v.

For signal parameters which use the Laplacian (rather than the Gaussian process) prior, the spatial smoothness precision parameter α_k can be found from $\mathbf{w}_v$ using the update equations from Penny et al. (2005):

$$\alpha_k = Ga(\alpha_k; g_k, h_k) \tag{4.21}$$

$$\frac{1}{g_k} = \frac{1}{2}\mathbf{m}_k^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{L}\mathbf{m}_k + \frac{1}{2}\mathrm{Tr}\left[\widehat{\mathbf{\Sigma}}_k\mathbf{L}^{\mathrm{T}}\mathbf{L}\right] + \frac{1}{q_1} \tag{4.22}$$

$$h_k = \frac{V}{2} + q_2 \tag{4.23}$$

where $\mathbf{L}$ is the regularized unweighted Laplacian and a non-informative prior is used: $\alpha_k \sim Ga(q_1, q_2)$, $q_1 = 10^{12}$, $q_2 = 10^{-12}$. $\widehat{\mathbf{\Sigma}}_k$ is a $V \times V$ diagonal matrix built up from the $(k,k)^{\mathrm{th}}$ element of each $\mathbf{\Sigma}_v$; this discards the covariance information in order to match the definition used by VB. Since this new technique will be compared to existing methods that use pure VB, this ensures that α_k is always inferred in the same way, regardless of the priors used on the other signal parameters.

4.6 Overview of Results

The next three sections demonstrate the benefits of combining non-spatial and spatial prior information by considering a linear simulation and two non-linear modelling applications. The first non-linear application is a model of the haemodynamic response function (HRF) in FMRI using a linear basis set, with priors on the allowable HRF shapes (ratios of the regressors in the basis). The second is on resting-state arterial spin labelling (ASL) data, where the perfusion response is estimated by acquiring images with different inversion times (TIs). Both have the same basic form: a scale factor that determines the size of the signal, and a shape parameter that has a useful biophysical prior. In both cases, the recommended approach is to use a Laplacian spatial prior on the scale parameter and a combined Gaussian process prior on the shape parameter (called "model C"). These are compared against using

Data set	Voxels	Iterations	Model A	Model B	Model C
Linear simulation	400	20	1s	1s	69s
HRF Simulation 1	400	20	6s	5s	34s
HRF Simulation 2	400	20	5s	6s	46s
Pain FMRI data (AR1)	1062	15	48s	51s	3.5 h
ASL Data	1087	100	70s	70s	6.3 h

Table 4.1: Computation times for the inference methods used by models A, B, and C. These were implemented in C++ and performed on a single core of a 3GHz Intel Xeon X5365 processor. Mean times are given when there were multiple computations on the same model. Note in particular the dramatic effect of increasing voxel numbers on computation time; however, a partitioning-based approach could potentially eliminate this problem (see discussion). The number of iterations was chosen arbitrarily; typically there was little change after 20 iterations. The EO step stopped once δ_k was known to within $\pm 1\%$. Later iterations are considerably faster than early iterations because reaching these bounds required fewer evaluations of the evidence expression.

either a non-spatial prior on shape ("model A") or the Laplacian spatial prior on shape ("model B").

Voxel numbers and computation times for each of these results are shown in table 4.1.

4.7 Single-parameter Linear Model Simulations

In the first simulation, the signal is spatially-smoothed white noise (FWHM = 4.7 units) that has a (post-smoothing) marginal distribution of $N(0, 1)$. Values that are between -1 and 1 are set to zero, so the image consists of a few positive and negative islands: see figure 4.2(a). This static signal is repeated through 100 volumes and white noise (of amplitude 10 or 30) is added to the data set. A very simple linear model is used with only one regressor (all ones). To further simplify the simulation, the noise precision was fixed at the true value.

Figure 4.2 shows the improvement in discrimination that is provided by the combined prior. This is particularly evident in the high-noise situation. Since the combined prior is able to use the biophysical prior, it keeps the outputs constrained

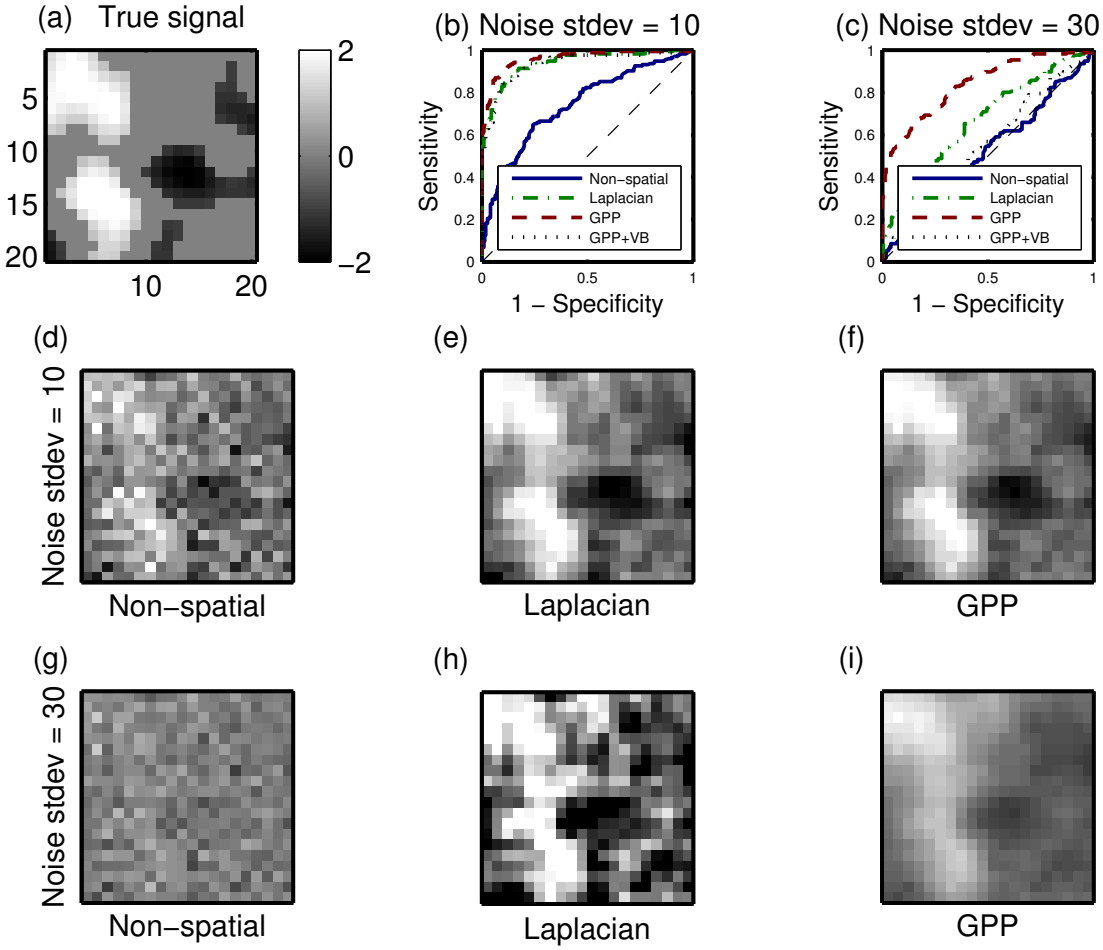


Figure 4.2: The single-parameter simulation. The true signal (a) is repeated in all 100 volumes. This is spatially-smoothed white noise, renormalized to have a marginal distribution of $N(0,1)$ in all voxels. Absolute values less than 1 have been set to zero. The same [-2,2] colour scale is used in all images in this figure. White noise of amplitude 10 is added and the data set is analysed using the fixed non-spatial prior (d), the Laplacian spatial prior (e) and the GPP (f). Using the absolute value of each of these parameter images, discrimination performance is shown in the receiver operating characteristic (ROC) curves (b). The GPP+VB line shows the GPP using an alternative inference method; see appendix B.1. This analysis is repeated using amplitude-30 noise and the same results are shown in (g-i) and (c).

to a sensible range (-1.1 to +2.0), while also providing spatial regularization. The spatial prior does not have the benefit of the biophysical prior and produces large estimates (-3.7 to +4.7) that are well outside the true range (-2.1 to +2.6). Since it allows these large fluctuations in value, the spatial prior also provides considerably less smoothness. The combined prior provides the best discrimination of the three methods.

Interestingly, the use of a linear model and fixed noise precisions mean that the hybrid VB-EO method reduces to pure EO, so it is not necessary to iterate the method. The other models are pure VB and are iterated 20 times.

4.8 Constrained Linear Basis Sets for Modelling HRF Shape Variations

When performing standard GLM analysis of FMRI data, the main HRF regressor $\mathbf{X}_1$ is usually accompanied by one or more variation terms $\mathbf{X}_2 \ldots \mathbf{X}_K$; often this is just a single term, the temporal derivative. These help to fit the natural variation in HRF shape more accurately, but have the downside of potentially overfitting to the noise. In Woolrich et al. (2004) it was shown that only limited linear combinations of the basis functions produce biophysically plausible shapes, and that this concept can be encoded in a fixed non-spatial prior resulting in increased sensitivity.

This model is parameterized as $\mathbf{g}(\beta_1, r_2, \ldots, r_K) = \beta_1\mathbf{X}_1 + r_2\beta_1\mathbf{X}_2 + \ldots + r_K\beta_1\mathbf{X}_K$. (To express the true regressor weight on $\mathbf{X}_k$ one can use $\beta_k = r_k\beta_1$, for $k = 2 \ldots K$.) An informative non-spatial prior also needs to be provided on the shape ratio parameters $r_2 \ldots r_K$.

Normally one might use this non-spatial prior on the HRF shape and a spatial smoothness prior on the activation scale (since "blobs" of activation are expected). However, this model could also benefit from the use of a spatial prior on the HRF shape terms. It is reasonable to assume that the HRF shape (parameterized by

$r_2 \ldots r_K$) is quite spatially consistent, at least within compact functional areas (Makni et al., 2008). Using the combined Gaussian process prior, both pieces of prior information can be encoded simultaneously.

4.8.1 Results on Simulated Data

In generating simulated data, it is assumed that activations are a linear mixture of a canonical regressor $\mathbf{X}_1$ and its temporal derivative $\mathbf{X}_2$: the forward model is $\mathbf{g}(\beta_1, r_2) = \beta_1 \mathbf{X}_1 + r_2 \beta_1 \mathbf{X}_2$. The informative prior on shape is $r_2 \sim N(0, 1)$, corresponding to a time-to-peak standard deviation of ± 1 second from the canonical HRF.

Using this model, a simulated BOLD FMRI data set was simulated using unit-amplitude white noise, and three models were evaluated:

Model A: Laplacian prior on β_1 with a non-spatial prior on r_2,

Model B: Laplacian prior on β_1 with a Laplacian prior on r_2, and

Model C: Laplacian prior on β_1 with the combined prior on r_2.

Since there is no non-spatial prior defined for activation size β_1, it makes sense to use the spatial smoothness prior in all models. The only difference is the choice of prior on the shape parameter r_2.

The first simulation consists of four Gaussian blobs of activation (FWHM of 3 to 5 voxels), each with a slightly different "sensible" HRF shape (see figure 4.3). These vary by the amount of derivative term added to the canonical HRF. The time-series consists of 200 scans with TR=3s and a single-event design with a random inter-stimulus interval of 5 to 20 seconds. White noise is added with an amplitude of 1, and a white noise model is used by fixing all AR parameters $a_v = 0$.

The results are shown in figure 4.4. All three models use the same spatial smoothness prior on activation strength β_1, so the estimates are nearly identical.

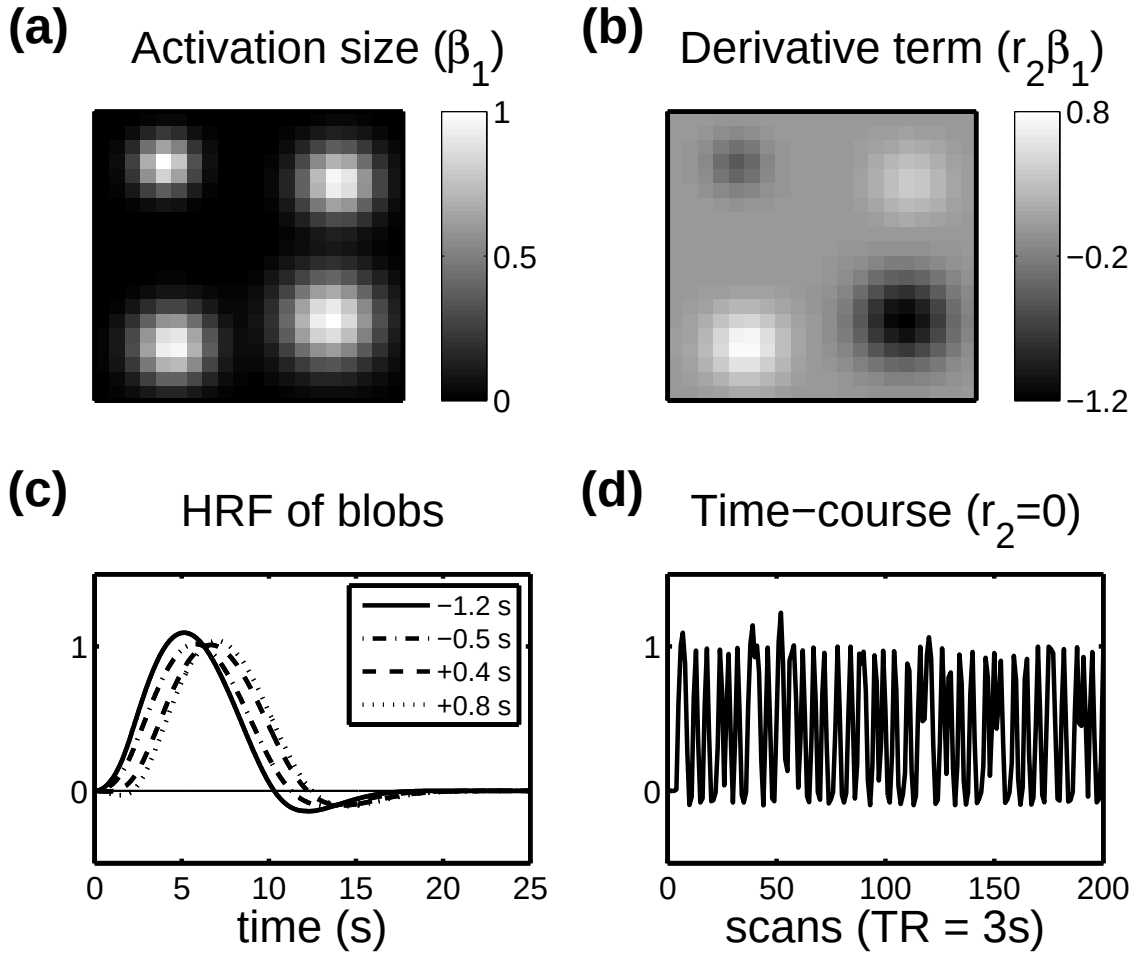
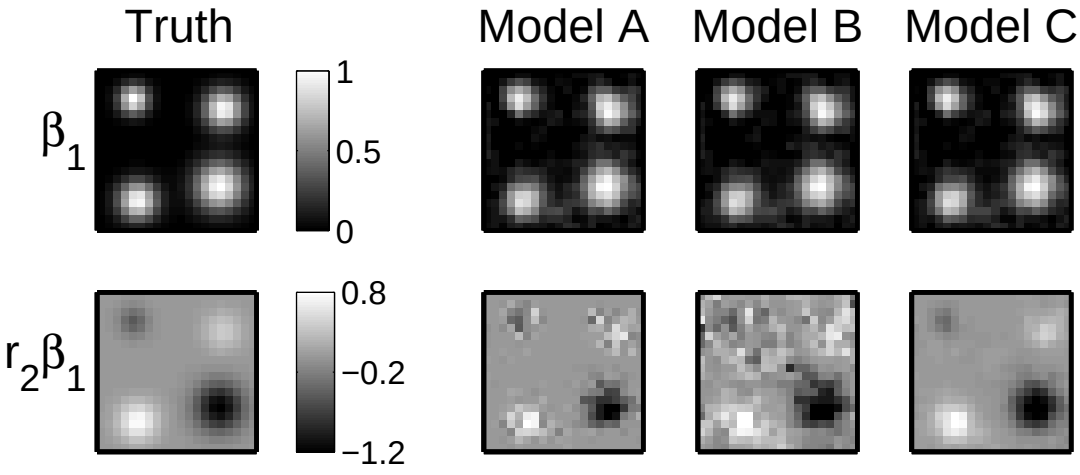


Figure 4.3: The simulated data used with the constrained linear HRF shape models. (a) The value of the activation parameter β_1, representing the amount of regressor $\mathbf{X}_1$ used in each voxel. (b) The value of $r_2\beta_1$, showing the weight of regressor $\mathbf{X}_2$ in each voxel. (c) HRF shapes in each blob. (d) The signal time-course, with no delay and before adding noise. This is the $\mathbf{X}_1$ regressor.



	Model A: *Laplacian* β_1, *Non-spatial* r_2	**Model B:** *Laplacian* β_1, *Laplacian* r_2	**Model C:** *Laplacian* β_1, *GPP* r_2
β_1 image	0.0641	0.0650	**0.0610**
$r_2\beta_1$ image	0.1491	0.2094	**0.0709**

Figure 4.4: Results for analysing simulated FMRI data with the constrained HRF shape model: $\mathbf{g}(\beta_1, r_2) = \beta_1\mathbf{X}_1 + r_2\beta_1\mathbf{X}_2$. Top row: The activation size β_1: truth, Model A, Model B, and Model C. Second row: Second regressor estimate β_2 in the same models. Table: Errors between each of the estimated images and the ground-truth image, expressed as RMS difference as a fraction of true RMS amplitude. Model C produces the most accurate estimates for both parameter images (indicated by the bold values).

However, for the weight of $\mathbf{X}_2$ there are obvious improvements over either the biophysical (non-spatial) or Laplacian (spatial) prior alone. The root-mean-square (RMS) error is halved, and the qualitative appearance is also much closer to the true $r_2\beta_1$ image.

The most obvious difference between models B (spatial prior on r_2) and C (combined prior on r_2) is the noise in the $r_2\beta_1$ estimates where β_1 is small. This can only occur in model B because there is no informative prior to restrain r_2.

4.8.2 Results on Simulated Data with Extreme HRF Shapes

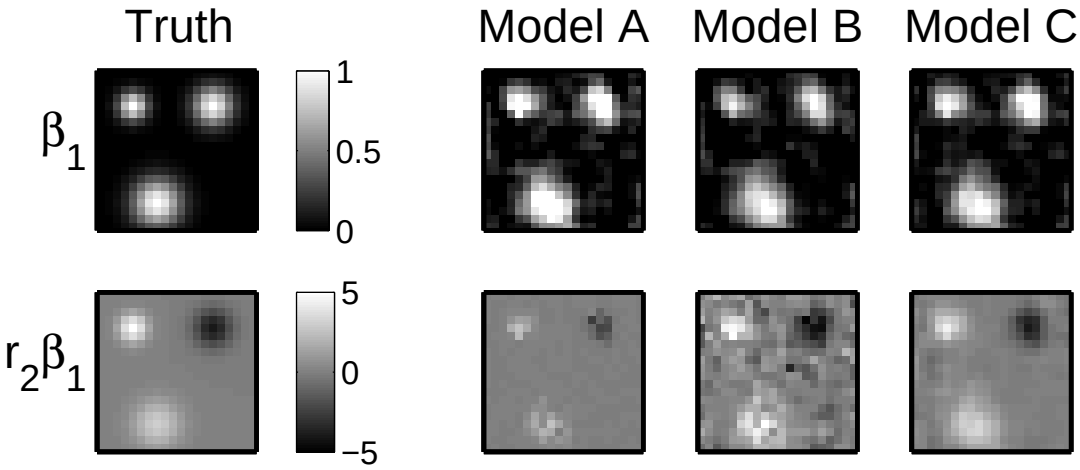
A common criticism of informative priors is that if they are carelessly made too tight, they can completely overwhelm the data. It may therefore be considered beneficial if sufficiently clear data can dominate over a prior that is somewhat too strong, yielding a more data-driven result. The second simulation tests this scenario by picking values for r_2 that are in the tails of the prior distribution, at +3, -4, and +5 standard deviations. (The white noise is also increased to an amplitude of 3.)

Figure 4.5 shows the results. In model A, the overly-tight non-spatial prior means that the magnitude of $r_2 = \beta_2/\beta_1$ is consistently underestimated, leading to underestimates of $|\beta_2|$ and overestimates of β_1. In model B, the spatial prior on r_2 does not know about the biophysical prior information, and provides good estimates of shape; and as before, it suffers from noisy β_2 estimates where there is no activation. The combined prior C gives good estimates of both β_1 and β_2 images.

Unlike the non-spatial prior, the combined prior is able to pool data from nearby voxels and collectively overpower the biophysical prior information. In this way, the entire δ_k-radius region suffers the prior "cost" only once, rather than once per voxel.

4.8.3 Results on Real FMRI Data

The constrained HRF shape model is applied to a slice of real FMRI data. A three-element basis set $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ is used to describe the HRF. These account for



	Model A: *Laplacian* β_1, *Non-spatial* r_2	**Model B:** *Laplacian* β_1, *Laplacian* r_2	**Model C:** *Laplacian* β_1, *GPP* r_2
β_1 image	0.1817	0.1514	**0.1431**
$r_2\beta_1$ image	0.6700	0.7806	**0.3190**

Figure 4.5: Results for the second HRF simulation, with shapes that are somewhat outside the non-spatial biophysical prior (r_2 = +3,–4,+5 stdevs). Top row: The activation size β_1: truth, Model A, Model B, and Model C. Second row: Second regressor estimate $r_2\beta_1$ in the same models. Table: Errors between each estimated images, expressed as RMS difference as a fraction of true RMS amplitude. Model C produces the most accurate estimates for both parameter images (indicated by the bold values).

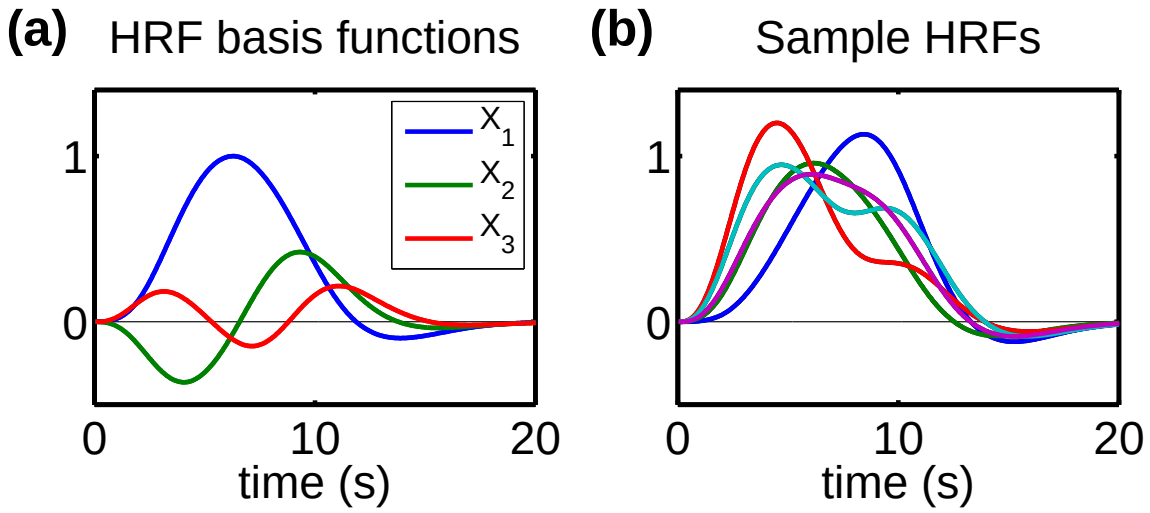


Figure 4.6: The constrained HRF model used for real data. (a) The three linear regressors, parameterized as $\mathbf{g}(\beta_1, r_2, r_3) = \beta_1\mathbf{X}_1 + r_2\beta_1\mathbf{X}_2 + r_3\beta_1\mathbf{X}_3$. The regressors are scaled so that typical HRFs are described by $N(0,1)$ distributions on r_2 and r_3. Right: Some typical HRF shapes from this model, sampling r_2 and r_3 from their informative $N(0,1)$ priors. These samples are not all plausible (e.g. the red line has a double peak) but the contraint greatly reduces the possible space of unrealistic shapes.

expected variation in HRF shape and roughly correspond to the stimulus convolved with the canonical HRF, its temporal derivative, and its dispersion (width) derivative. The model is parameterized as $\mathbf{g}(\beta_1, r_2, r_3) = \beta_1\mathbf{X}_1 + \beta_1 r_2\mathbf{X}_2 + \beta_1 r_3\mathbf{X}_3$ since the biophysical prior is defined on the shape ratio parameters r_2 and r_3. These regressors and reasonable priors on the ratios were generated using the half-cosine HRF model described in Woolrich et al. (2004). Figure 4.6 shows the basis set and some HRF shapes sampled from the priors on r_2 and r_3.

The data set is a single slice from an FMRI scan of a single-event pain experiment, for which the stimulus was the thermal noxious stimuli of 3 s duration administered to the dorsum of the volunteer's left hand with varying inter-stimulus interval (between 30 and 50 s). There are 1062 voxels in this slice and a total of 268 volumes used (TR=3 s). The scan was motion-corrected, and six additional linear motion regressors were included in the model. Note that using a spatial prior on these nuisance

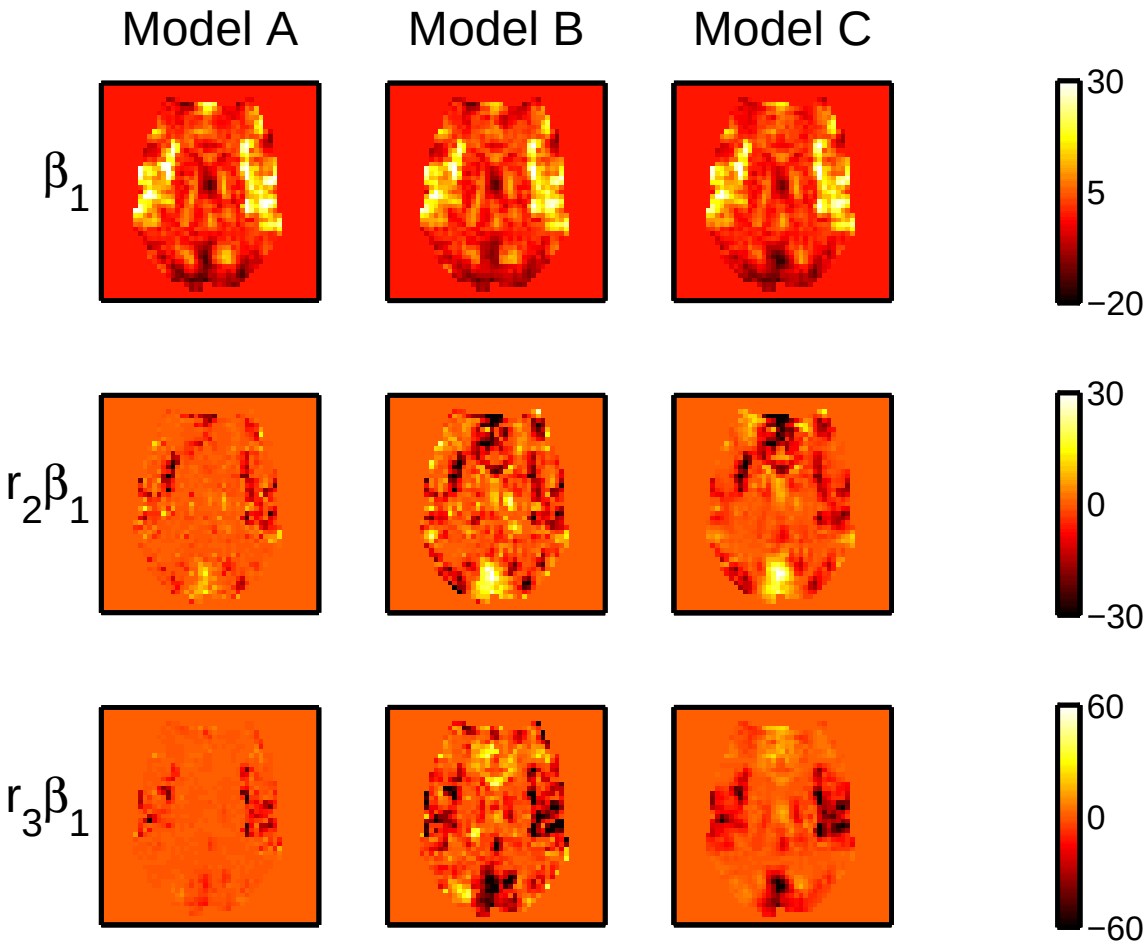


Figure 4.7: Analysis of a pain data set using three regressors, with weights β_1, $r_2\beta_1$, and $r_3\beta_1$. All three rows use a spatial smoothness prior on β_1, but differ in the prior on the $r_{2,3}$ values. Model A: Non-spatial priors are used on r_2 and r_3. Model B: spatial smoothness priors on all parameters. Model C: Combined prior on r_2 and r_3. All units are arbitrary.

regressors would be inappropriate since motion effects occur mainly on the edges; instead, an uninformative non-spatial $N(0, 10^{12})$ prior is used for these regressors.

As with the simulated data, three models are compared: model A uses the spatial smoothness prior on β_1 and non-spatial priors on r_2 and r_3; B uses spatial smoothness on β_1, r_2 and r_3; and C uses spatial smoothness on β_1 and the combined prior on r_2 and r_3.

The resulting parameter estimates are shown in figure 4.7. All three models show very similar patterns of activation, with relatively subtle differences between

them. In model A, the second and third regressors are almost unused because the biophysical prior information dominates, while in C the inferred smoothness on the r_2 and r_3 images shows spatially-extended areas where the shape is best explained by reasonable amounts of these additional basis terms. B produces similar-looking but noisier parameter images, without making use of the biophysical prior at all.

4.9 Perfusion Modelling with Multi-inversion ASL

Arterial spin labelling (ASL) is a non-invasive method for studying blood perfusion in the brain. ASL works by magnetically tagging blood in the neck to create an endogenous contrast agent, which is then observed in the slice of interest a short time later. For analysis of perfusion data, this section uses a simple non-linear generative model $\mathbf{g}(f, \Delta t)$ from Buxton et al. (1998) with just two parameters: cerebral blood flow (CBF) f and bolus arrival time Δt. Note that CBF is often used in stroke imaging, and the time delay is a confound in CBF estimation that may also be useful in itself, since it gives additional information about the perfusion into an area (e.g. whether the blood is taking a direct route or a circuitous route that indicates a blockage). The perfusion time-course is illustrated in figure 4.8. In order to observe the full shape of this curve, various inversion times (TIs) are used to vary the time between tagging and acquisition, as proposed in (Figueiredo et al., 2005). These times are indicated by the red arrows in that figure.

The multi-inversion-time ASL data was acquired on a 3 Tesla scanner with a healthy volunteer at rest, using the Q2TIPS sequence with saturation after 0.7 seconds (Luh et al., 1999). The data consists of tag-control pairs (already subtracted) taken at ten TIs, with the whole process repeated 30 times. The data set consists of a single slice with 1087 voxels within the brain. This is the same preprocessed data set from Chappell et al. (2009), analysed using the forward model implementation provided by Dr. Michael Chappell.

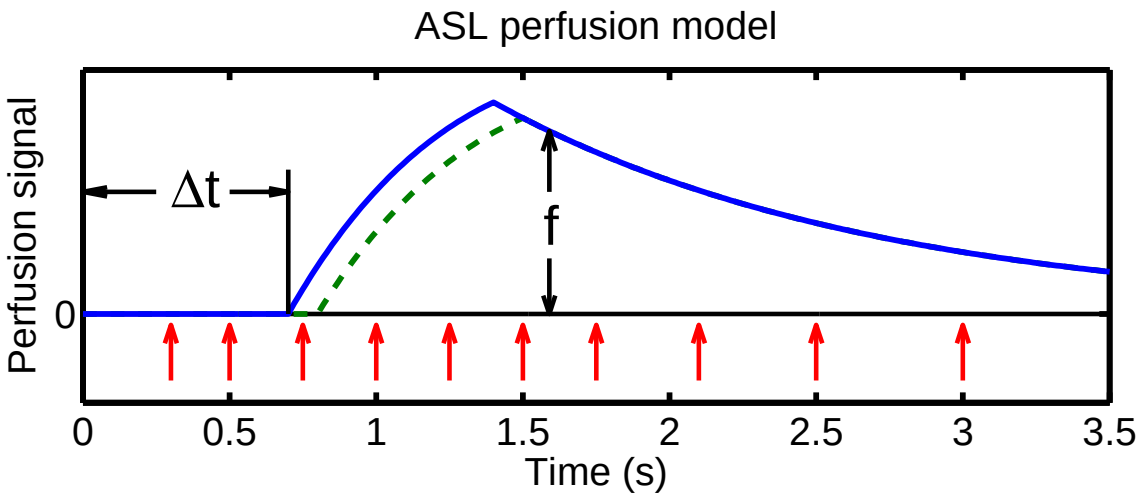


Figure 4.8: The perfusion model with two inferred model parameters: flow f and bolus arrival time Δt (Buxton et al., 1998). For times $t < \Delta t$, no tagged blood enters the voxel, then as the bolus flows into the voxel the signal grows, and after the bolus has fully arrived the signal decays with the T_1 of blood. Time-courses are shown for $\Delta t = 0.7$ s (solid line) and 0.8 s (dashed line); note that changing arrival time Δt only changes the signal in the middle range of the timepoints. The flow parameter f controls the overall signal size. The points where the perfusion curve is sampled (using different inversion times) are shown by the ten vertical arrows. Here the "perfusion signal" is the tag-control differenced data.

In practice, there is useful *a priori* information about the approximate arrival time of the bolus. The informative biophysical prior on the arrival time is a normal distribution with a mean of 0.7 s and standard deviation of 0.3 s. Previous analysis of this data set has shown that using this prior is reasonable and improves the quality of the resulting images (Chappell et al., 2009).

There is no meaningful prior available for CBF, so as before the spatial smoothness prior is used on this parameter in all models. This is evaluated using three different priors on the arrival time, in four models:

Model A: Spatial prior on f and non-spatial prior on Δt,

Model A′: Same as A, using presmoothed data (6mm FWHM),

Model B: Spatial prior on f and the same spatial prior on Δt, and

Model C: Spatial prior on f and combined prior on Δt.

All models are run directly on the unsmoothed data, except A′ uses data preprocessed with a fairly typical smoothing kernel. Using the full data set, the final model fits are shown in figure 4.9.

In these results, models B and C give very similar results for both parameter images. The Model A estimate of CBF is similar, but arrival time is quite noisy. When using the same model with 6 mm presmoothing (model A′), the arrival time image looks good but CBF is clearly blurred. The comparison of models A and A′ illustrates the compromise that is sometimes involved in presmoothing the data.

To evaluate each method across a range of noise conditions, degraded versions of this scan were constructed by using subsets of the full 30-repeat data set. This is more representative of data acquired in a clinical scan, where there would be less time available for data collection. Note that the full set of measurements (different TIs) were used in all cases, but there were simply fewer repeats of each measurement. To

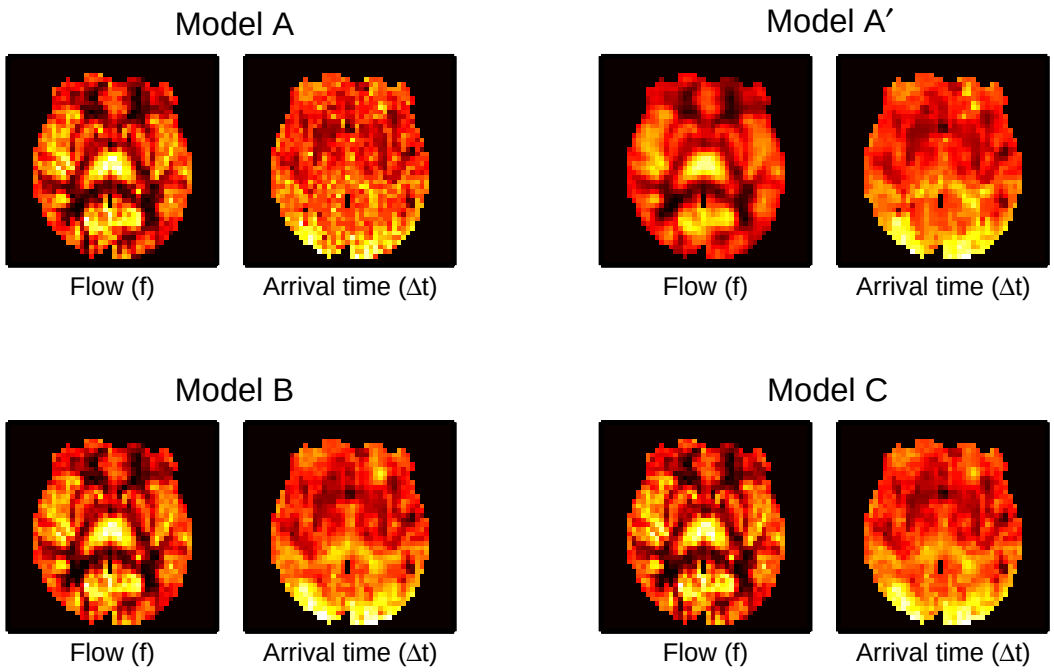


Figure 4.9: Flow and delay-time estimates using the full data set for each of the models. In all cases, the spatial smoothness prior was used on total flow f, while the prior on Δt was either non-spatial (models A and A′), spatial smoothness (model B), or the combined GPP (model C). Model A′ uses fixed 6mm FWHM smoothing while the others use none. These image pairs provide the ground truth used in figure 4.11. Images of flow f range from 0 (black) to 25 (white) in arbitrary scanner units, and arrival times Δt images range from 0 to 1.5 seconds.

avoid bias due to physiological drift, the scans in each subset were spread throughout the whole data set as much as possible. Subsets of 1, 2, 5, 10, and 15 evenly-spaced repeats were extracted, and all were analysed using exactly the same set of methods. A few of these resulting Δt images are shown in figure 4.10.

The gradual reduction in quality is evident as fewer repeats are used. However, each of the methods degrades differently, falling back increasingly on the prior information. For example, the models that use the non-spatial prior information (A, A′, and C) tend towards the prior mode of 0.7 seconds, while the Laplacian spatial prior method (B) makes physically-unrealistic guesses. The combined prior adopts relatively high spatial smoothness and stays within the biophysical prior range.

Figure 4.11 evaluates how accurately each method is able predict its own results

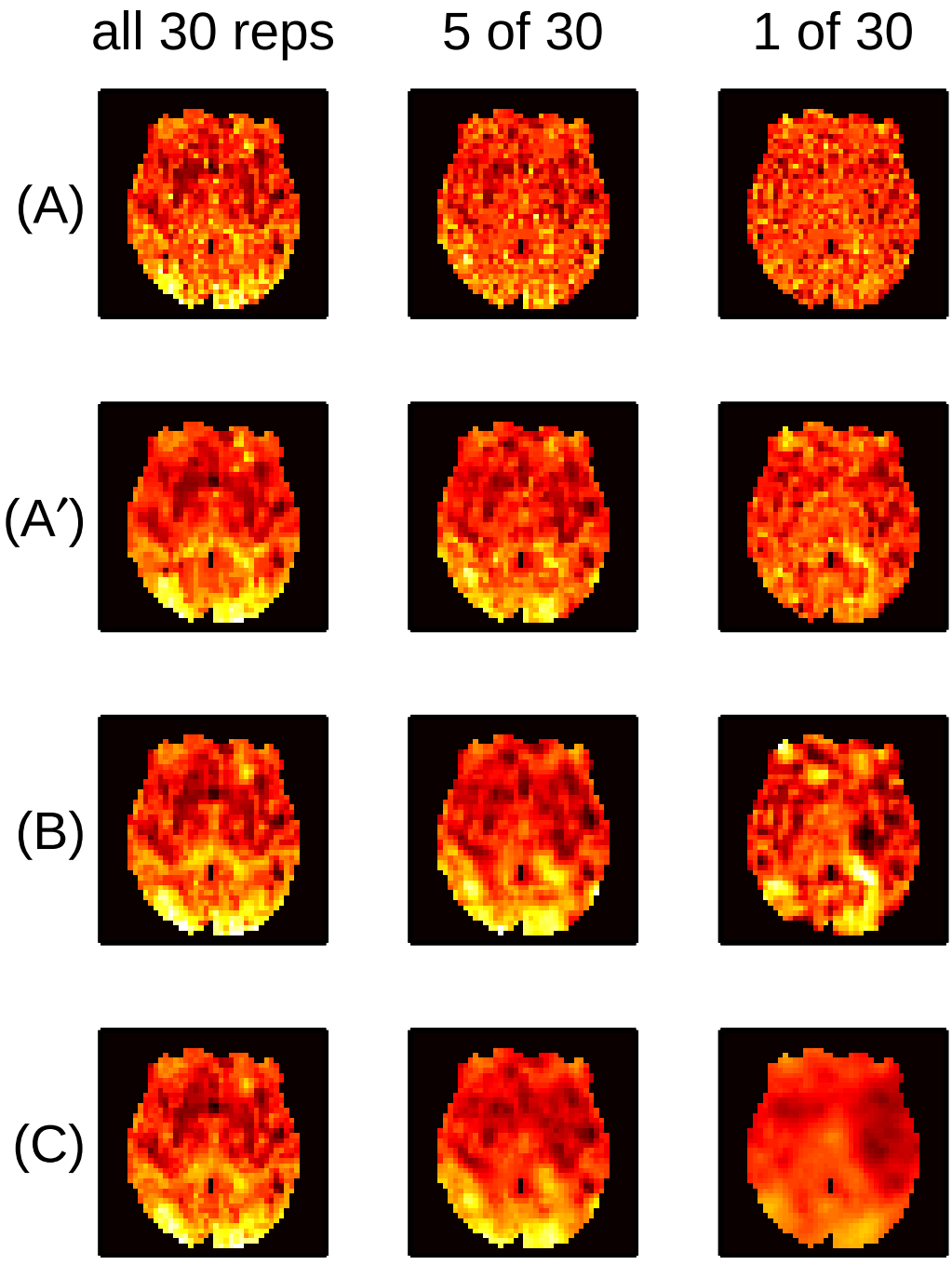


Figure 4.10: Typical images of arrival time Δt as the data is degraded (by reducing the number of repeats from 30 to 5 and to 1). The corresponding images of the flow parameter f are not shown. (A) non-spatial prior on Δt, (A′) non-spatial prior on Δt with 6mm presmoothing, (B) Spatial prior on Δt, (C) Combined GPP on Δt. Notice the different ways in which the estimated Δt images degrade. The colour scales run from 0 to 1.5 seconds.

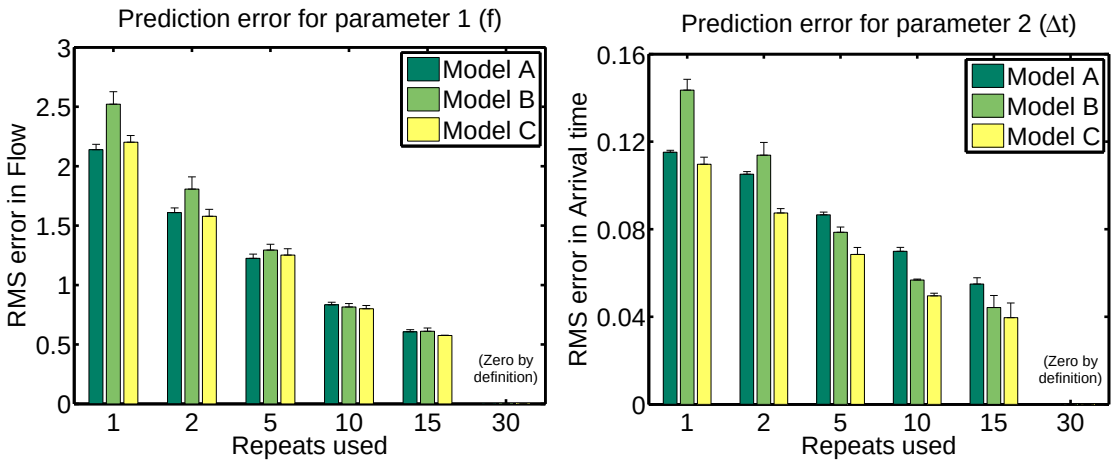


Figure 4.11: RMS difference between partial and complete data sets using the same method; lower is better. There are several runs (on non-overlapping subsets) used for each size; the bar shows the mean over all runs of a given size, with the error bars showing standard error of the mean (based on multiple disjoint subsets of the complete 30-repeat data set). Left: flow estimates. Right: shape estimates.

from limited data. With real data there is no ground truth available, so the images from figure 4.9 are used as a "gold standard" for each method. The SNR is reduced naturally by simply taking subsets of the data (between one and 15 repeats used in each analysis). The values shown are root-mean-square (RMS) errors, averaged over a many subsets of the full data set.

For estimating the time delay, using the combined prior (C) produces better predictions than the spatial prior (B) and the biophysical prior (A). When there are few repeats, the biophysical prior information is particularly useful and the spatial smoothness prior suffers because it cannot use this information. However, the spatial smoothing provided by the combined prior helps to reduce the noise in these images and provides better prediction than the non-spatial prior alone.

For the flow parameter, recall that the same prior is used in all models; any differences between these results is due to the indirect effect of having different arrival time estimates. In this respect there appears to be very little difference between using a non-spatial prior (A) or the combined prior (C) on Δt. However, using a spatial

prior (B) on Δt can cause a large reduction in the quality of the flow estimates, especially when the data is very noisy (1 or 2 repeats).

There are limitations to this approach of using a different gold standard for each method. In particular, the 6mm-smoothing results (A′) appears to perform very well (not shown), but it is clear from figure 4.9 that the model A′ target images (particularly of f) are not very good ones to aim for. The other methods (models A, B, and C) are of similar quality to each other, making this comparison more valid. In addition, it was found that using a single gold standard (B or C) for all methods gives very similar results to those shown.

4.10 Discussion

The combined prior provides a principled and adaptive way to use both non-spatial and spatial prior information simultaneously. It uses a parameterized covariance matrix which keeps a fixed variance value on the diagonal, while still allowing adjustable spatial smoothness. The modularity of the hierarchical framework means that different priors can be used on each parameter image; parameters without an informative non-spatial prior should use the spatial smoothness prior instead.

Since the adaptive parameter δ_k is buried in the combined prior's covariance matrix, $\mathbf{C}_k = \sigma_k^2 \exp(-\boldsymbol{\Delta}/\delta_k)$, rather than being easily factored out (as is the case with the spatial smoothness prior's adaptive parameter, in $\mathbf{C}_k = \alpha_k^{-1}(\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}$), analytic VB updates cannot be found. Instead, evidence optimization has been demonstrated as an effective way to infer on the distance scale factor δ_k and the forward model parameters $\mathbf{w}_*$, while the VB updates are still used for noise parameters ϕ_v and a_v and for any spatial smoothness weights α_k.

A simple approach makes it possible to infer on non-linear models without modifying the framework. The non-linear function $\mathbf{g}$ is replaced by its linear Taylor expansion $\overline{\mathbf{g}}$, and after each iteration it is re-centered on the latest point estimate to

minimize the errors between the two. This is very effective, and on these (reasonably well-behaved) models no convergence problems are observed.

In two neuroimaging applications, this work demonstrates that the combination of spatial and non-spatial prior information can provide noticeable benefits over either option taken individually. Consistent good behaviour is shown across a range of problems (from the non-spatial prior dominating to the spatial prior being most useful) and inference accuracy is noticeably improved when both pieces of information are relevant. This is demonstrated quantitatively using simulated FMRI data (constrained HRF shape model) and real multi-inversion ASL data (perfusion modelling). The benefits extend beyond the parameter that makes use of the combined prior. In both applications, estimates of the scale parameter (i.e. β_1 or f) were often improved when the combined prior was used on the shape parameter (r_2 or Δt) – presumably as an indirect effect of the improved estimates on the latter parameter.

It should be noted that the between-*parameter* correlations are taken into account, for example when calculating the updates on the signal parameters. The only time these covariances are discarded is in appendix B.2, where this approximation is used to speed up the estimation of the δ_k parameters.

Free energy

All of the quantitative analysis in this paper has looked at the estimated parameter images $\mathbf{m}_k$ and compared them to some ground truth image. Another useful evaluation technique would be Bayesian model comparison, which has been previously applied to spatial FMRI models (Penny et al., 2007). There are a potential complication in applying this technique to the models presented here. For non-linear forward models $\mathbf{g}$, it is worth noting that the free energy (for VB) or conditional evidence (for EO) is no longer a bound on the evidence of the actual model; using the approximation $\mathbf{g} \approx \overline{\mathbf{g}}$ can increase or decrease these estimates. Every time the model

is re-centered (setting $\mathbf{z}_v = \mathbf{m}_v$), the Jacobian $\mathbf{J}$ and the intercept $\overline{\mathbf{g}}(\mathbf{0})$ change, which is enough to make model comparison inexact. This concern may not be a problem in practice, and sampling-based inference methods such as Markov-chain Monte Carlo can be used to validate evidence estimates when non-linear models are involved (Friston et al., 2007).

Apart from these concerns relating to non-linear forward models, the full free energy can be used for model comparison or model averaging. The free energy for the hybrid method would have the same form as the VB free energy, except that it does not factor $N(\mathbf{m}_*, \mathbf{\Sigma}_*)$ over voxels. It is therefore also a lower bound on the model evidence (although a correction should be applied to account for the point estimate on δ_k).

This definition of free energy is also guaranteed to increase (or stay the same) in each of the update steps. In particular, note that maximizing the conditional evidence (using equation 4.15) is equivalent to the maximizing the free energy (using equation B.4) – only the estimation method is different. It can also be shown that the EO estimates of the signal parameter posterior distribution $N(\mathbf{m}_*, \mathbf{\Sigma}_*)$ are identical to a VB update applied to their joint distribution $q(\mathbf{w}_*)$.

There is one minor exception here: the use of the diagonal matrix $\widehat{\mathbf{\Sigma}}_k$ instead of full $\mathbf{\Sigma}_k$ in the α_k updates means that these updates maximize a slightly different definition of free energy (one that is factored over voxels). The main impact of this would be that smoothness hyperparameter α_k is somewhat underestimated. (It is therefore theoretically possible that the full free energy could be reduced by an α_k update step, but since the bias is consistent from iteration to iteration it would not affect convergence.) Methods that factor over voxels do not have access to the full $\mathbf{\Sigma}_k$ and therefore use $\widehat{\mathbf{\Sigma}}_k$ everywhere (including in their definition of free energy). For the hybrid VB-EO approach it would be possible to use the full $\mathbf{\Sigma}_k$ in place of $\widehat{\mathbf{\Sigma}}_k$, but that has not been done here because that would give model C an unfair advantage in

estimating the smoothness of the parameter 1 (the "scale" parameter) compared to models A and B.

Further validation

Using a Markov Chain Monte Carlo (MCMC) approach to validate the full model would likely be very difficult because of the strong dependencies across voxels, and between the voxels and the hyperparameters. A naive implementation could easily suffer from the same problems as the free energy maximization approach described in appendix B.1. However, MCMC validation can be a good way to assess the impact of non-linearity on the inference, as shown in the previous chapter and by Chappell et al. (2009). If non-linearity has a negligible effect on non-spatial inference then it is likely not a problem for this inference method either.

Instead, the hybrid inference method could be further validated by using a full evidence optimization approach to inference. A fully-EO method would seek to maximize the evidence expression $\mathscr{E}(\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a})$ (from equation B.10) by selecting point estimates for all of the hyperparameters and all of the noise estimates. Since this is a standard approach, it would alleviate any potential concerns to do with the factorization and the hybrid approach. However, it would have to be run repeatedly to ensure that the final linearization is kept close to the posterior mean.

Extensions to the combined prior

In each of the models used in this paper, the the first parameter (the scale parameter) is unsuitable for use with the combined prior because it has no informative non-spatial prior. Instead, the spatial smoothness prior ($\mathbf{C}_k = \alpha_k^{-1}(\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}$) is reused because its scaling parameter makes this prior automatically adapt to the scale of typical parameter values in $\mathbf{w}_k$. This means that the only difference between models A, B, and C is the prior used on the remaining parameters (the shape parameters). Another option would be to use a conventional Gaussian process prior such as the automatic

smoothness determination prior used by Sahani and Linden (2003), which infers both the length scale and a prior weight from the data. The modular approach described in this paper makes it easy to use whichever prior is most suitable for each parameter in the forward model.

Surprisingly, the combined prior could still provide effective spatial regularization even if the fixed non-spatial prior is quite weak. A purely non-spatial prior is ignored when the typical parameter values in $\mathbf{w}_k$ are much smaller than the prior standard deviation σ_k. This suggests that the combined GPP would also be ignored because it uses the same fixed variance. In fact, very large δ_k values are inferred in this situation, which brings the between-voxel correlations very close to 1 and this helps to compensate for the large prior variance. In the 1-D case it can be shown that this approaches a Markov random field prior ($\mathbf{C}_k = \alpha_k^{-1}\mathbf{L}^{-1}$) with weight $\alpha_k = \delta_k/2\sigma_k^2$. Numerical problems can occur as $\mathbf{C}_k$ becomes singular ($\delta_k \gg 10^3$), so in practice one should avoid using the combined prior with a very weak non-spatial component.

Computation speed

Inference using the spatial smoothness prior is very quick (comparable to completely non-spatial inference) due to the VB framework and the fact that the full covariance matrix $\mathbf{C}_k = (\alpha_k \mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}$ is never actually calculated. In contrast, inferring on the combined prior requires the repeated calculation of a $V \times V$ matrix inverse. This currently limits the size of problems that can use the combined prior to a few thousand voxels. There may be numerical approximations to make the full-sized problems computationally feasible. One immediate solution is to partition the data into separate inferences, such as analysing each slice separately. Since partitioning prevents spatial regularization across the boundary, a more structure-driven approach can be beneficial. Flandin et al. (2002) proposes parcellation based on clustered functional data, and Harrison et al. (2008) have developed a data-driven approach to

partitioning that tends to place the partition boundaries along the natural edges in the image (i.e., along the lines of steepest spatial gradient).

Conclusions

Despite their simple form, Gaussian process priors provide a great deal of flexibility and precise control over image structure. This makes it possible to create Bayesian formulations of advanced image analysis techniques, such as non-stationary smoothing based on the diffusion kernel (Harrison et al., 2007). By expressing priors in Gaussian process form, they can be incorporated into efficient approximate frameworks such as VB and EO and can become practical tools for analysing neuroimaging data in a probabilistic framework.

Chapter 5

Integrated Bayesian Decomposition and Decoding

5.1 Introduction

Multivariate methods for decoding brain signals have been increasingly used for decoding distributed signals in FMRI data and other neuroimaging modalities. This "brain reading" approach uses supervised learning techniques to find patterns that are predictive of the target variables of interest and may not appear in a mass-univariate analysis (Haynes and Rees, 2006). This is believed to be a powerful approach for finding distributed representation patterns because it relies on different combinations of spatial regions being activated in each state, rather than entirely independent spatial regions being used (Haxby et al., 2001; Norman et al., 2006).

As a concrete example of this, Haynes and Rees (2005) used a linear discriminant to predict whether a flashing visual stimulus was in one of two possible orientations. The target in this case was orientation of the stimulus in each trial (consisting of a 30-second block), and the neuroimaging data for each trial was the average activation voxel values of all voxels in a single visual area (either V1, V2, or V3) during the trial. They showed that the orientation could be decoded from each of the three regions with high accuracy (70–80%). When they made the stimulus imperceptible to the subject by adding masking, the information could still be strongly decoded from V1

($\sim 60\%$) while decoding from V2 and V3 was no better than chance. This shows that these "invisible" stimuli evoke a response in the early visual cortex, but the signal may not propagate into later areas.

Often this analysis is performed on classifying single FMRI volumes (e.g. Mourao-Miranda et al. 2005), but more generally the experiment is broken down into a series of trials, with some neuroimaging data associated with each trial.

The same basic approaches can be considered for other neuroimaging modalities, such as magnetoencephalography (MEG), electroencephalography (EEG), and Local Field Potential (LFP) recordings. In this case, the neuroimaging data needs to be divided into trials, with the neuroimaging data for each trial being an epoch of temporal or spatio-temporal recordings, generally time-locked to the event of interest in each trial. In this case, Event Related Potentials (ERPs) take the place of spatial decoding patterns. Spatio-temporal analysis is also possible in FMRI; for example, Mourao-Miranda et al. (2007) used a block of 14 consecutive FMRI volumes (concatenated into a single vector) as the neuroimaging data for each trial.

The main goals of a decoding-based analysis are to determine whether a behavioural variable can be decoded from the neuroimaging data (and how accurately) and to map the patterns of activity in which the variable is encoded in order to gain insight into the underlying neuronal representations. The challenge for these methods is that generalizable patterns must be found in a limited number of repeats of training data in the presence of both structured and unstructured noise.

This chapter presents a novel supervised learning approach in which the data decomposition and sparse decoding are learned simultaneously, as part of the same Bayesian model. This permits the decomposition to be shaped by the needs of decoding but balanced by the need to describe the data succinctly. This aims to improve the interpretability of the decomposition and the accuracy of decoding.

5.1.1 Notation and Terminology

The training data consists of a set of R trials (repeats), indexed by $r = 1 \ldots R$. Regardless of the type of neuroimaging data used, the data for trial r will be expressed as a $T \times 1$ vector, $\mathbf{y_r}$. In this chapter the trial data $\mathbf{y_r}$ is a purely temporal pattern, but it could equally well be a vector of spatial or spatio-temporal data corresponding to that trial. These are horizontally concatenated to produce the $T \times R$ matrix $\mathbf{Y}$.

To provide a common terminology for both generative and discriminative models, the trial-by-trial decoding targets will be referred to as "behavioural variables": these can be related to either stimulus or response measures, but they are generally observable and related to the behavioural (non-neuroimaging) aspects of the experiment. In a GLM these would be the explanatory variables (regressors) and in a decoding model one behavioural variable would be chosen as the target. There can be several behavioural variables in the experiment, together describing the properties of each trial. Trial r's behavioural variables are given by the $B \times 1$ vector $\mathbf{v}_r$, where B is the number of behavioural variables to be decoded simultaneously (in traditional decoding methods, $B = 1$). $\mathbf{V}$ is the $B \times R$ matrix $[\mathbf{v}_1|\mathbf{v}_2|\cdots|\mathbf{v}_R]$. Given the training data $\{\mathbf{Y}, \mathbf{V}\}$, the models learn their internal parameters, denoted by $\mathbf{\Theta}$.

5.1.2 Generative and Discriminative Models

A purely generative model (such as the general linear model) would generate the data distribution as a function of the behavioural variables, i.e.

$$\mathbf{y}_r \sim g(\mathbf{v}_r, \mathbf{\Theta}). \tag{5.1}$$

In the case of a GLM, the behavioural variables $\mathbf{V}^{\mathrm{T}}$ are used as one or more EVs and $\mathbf{\Theta}$ contains the regression coefficients and noise parameters characterising $\mathbf{e}_\mathbf{r}^{(\mathbf{Y})}$ and $\mathbf{e}_\mathbf{r}^{(\mathbf{V})}$. Typically these are evaluated to obtain measures of statistical significance; in spatial FMRI this produces maps of significantly-activating voxels, and in temporal

data this would show which parts of the signal are significantly varying with the behavioural measure. Generally this is not used as a predictive model (i.e. cross-validation with withheld test data), but instead the statistical significance of the regression coefficients is assessed directly to generate spatial maps.

As illustrated in figure 5.1(a–b), a decoding model can be considered the inverse of a generative model because it learns to decode a trial's behavioural variables $\mathbf{v}_r$ from that trial's neuroimaging data $\mathbf{y}_R$:

$$\mathbf{v}_r \sim f(\mathbf{y}_r, \boldsymbol{\Theta}) \tag{5.2}$$

where, as before, a new target vector $\hat{\mathbf{v}}$ is predicted by the new data vector in $f(\hat{\mathbf{y}}_r|\boldsymbol{\Theta})$. Linear decoding models try to find a pattern in the neuroimaging data that maps onto the behavioural variables as accurately as possible:

$$\mathbf{v}_r \approx \mathbf{D}\mathbf{y}_r + \mathbf{d_0} \tag{5.3}$$

where $\mathbf{D}$ is a $B \times T$ "decoding matrix" (each column of which is a decoding vector), and $\mathbf{d_0}$ is a $B \times 1$ offset to account for the means of the neuroimaging data and behavioural variables. The decoding vectors themselves are useful to examine because they should approximate the pattern of activity related to the behavioural variable, weighted to avoid noise. Finding $\mathbf{D}$ is an ill-posed problem and regularization methods are needed to obtain a generalizable solution. A common approach is to assume sparseness, for example using Bayesian techniques such as automatic relevance determination (ARD) and the relevance vector machine (RVM) (Tipping, 2000) or non-Bayesian methods like support vector machines (SVMs) (Bishop, 2006).

Friston et al. (2008a) proposed a sparse Bayesian decoding approach which solves the ill-posed problem by using only those features that increase the free energy. This automatically tunes the model to be just complicated enough to explain the data, and this tends produce generalizable decodings; this built-in tuning is a benefit compared to methods like SVMs which need to use nested cross-validation

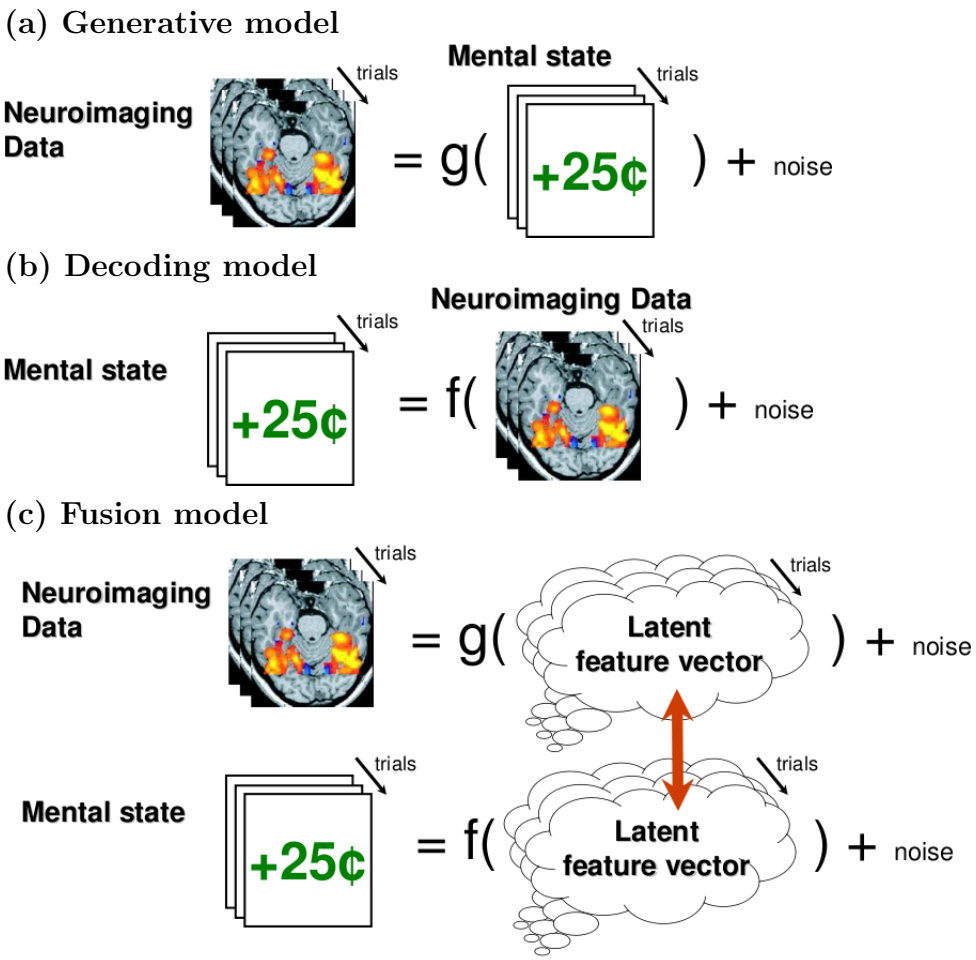


Figure 5.1: Three approaches to analysing neuroimaging: (a) Generative modelling (including the GLM) use the changing value of each behavioural variable in each TR (or trial) as a regressor to explain the time-courses (trial-courses) of the observed neuroimaging data. In this illustration, the behavioural variable is whether or not a reward was given in the task, and the result of the GLM inference is a map showing where the voxel's data is significantly related to that behavioural variable. (b) Decoding models attempt to find the patterns in the neuroimaging data that predict or explain the behavioural variable. (c) The proposed fusion approach simultaneously decomposes the data into a short "latent feature vector" and uses that same feature vector to decode the behavioural variables. By performing both parts of the inference simultaneously, the patterns selected in the decomposition part will be informed by the decoding and this may result in different patterns being inferred that are more interpretable or may be better tuned to the task of decoding.

to maximize generalizability. A number of potential basis sets were considered in their research, including smooth basis functions, singular vectors (found by an SVD decomposition of the data set) and support vectors (where the feature set was defined by the voxel patterns in each trial). Using FMRI data, they found that the best feature set was individual voxels. To deal with the enormous number of hyperparameters in this model (one per voxel) an iterative greedy search mechanism was used to assist with dimensionality reduction. The SVD was the next best feature set to use.

However, the SVD features used by this approach are fixed point estimates determined directly from the data. By inferring them probabilistically as part of the model, it may be possible to refine them using information from the behavioural variable decoding. This chapter presents a "fusion" approach that simultaneously infers the decomposition of the neuroimaging data and the decoding of the behavioural variables.

5.1.3 Fusion Approach

This fusion approach is implemented in a novel Bayesian framework which links a generative model of the data with a decoding model of the behavioural variables, by introducing a shared, hidden latent state for each trial that links the two models probabilistically. This is illustrated in figure 5.1(c).

This hidden state in trial r is given by an $L \times 1$ vector, denoted by $\mathbf{h}_r$:

$$\mathbf{y}_r \sim g(\mathbf{h}_r, \boldsymbol{\Theta}^{(\mathbf{Y})}) \tag{5.4}$$

$$\mathbf{v}_r \sim f(\mathbf{h}_r, \boldsymbol{\Theta}^{(\mathbf{V})}) \tag{5.5}$$

Importantly, the learned model parameters for f and g are completely separate; $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}^{(Y)}, \boldsymbol{\Theta}^{(V)}\}$. Although the latent state $\mathbf{h}_r$ can never be measured directly, it provides a probabilistic link between the measurable variables by balancing the incoming information provided by both the data and target vectors and feeding this

shared information back to both equations. The individual elements of $\mathbf{h}$ have no fixed meaning and their interpretation depends entirely on the correspondence between the learned model parameters in $\boldsymbol{\Theta}^{(\mathbf{Y})}$ and $\boldsymbol{\Theta}^{(\mathbf{V})}$. For some purposes, $\mathbf{h}_r$ can even be integrated out completely; see section 5.3.2. Having a simple intermediate space means that there is a great deal of flexibility to use the type of model that is most appropriate for each data type.

5.2 Linear Data Fusion Model

The fusion model follows from equations 5.4–5.5 by assuming that the functions g and f are linear:

$$\mathbf{y}_r = \boldsymbol{\mu}^{(\mathbf{Y})} + \mathbf{W}\mathbf{h}_r + \mathbf{e}_r^{(\mathbf{Y})} \tag{5.6}$$

$$\mathbf{v}_r = \boldsymbol{\mu}^{(\mathbf{V})} + \mathbf{A}\mathbf{h}_r + \mathbf{e}_r^{(\mathbf{V})} \tag{5.7}$$

Using limited training data, the model must learn the matrices $\mathbf{W}$ and $\mathbf{A}$, the means $\boldsymbol{\mu}^{(Y)}$ and $\boldsymbol{\mu}^{(V)}$, and the noise parameters. To provide regularization to this ill-posed problem, adaptive priors are place on $\mathbf{W}$ and $\mathbf{A}$ to encourage sparsity (i.e. using only those elements of the latent space $\mathbf{h}_r$ that are useful). Note that since both of these priors encourage sparsity, they will automatically eliminate components that are not needed; therefore it is useful to start with a large number of components L and have some of them eliminated. The full matrix diagram is shown in figure 5.2.

The two noise models are not the same because they are modelling different types of data:

$$\mathbf{e}_r^{(\mathbf{Y})} = \mathrm{N}\big(\mathbf{0}, \lambda_r^{-1}\mathbf{I}\big) \tag{5.8}$$

$$\mathbf{e}_r^{(\mathbf{V})} = \mathrm{N}\big(\mathbf{0}, \boldsymbol{\Gamma}^{-1}\big) \tag{5.9}$$

The model parameters to be learned are grouped together as $\boldsymbol{\Theta}^{(\mathbf{Y})} = \big\{\mathbf{W}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \boldsymbol{\mu}^{(Y)}\big\}$ and $\boldsymbol{\Theta}^{(\mathbf{V})} = \big\{\mathbf{A}, \boldsymbol{\Gamma}, \boldsymbol{\alpha}, \boldsymbol{\mu}^{(V)}\big\}$, where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_R]^{\mathrm{T}}$, and $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ are the ARD hyperparameters for $\mathbf{A}$ and $\mathbf{W}$ respectively.

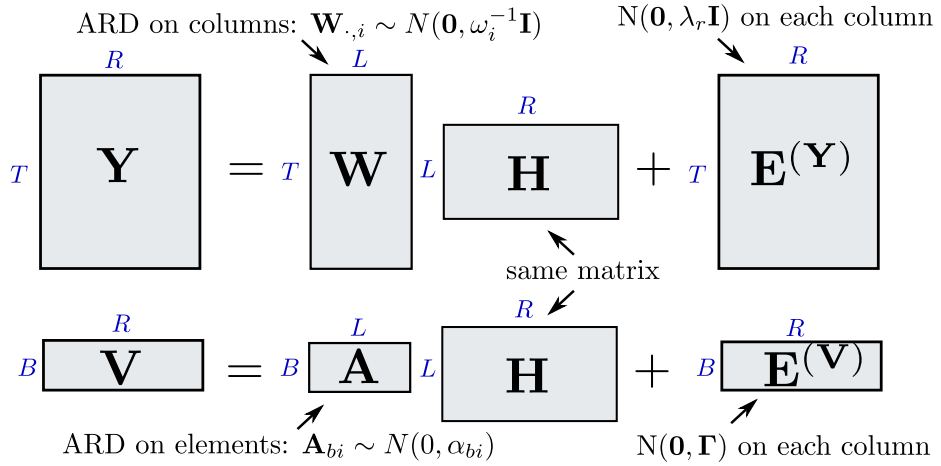


Figure 5.2: Matrix diagram of the fused PCA and sparse decoding model. The top equation models the neuroimaging data $\mathbf{Y}$ while the bottom equation decodes the behavioural variable(s) $\mathbf{V}$.

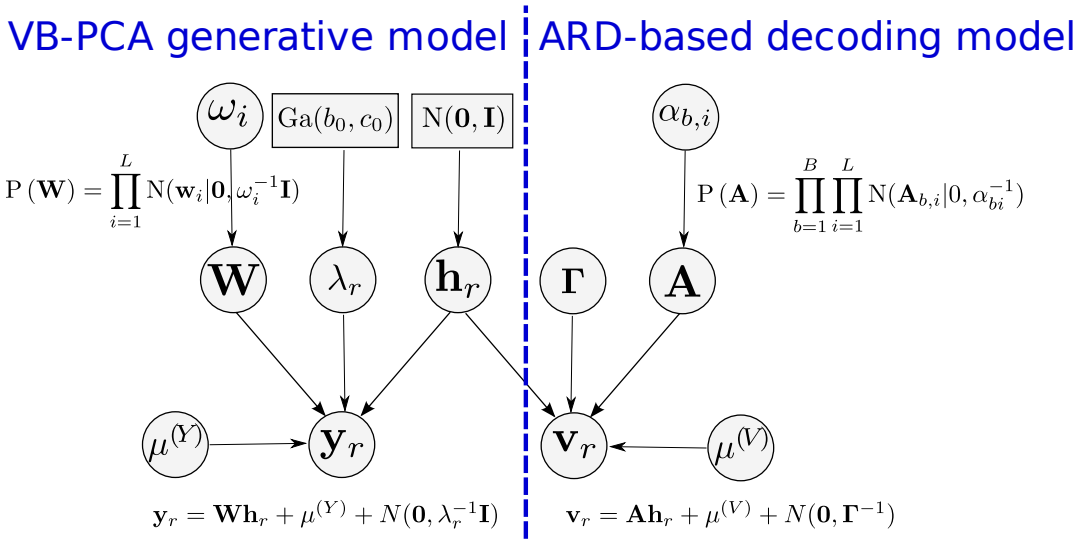


Figure 5.3: A graphical illustration of the probabilistic dependencies underlying the fusion model. The left side is a generative VB-PCA model of the neuroimaging data $\mathbf{y}_r$ and the right side shows the sparse multivariate decoding of the behavioural variables $\mathbf{v}_r$. The two sides are linked only through shared latent state vectors $\mathbf{h}_r$. The quantities in circles are random variables, while informative priors are shown in rectangles and non-informative priors have been omitted for simplicity. The equations show some of the key prior distributions over model parameters. Elements with subscripts are replicated across repeats (r), components (i) and/or target variables (b).

The full graphical model is shown in figure 5.3. The likelihood is the product of two parts:

$$\mathrm{P}\big(\mathbf{Y}|\mathbf{H},\boldsymbol{\Theta}^{(\mathbf{Y})}\big) = \prod_{r=1}^{R}\mathrm{P}\big(\mathbf{y}_r|\mathbf{h}_r,\mathbf{W},\lambda_r,\boldsymbol{\mu}^{(Y)}\big) \tag{5.10}$$

$$\mathrm{P}\big(\mathbf{V}|\mathbf{H},\boldsymbol{\Theta}^{(\mathbf{V})}\big) = \prod_{r=1}^{R}\mathrm{P}\big(\mathbf{v}_r|\mathbf{h}_r,\mathbf{A},\boldsymbol{\Gamma},\boldsymbol{\mu}^{(V)}\big) \tag{5.11}$$

where $\mathbf{H} = [\mathbf{h}_1|\mathbf{h}_2|\cdots|\mathbf{h}_R]$ and the prior $\mathrm{P}(\boldsymbol{\Theta}) = \mathrm{P}\big(\boldsymbol{\Theta}^{(\mathbf{Y})}\big)\,\mathrm{P}\big(\boldsymbol{\Theta}^{(\mathbf{V})}\big)$ is composed of

$$\mathrm{P}\big(\boldsymbol{\Theta}^{(\mathbf{Y})}\big) = \mathrm{P}\big(\boldsymbol{\mu}^{(Y)}\big)\,\mathrm{P}(\mathbf{W}|\boldsymbol{\omega})\,\mathrm{P}(\boldsymbol{\omega})\,\mathrm{P}(\boldsymbol{\lambda}) \tag{5.12}$$

$$\mathrm{P}\big(\boldsymbol{\Theta}^{(\mathbf{V})}\big) = \mathrm{P}\big(\boldsymbol{\mu}^{(V)}\big)\,\mathrm{P}(\mathbf{A}|\boldsymbol{\alpha})\,\mathrm{P}(\boldsymbol{\alpha})\,\mathrm{P}(\boldsymbol{\Gamma})\,. \tag{5.13}$$

The remainder of this section explains the parts of this model in greater detail.

5.2.1 VB-PCA Generative Model of Neuroimaging Data

Equation 5.6 models the neuroimaging data's decomposition into linear components. In isolation, it corresponds to a variational PCA (Bishop, 1999) with the following priors:

$$\mathrm{P}(\mathbf{W}|\boldsymbol{\omega}) = \prod_{i=1}^{L}\mathrm{N}\big(\mathbf{W}_{\cdot,i}|\mathbf{0},\omega_i^{-1}\mathbf{I}\big)\,, \qquad \mathrm{P}(\mathbf{H}) = \prod_{r=1}^{R}\prod_{i=1}^{L}\mathrm{N}(\mathbf{H}_{ir}|0,1)\,. \tag{5.14}$$

The ARD priors have a non-informative scale-free Gamma-distributed prior on each element: $\mathrm{P}(\boldsymbol{\omega}) = \prod_{i=1}^{L}\mathrm{Ga}(\omega_i|10^{12},10^{-12})$. Therefore, for the i^{th} component, the timecourse ($\mathbf{W}_{\cdot,i}$) tends to absorb the scale while the trial weightings ($\mathbf{H}_{i,\cdot}$) tend to have unit scale (root-mean-square). When a component is eliminated, $\langle\boldsymbol{\omega}\rangle \to \infty$ causes $\langle\mathbf{W}_{\cdot,i}\rangle \to \mathbf{0}$ and $\langle\mathbf{H}_{i,\cdot}\rangle \to \mathbf{0}^{\mathrm{T}}$.

Of course, in this work equation 5.6 is not inferred in isolation but in the context of an integrated decomposition and decoding model, so equation 5.7 also has an influence on the inferred latent state $\mathbf{H}$.

5.2.2 Per-trial Noise Estimates

Some types of neuroimaging data suffer from the presence of occasional but strong external artefacts. In practice these artefacts play a strong role in defining the PCA components thus making them sub-optimal for explaining brain activity. These noisy trials can be considerably downweighted by modelling a different noise precision λ_r for each trial of the data. The prior distribution is given by

$$\mathrm{P}(\boldsymbol{\lambda}_r) \quad = \quad \mathrm{Ga}(\boldsymbol{\lambda}_r | b_{\lambda 0}, c_{\lambda 0}) \tag{5.15}$$

Ideally, an uninformative prior would be used (e.g. $b_{\lambda 0} = 10^{12}$, $c_{\lambda 0} = 10^{-12}$) to avoid biasing the noise estimate. However, it was found that these unconstrained noise parameters can interact badly with the PCA model to produce a degenerate result: a small number of trials were inferred to have essentially no noise at all, and the columns of $\mathbf{W}$ modelled the neuroimaging data in these trials exactly.

Using a strong prior is one way to provide regularization to avoid this. The prior was determined empirically, with a mean roughly equal to the noise precision (estimated using the same noise for all repeats) and the shape parameter being just strong enough to reliably avoid the degenerate solution. For the LFP data that will be analysed, the prior used is $\mathrm{P}(\lambda) = \mathrm{Ga}(10^{-4}, 10^2)$. This effectively sets a noise floor which avoids the problem of PCA components overfitting the precise shape of a small number of trials. This may not be a problem with the model itself, but may instead occur as a result of the VB factorization used for inference (which assumes there is no posterior correlation between $\boldsymbol{\lambda}$ and $\mathbf{W}$).

One side effect of using a per-trial noise model is that it becomes important to explicitly model the mean $\boldsymbol{\mu}^{(\mathbf{Y})}$ rather than just demeaning the data as a preprocessing step. This is because the estimated mean may change when the individual trials are re-weighted based on their noise precisions.

5.2.3 Multivariate Behavioural Noise Model

A general Gaussian noise model is used to model different noise levels for each behavioural variable as well as the noise correlation between the variables. The correlation matrix $\mathbf{\Gamma}^{-1}$ can be any positive-definite matrix, so the full correlation structure is learned from the data. This is feasible because the number of behavioural variables B is quite small.

In principle, the full multivariate noise model can even deal gracefully with the maximally-degenerate case where the same EV has been included twice – the noise will be perfectly correlated and the data from those EVs will be weighted almost the same as if the EV had been included only once. Under an uncorrelated noise model this situation would artificially double the weight of the data component and lead to overfitting, generally by using a component of the latent space $\mathbf{H}$ just to model the behavioural variables' residuals. For example, in a decision-making experiment it might be interesting to decode both the "reward/non-reward" binary variable alongside the "amount won/lost" continuous behavioural variable. Using the correlated noise model means that several related behavioural variables like this can be included without worrying about them biasing the results in that way.

The natural extension of the Gamma distribution to positive-definite precision matrices is the Wishart distribution. The uninformative $\mathrm{P}(\mathbf{\Gamma}) = \mathrm{Wishart}(10^{12}\ \mathbf{I}, 10^{-12})$ is used.

5.2.4 Sparse Multivariate Regression

Encouraging sparsity in the sparse weight matrix $\mathbf{A}$ automatically eliminates components that are only weakly useful for decoding and thereby reduces the risk of overfitting. A separate ARD prior is placed on each *element* of the regression

matrix $\mathbf{A}$:

$$\mathrm{P}(\mathbf{A}|\alpha) \;=\; \prod_{i=1}^{L}\prod_{b=1}^{B}\mathrm{N}\big(\mathbf{A}_{b,i}|0,\alpha_{b,i}^{-1}\big) \tag{5.16}$$

with an uninformative scale-free prior on the ARD hyperparameters

$$\mathrm{P}(\alpha) \sim \mathrm{Ga}\big(10^{12},10^{-12}\big)\,. \tag{5.17}$$

This means that each behavioural variable b has to separately justify its use of a particular component i of $\mathbf{H}$.

Another option would be to put ARD on each column, i.e.

$$\mathrm{P}(\mathbf{A}|\alpha) \;=\; \prod_{i=1}^{L}\mathrm{N}\big(\mathbf{A}_{\cdot,i}|0,\alpha_{i}^{-1}\mathbf{I}\big)\,. \tag{5.18}$$

This basically treats the behavioural variables $\mathbf{v}_r$ as a single rotation-invariant vector, even though the variables might have completely different scales. By forcing all behavioural variables to have the same sparsity structure, this would also allow a weak variable to piggyback off a stronger variable, which could cause overfitting in some variables. A less obvious consequence of this is that it would increase the complexity cost of including a component that is only useful for decoding one of the behavioural variables, because the model must allow that component to explain *all* B variables rather than just the one. In practice, this tends to lead to the model eliminating too many components.

As a result of this, the elementwise prior (equation 5.16) is used.

5.3 Variational Bayes Implementation

Inference on the fused model is performed using Variational Bayes. The posterior distribution is approximated by the factorized posterior

$$P(\boldsymbol{\Theta},\mathbf{H}|\mathbf{Y},\mathbf{V}) \approx \mathrm{P}'\big(\mathbf{W},\boldsymbol{\mu}^{(\mathbf{Y})}\big)\,\mathrm{P}'(\boldsymbol{\omega})\,\mathrm{P}'(\boldsymbol{\lambda})\,\mathrm{P}'\big(\mathbf{A},\boldsymbol{\mu}^{(\mathbf{V})}\big)\,\mathrm{P}'(\boldsymbol{\alpha})\,\mathrm{P}'(\boldsymbol{\Gamma})\,\mathrm{P}'(\mathbf{H})\,. \tag{5.19}$$

This factorization leads to analytic expressions for all of the VB updates. The posterior distribution of these components take the form of a matrix normal distribution on $[\boldsymbol{\mu}^{(V)}, \mathbf{W}]$, a multivariate normal distribution on each $\mathbf{h}_r$, a $BL \times 1$ multivariate normal distribution on $\text{vec}[\boldsymbol{\mu}^{(Y)}, \mathbf{A}]$, Gamma distributions on $\boldsymbol{\omega}_i$, $\boldsymbol{\alpha}_{b,i}$ and $\boldsymbol{\lambda}_r$, and a Wishart distribution on $\boldsymbol{\Gamma}$. These updates are stated and explained in appendix C.1.

5.3.1 Initialization

$\mathbf{W}$ and $\mathbf{H}$ are initialized using a standard PCA decomposition (Jolliffe, 2002). Very similar results can be obtained by initializing with random numbers, but this slows convergence and produces components in a random order. The $\mathbf{A}$ matrix is initialized using its VB update (with the hyperparameters $\boldsymbol{\alpha}$ initially set close to 0, i.e. an unconstrained prior).

5.3.2 Equivalent Linear Decoding Matrix

In order to make predictions on unseen training data, it is convenient to produce a linear decoding matrix analogous to the $\mathbf{D}$ matrix described in equation 5.3. The rows of this can also be plotted to see the equivalent decoding vector that maps the neuroimaging data onto each of the B behavioural variables.

First, note that the fusion model can be stated as a joint Gaussian distribution between each repeat's neuroimaging data $\mathbf{y}_r$ and its behavioural variables $\mathbf{v}_r$, which removes the latent state vectors $\mathbf{h}_r$ from the equations. By stacking equations 5.6 and 5.7, the joint likelihood $\text{P}(\mathbf{v}_r, \mathbf{y}_r | \boldsymbol{\Theta}, \mathbf{h}_r)$ is expressed as

$$\begin{bmatrix} \mathbf{y}_r \\ \mathbf{v}_r \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(\mathbf{Y})} \\ \boldsymbol{\mu}^{(\mathbf{V})} \end{bmatrix} + \begin{bmatrix} \mathbf{W} \\ \mathbf{A} \end{bmatrix} \mathbf{h}_r + \text{N}\left(\mathbf{0}, \begin{bmatrix} \lambda_r^{-1}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}^{-1} \end{bmatrix}\right) \tag{5.20}$$

and integrating out $\mathbf{H}_r$ (recall that the prior on $\mathbf{H}_r$ is $\text{P}(\mathbf{h}_r) = \text{N}(\mathbf{h}_r | \mathbf{0}, \mathbf{I})$) yields

$$\begin{bmatrix} \mathbf{y}_r \\ \mathbf{v}_r \end{bmatrix} \sim \text{N}\left(\begin{bmatrix} \boldsymbol{\mu}^{(\mathbf{Y})} \\ \boldsymbol{\mu}^{(\mathbf{V})} \end{bmatrix}, \begin{bmatrix} \mathbf{W}\mathbf{W}^{\text{T}} + \lambda_r^{-1}\mathbf{I} & \mathbf{W}\mathbf{A}^{\text{T}} \\ \mathbf{A}\mathbf{W}^{\text{T}} & \mathbf{A}\mathbf{A}^{\text{T}} + \boldsymbol{\Gamma}^{-1} \end{bmatrix}\right) \tag{5.21}$$

Now that the latent state is removed, this same joint distribution can be assumed for an unseen test trial with neuroimaging data $\hat{\mathbf{y}}$ and behavioural variables $\hat{\mathbf{v}}$, using the posterior distributions on $\boldsymbol{\Theta}$ inferred from the training data.

This can be rearranged (see appendix C.2) to find that $\mathrm{P}(\hat{\mathbf{v}}|\hat{\mathbf{y}}, \boldsymbol{\Theta})$ is given by a normal distribution with mean

$$\mathbf{m}_r = \mathbf{A}\mathbf{W}^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \lambda_r^{-1}\mathbf{I})^{-1}\mathbf{y}_r. \tag{5.22}$$

This expression works for point estimates of $\mathbf{W}$ and $\mathbf{A}$, but not if $\mathbf{W}$ is probabilistic. The natural way to compute this from a VB posterior is to first infer the latent state for the test trial ($\mathrm{P}'\left(\hat{\mathbf{h}}\right)$) from $\hat{\mathbf{y}}$ (assuming all of the other posteriors remain fixed) and then calculate $\mathrm{P}'(\hat{\mathbf{v}})$ from that. This results in the normal predictive distribution with mean

$$\begin{aligned}
\mathrm{P}'(\hat{\mathbf{v}}) &= \mathrm{N}\left(\hat{\mathbf{m}}, \hat{\mathbf{S}}\right) && (5.23)\\
\hat{\mathbf{m}} &= \langle \mathbf{A} \rangle \left\langle \hat{\mathbf{h}} \right\rangle && (5.24)\\
&= \langle \mathbf{A} \rangle \langle \mathbf{W} \rangle^{\mathrm{T}} \left(\left\langle \mathbf{W}\mathbf{W}^{\mathrm{T}} \right\rangle + \lambda^{-1}\mathbf{I} \right)^{-1} (\hat{\mathbf{y}} - \boldsymbol{\mu}^{(Y)}) + \boldsymbol{\mu}^{(V)} && (5.25)
\end{aligned}$$

and therefore $\hat{\mathbf{m}} = \mathbf{D}\hat{\mathbf{y}} + \mathbf{d_0}$ with

$$\begin{aligned}
\mathbf{D} &= \langle \mathbf{A} \rangle \langle \mathbf{W} \rangle^{\mathrm{T}} (\langle \mathbf{W}\mathbf{W}^{\mathrm{T}} \rangle + \lambda^{-1}\mathbf{I})^{-1} && (5.26)\\
\mathbf{d_0} &= \boldsymbol{\mu}^{(V)} - \mathbf{D}\boldsymbol{\mu}^{(Y)} && (5.27)
\end{aligned}$$

such that the maximum a posteriori estimate is given by $\langle \hat{\mathbf{v}}|\hat{\mathbf{y}} \rangle = \mathbf{D}\,\hat{\mathbf{y}} + \mathbf{d}_0$.

The decoding vector is useful in itself since it shows exactly what patterns of the data are used for predicting the behavioural data from the neuroimaging data.

5.3.3 Simplified Approaches for Comparison

Two simplified methods were implemented to provide a baseline for assessing performance. The first approach, referred to as the "Fixed PCA" approach, uses

the standard PCA decomposition on $\mathbf{Y}$ to create the fixed temporal basis set $\mathbf{W}$ and the latent feature data $\mathbf{H}$. This is not a lossless PCA decomposition, but instead the number of components is fixed in advance. In this model, the maximum number of components ($L = 10$ for simulated data and $L = 30$ for real data) was kept. The latent feature matrix $\mathbf{H}$ is then taken as a constant and inference proceeds using only the decoding equation 5.7 and the related sparse priors.

It is also interesting to see which benefits come simply from using the VB-PCA instead of the standard PCA in equation 5.6. Therefore, a second approach, the "Sequential" approach, is also evaluated. This method first solves the VB-PCA equation 5.6 without equation 5.7. This will automatically pick the appropriate number of components to keep and will also infer the trial-by-trial noise modelling. The resulting fixed point estimates of $\mathbf{H}$ are then used as constant inputs to equation 5.7, which is solved as with VB using the same priors as usual.

The full "Fusion" method infers equations 5.6 and 5.7 simultaneously using VB, which passes probabilistic information between the equations via the posterior distribution on the $\mathbf{H}$ parameters. This difference between the sequential method and the full fusion method is that the sequential model prevents this feedback between the equations from occurring.

5.4 Simulated Data

A simulated data set was constructed to test the decoding performance of the method under controlled conditions. The simulated timecourse data has 100 timepoints within each trial and seven components; for easy visualization, these component patterns are non-overlapping rectangular blocks 10 timepoints long, with heights of 1, 0.9, 0.8, 0.7, 0.6, 0.5 and 0.4. The fact that this is a very simple structure should not influence the PCA decomposition because PCA is invariant to rotations of the input space; only the magnitude and covariance of the components are relevant. There is no spatial

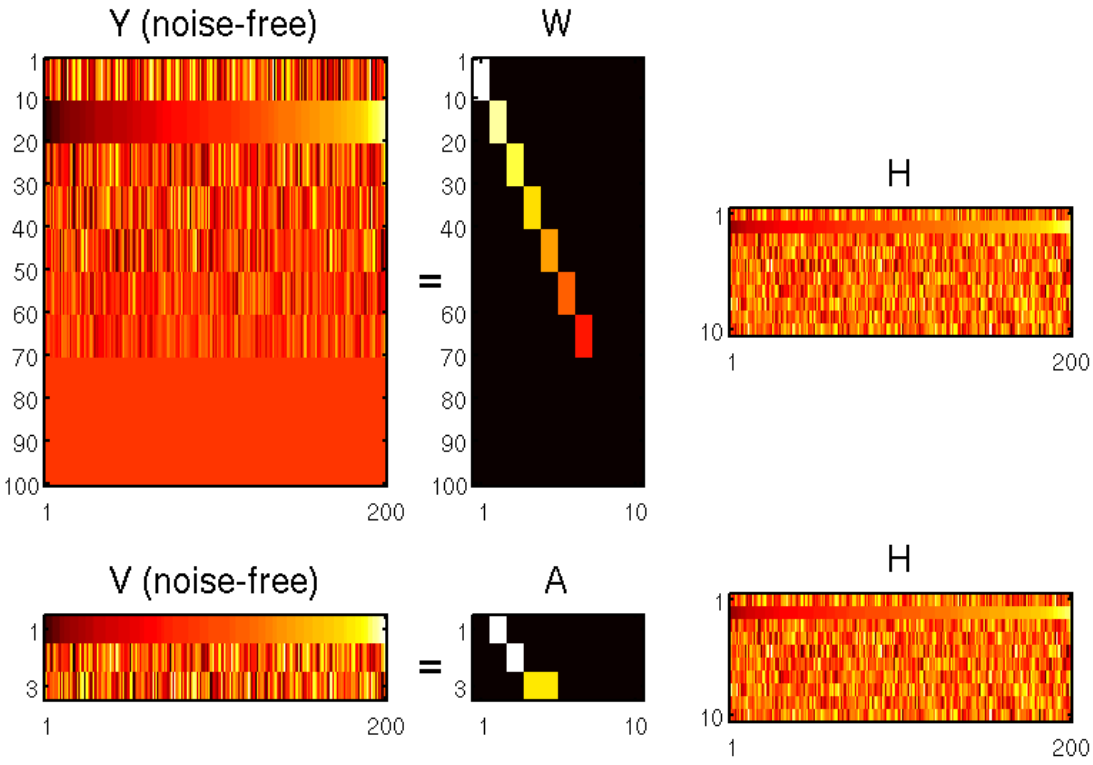


Figure 5.4: Matrix diagrams showing the simulated data. The true $\mathbf{W}$ and $\mathbf{A}$ matrices are fixed (see text) and $\mathbf{H}$ is unit white noise. For this illustration, the repeats are sorted by the second row of $\mathbf{H}$ to show how this row propagates into rows (timepoints) 11–20 of the neuroimaging data $\mathbf{Y}$ and row 1 of the three behavioural variables in $\mathbf{V}$. The data used for inference has Gaussian noise added to this, and $\mathbf{H}$ is discarded, so the components have to be re-learned from the structure of the data.

dimension in this data (i.e. it is like decoding from a single FMRI voxel or MEG sensor).

The neuroimaging data vector $\mathbf{y}_r$ for each trial was generated by multiplying each of these component patterns by the appropriate element of a randomly-drawn latent state $\mathbf{h}_r \sim N(0, \mathbf{I})$. A vector of behavioural variables ($B = 3$) was also generated for each trial: $\mathbf{v}_r = \left[(\mathbf{h}_r)_2, (\mathbf{h}_r)_3, \frac{1}{\sqrt{2}}\left((\mathbf{h}_r)_4 + (\mathbf{h}_r)_5\right)\right]^{\mathrm{T}} + \mathrm{N}(\mathbf{0}, \mathbf{\Gamma}^{-1})$. Each behavioural variable therefore has zero mean and unit variance (before noise is added). The corresponding $\mathbf{W}$ and $\mathbf{A}$ matrices that produced this data set are shown in the centre of figure 5.4. The model is kept simple by fixing the means ($\mu_{\mathbf{Y}} = \mathbf{0}$ and $\mu_{\mathbf{V}} = \mathbf{0}$) and

assuming that the noise level does not vary between repeats ($\lambda_r = \lambda$). A constant level of white noise was added to all trials, using $\lambda^{-1} = (1.5)^2$ and $\mathbf{\Gamma}^{-1} = (0.1)^2\mathbf{I}$. The training set consisted of $R = 200$ trials and the testing set had 10000 trials. This was repeated for 100 realizations of the simulation, i.e. using 100 different training and testing sets, all using the same true $\mathbf{W}$ and $\mathbf{A}$ matrices but different randomly-generated $\mathbf{H}$ matrices and additive noise.

5.5 Results on Simulations

Figure 5.5 shows the $\mathbf{W}$ and $\mathbf{A}$ matrices inferred by each method in a typical realization of the simulated data. Comparing the fusion model to the Fixed PCA model, the extracted data patterns $\mathbf{W}_{\cdot,i}$ more cleanly recover the true block structure. The weight matrices $\mathbf{A}$ are also sparser; this is because the timepoints that are useful for decoding are spread across several different components in the Fixed PCA, while in the fusion model the components have been rotated to reduce the amount of mixing. Sometimes other combinations were observed, for example figure 5.5(b) shows that the behavioural data has dominated and mixed the two components into a single component $i = 4$ (and a difference component, $i = 5$, which is not useful for decoding). This also provides a stable explanation of the data, and both of these would produce good decoding vectors.

Using the Sequential approach instead of the Fixed PCA reduces the number of components but does not noticeably change the component patterns. Interestingly, the components that were eliminated by the Sequential approach (e.g. $i = 6$ to $i = 10$ in figure 5.5(a) and (b)) are therefore not used for decoding $\mathbf{V}$, but in the Fixed PCA these components are not eliminated and were slightly useful for fitting the behavioural variables.

There is also somewhat reduced background noise in these components in the fused model: averaged across the first five components and 100 realizations, the fraction of

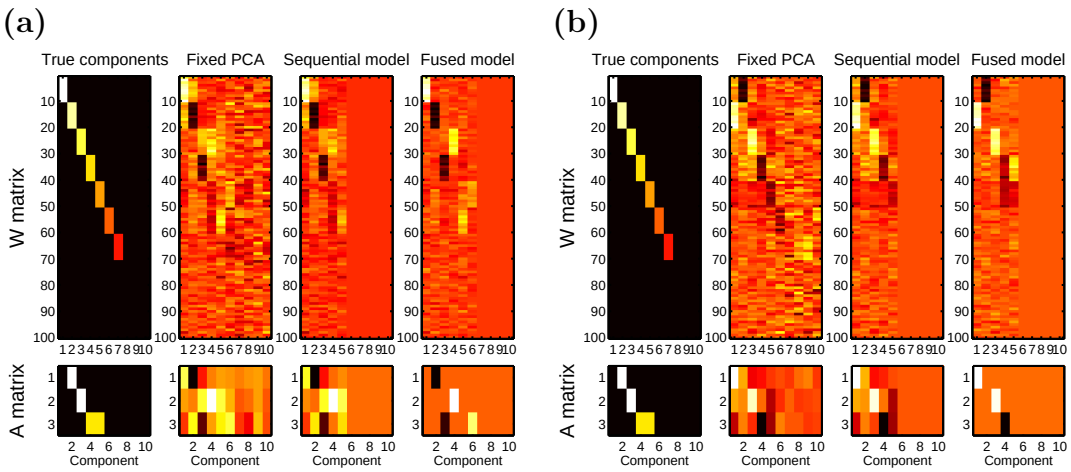


Figure 5.5: The $\mathbf{W}$ and $\mathbf{A}$ matrices inferred on simulated data by the three different approaches, shown for two typical realizations of the simulated data. In realization (a), the block structure of the components is recovered cleanly, assisted by the information from the decoding stage. In realization (b), the two components that are useful for decoding $b = 3$ are merged into a single component, because this yields an even sparser $\mathbf{A}$. In general, it was found that reducing the noise in $\mathbf{Y}$ leads to the left case (correctly describing the neuroimaging data) and increasing the noise in $\mathbf{Y}$ leads to the second case dominating most realizations (because $\mathbf{V}$ is more informative than $\mathbf{Y}$).

the component energy ($|\mathbf{w}_q|^2$) that was contained in the last 30 (pure noise) timepoints was 7.3% for the original PCA and the sequential VB-PCA, and 6.1% in the fused model. This suggests that the information from the decoding stage is being used to separate the signal from the noise in the components that are relevant for decoding.

Figure 5.6 shows how this improved component selection translates into better decoding vectors and more accurate prediction of unseen behavioural data. The gold standard in part (a) is the optimal decoding matrix $\mathbf{D_{opt}}$, found by using the true values for $\mathbf{W}$ and $\mathbf{A}$ in equation 5.22. This optimal decoding accuracy is the limit that would be reached by any appropriate learning method given an infinite amount of training data. In this simulation, the fusion model makes better use of the limited number of trials, producing a decoding vector that is closer to optimal than either of the other methods. Note that all three approaches appear to underestimate the size of the decoding vector. This is because these vectors are based on noisy estimates

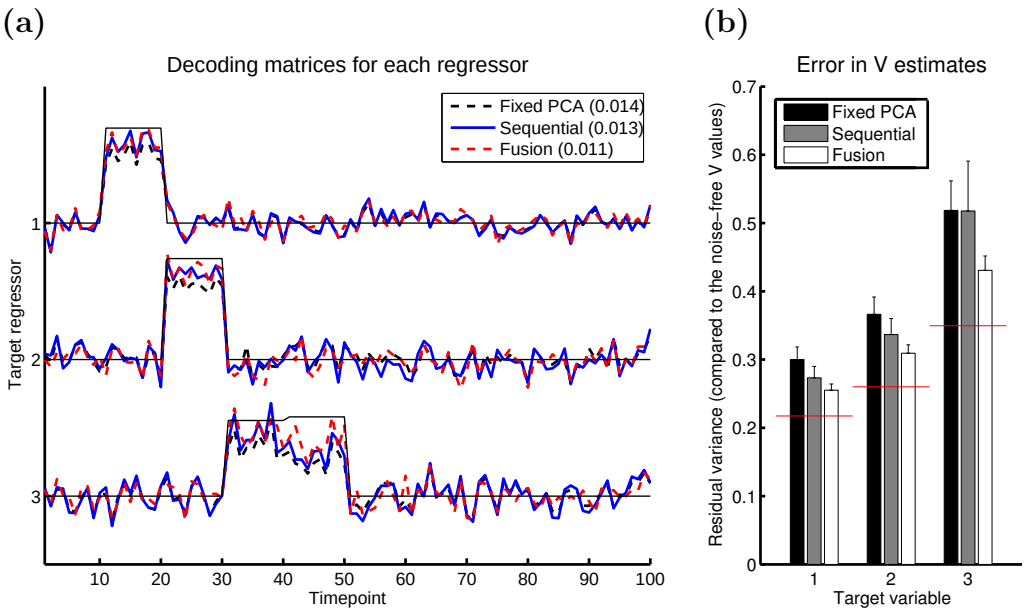


Figure 5.6: (a) Decoding vectors resulting from a typical realization of the simulated data. Although the sequential and fusion approaches give quite similar decoding vectors, the fusion approach's ones are slightly closer to the true optimal decoding vectors $\mathbf{D}_{\mathbf{optimal}}$ (the thin black line) – the RMS difference is the number given in parentheses in the figure legend. This is reflected in better accuracy on new test data: part (b) shows the residual errors in each method's estimates of the noise-free $\mathbf{v}_r$ target values – lower is better. Error bars show the *standard deviation* across 100 independent realizations of the simulation (the standard error of the mean is $1/10^{\text{th}}$ of this).The red line indicates the accuracy of optimal decoding, which is limited only by the noise in $\mathbf{y}_r$ (since noise in $\mathbf{v}_r$ is not used in this situation). Decoding using the fusion method (white bars) is more accurate than using a fixed PCA decomposition (black) at estimating all three targets. The sequential approach (grey) only gives part of this benefit.

of the component patterns, so there is a tradeoff; increasing the overall amplitude of a component would increase signal, but also increases the amount of noise that is added to the estimate.

The optimal decoding matrix illustrates the limits to decoding performance, limited only by noise, not by parameter estimation error. This is given by the red lines in figure 5.6(b). The fusion method is more accurate, and is actually about 30–50% closer to optimal decoding than either of the other approaches.

It is also interesting to note that the sequential approach performed better than the Fixed PCA approach, even though the only difference is that the Fixed PCA

approach has a few extra components. These are used for decoding and this reduces the Fixed PCA's overall decoding accuracy. The exception to this is in target variable 3: notice that instead of being halfway between the other two methods, the sequential approach is as bad as the Fixed PCA and has a much larger variance. The reason for this is that in a fraction of the cases, the 5^{th} component is visible in the Fixed PCA but eliminated by the VB-PCA. The Fusion method keeps this component because the value of this component in decoding the behavioural variable provides the feedback needed to ensure that that component is not eliminated.

5.6 Results on Local Field Potential Data

Two previously-acquired and preprocessed local field potential (LFP) data sets were analysed, from two patients who had recently had stimulus electrodes implanted. The data in each trial consists of a short epoch of LFP recording, time-locked to the event of interest. This is downsampled to only examine the baseband data; higher frequency oscillations were filtered out and are not examined in this analysis. The neuroimaging data $\mathbf{Y}$ is therefore decomposed into a set of temporal patterns (the component timecourses) which describe the Event Related Potentials (ERPs), rather than looking for spatial or spatio-temporal patterns as would be expected in an analysis of FMRI data.

Recordings were taken from Meditronics stimulation electrodes implanted into the dorsal Anterior Cingulate Cortex (dACC) of two patients for treatment of chronic pain. In patient 1, who had bilateral implants, the electrode that was most accurately targeted to dACC was used (left hemisphere); in patient 2, the implant was unilateral (left hemisphere). Signals were amplified (x10000) using isolated CED 1902 amplifiers, bandpass filtered (0.2–1000 Hz), and sampled using a CED 1401 data acquisition unit. Preprocessing involved removal of AC powerline noise from all trials and automatically discarding 5% of the trials automatically identified as containing

large electrical artefacts. To investigate the baseband ERPs, the timecourses were downsampled 80x from 2.95kHz (or 2.5kHz for patient 2) to 36.9Hz (or 31.25Hz) so the 1.5s interval contains 56 (or 47) timepoints. There were two runs per subject, but the first run with patient 1 was discarded because none of the classifiers performed better than chance, possibly because the electrical stimulation of the dACC had been shut off only minutes earlier.

The patients performed a simplified version of the decision-making experiment which has been previously used in MEG (Hunt et al., 2009) and FMRI (Behrens et al., 2007) experiments. Each subject performed two consecutive runs of 240 trials. In each trial, two options were presented on the screen ("blue" and "green"), along with a random reward value for each option; only one of these options would actually yield the stated reward, the other yielding nothing. The subject choses one of these options, and it is then revealed which option actually had the reward. If the subject close the correct option, the corresponding reward value is added to their score; reaching a certain score threshold would increase the subject's monetary payment. The probability of the blue or green option being rewarded varied slowly over time, so the subject was able to learn this probability based on past outcomes. Thus, the subjects make a decision in each round to maximize their expected yield, based on a balance between the learned probability of each option having a reward and the specified potential reward amount of each option. The blue vs. green probability remained stable for a large part of the experiment, and then started to flip back and forth more rapidly; ideally, subjects would learn the probability slowly in the stable parts of the experiment and learn quickly in the volatile parts.

Figure 5.7 shows the behavioural variables that were the targets of decoding. The volatility variable (derived from an "optimal Bayesian learner" performing the same decision-making task) shows how quickly the decision-making task is changing, and therefore how quickly the subject should be learning from the

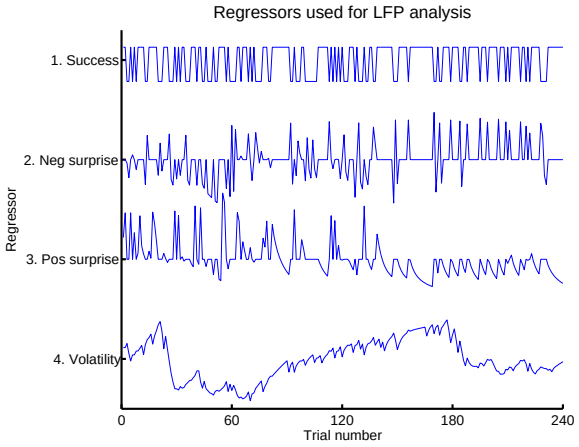


Figure 5.7: Regressors used in the decision-making task. The experiment goes through stable and unstable phases, and the best estimate of this is given by the "Volatility" EV; high volatility indicates that the learning rate should be higher. A result is surprising if it is not the outcome that would be expected based on recent trials. Surprising successes and surprising failures are modelled separately.

responses. In particular, volatility-related activity has been previously been found in the ACC in FMRI experiments (Behrens et al., 2007). The "surprise" variables show how unexpected the reward or non-reward was. Separate regressors are used to represent surprising rewards ("positive surprise") and surprising non-rewards ("negative surprise") because it is generally thought that positive and negative learning are encoded differently in the brain. All variables were demeaned and orthogonalized with respect to the main reward/non-reward variable.

Epochs of the LFP recordings were extracted, time-locked from -500ms to +1000ms from the display of the outcome, with the LFP timecourse being the average voltage over the four electrical contacts (spaced over $\sim$1cm) on one electrode to obtain a monopolar signal.

Figure 5.8 shows the result of analysing the data from subject 1, run 2, with all four behavioural variables. Part (a) shows the PCA components that were selected by each of the approaches. First, notice that most of the component timecourses, as well as the mean timecourse, are quite different between the fixed PCA approach and the VB-PCA techniques; this is because the latter ones inferred different noise

precision λ_r on each trial and these timecourses are reweighted accordingly. There are also some small differences in the component timecourses inferred by the fusion approach as opposed to the sequential approach. The nonzero elements of the weight matrix are shown in part (b), and the resulting decoding vectors are shown in (c). In this example, the ERPs inferred by the two approaches are quite similar.

The strongest signal in these recordings is the binary reward/non-reward response. For the other variables, the cross-validation predictions of the other EVs were actually less accurate than chance (i.e. they would have been more accurate to ignore the neuroimaging data completely and just guess the mean of the training data – i.e. $\langle \hat{\mathbf{v}} \rangle = \left\langle \boldsymbol{\mu}^{(V)} \right\rangle$). Figure 5.8(d) shows the RMS error (rescaled so that guessing the mean = 1) averaged across 5-fold cross-validation. The fact that these other variables are consistently slightly worse than chance, and were never observed to "admit defeat" by eliminating all of the components, suggests that they may be overfitting slightly.

With the reward/non-reward regressor, the results of leave-one-out cross-validation are shown in figure 5.9(a). Using a hard threshold at 0 these are shown as a classification accuracy in part (b). Note that chance is greater than 50% because the classes are not balanced in this experiment.

Figure 5.9(c) shows the decoding vectors produced by each of the methods on the data set for patient 1, run 2. All three approaches have the same basic shape but clearly the fixed PCA's decoding vector is much noisier. This is partially because the fixed PCA approach does not benefit from the denoising introduced by having different λ_r weights on each repeat, and partially because it has many more components to pick from and therefore more opportunities to overfit to noise. In this model the top 30 components of the PCA were kept; the sequential and fusion methods typically found 5–7 significant components from the LFP data.

Overall, the single-behavioural-variable regression results appear to be somewhat better for the fusion approach than the sequential approach, but there are not enough

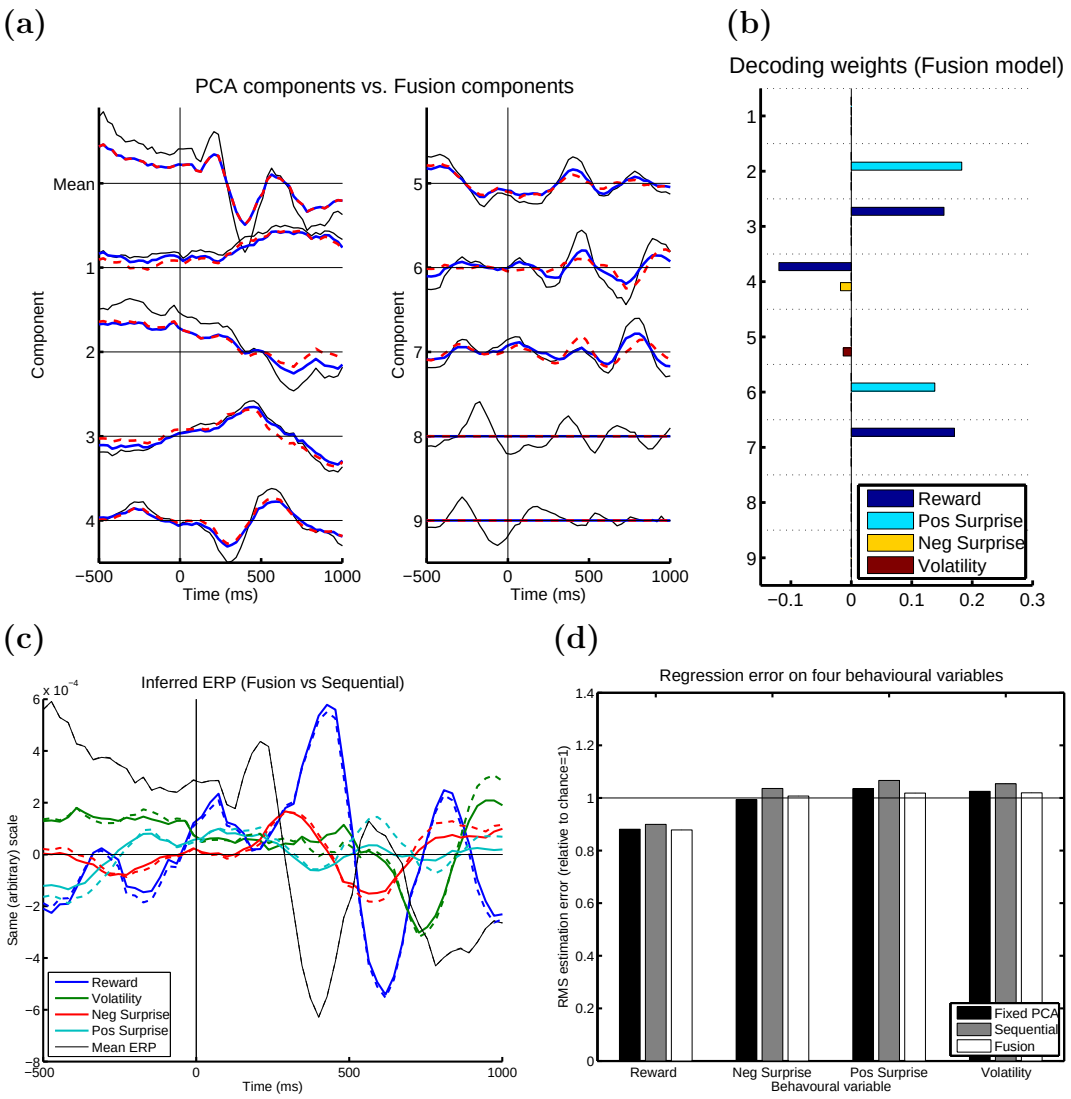


Figure 5.8: Decoding multiple regressors simultaneously in a real data set. (a) The fixed PCA components (thin black line) and VB-PCA components with per-trial noise weighting inferred sequentially (blue solid line) and using the simultaneous approach (red dashed line). (b) The weights of each of these components in decoding each regressor is given in the bar chart. (c) The decoding vectors $\mathbf{D}$ plotted using the fusion approach (solid lines) and the sequential approach (dashed lines). (d) Unfortunately, only the main reward/non-reward regressor was able to predict the behavioural data $v_{b,r}$ better than chance (i.e. better than guessing the mean of $\mathbf{v}_r$) in cross-validation tests.

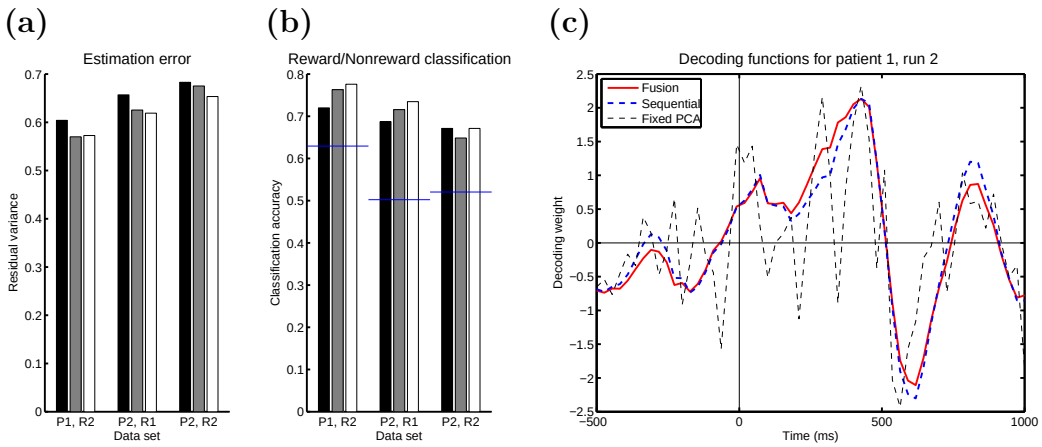


Figure 5.9: Analysis of the binary reward/non-reward regressor only. (a) The residuals when the target was estimated by each of the models: fixed PCA decomposition (black), VB-PCA preprocessing (grey), and the fused model (white). For comparison, simply guessing the mean would leave residual variances of 0.87, 1.0 and 1.0 for these three runs. (b) The classification performance of same three methods. The blue lines indicates chance (guessing the mean). (c) The decoding functions produced on one data set by each of the methods: fixed PCA (black dotted line), VB-PCA preprocessing (dot-dashed blue line) and the fusion model (red solid line).

independent runs to assess if this difference is significant.

5.7 Discussion

This chapter developed a novel Bayesian data fusion model that combined a generative PCA model of the data with a simple ARD-based decoding model to predict the values of a corresponding target vector. In simulations, it was clearly demonstrated that there can be benefits in allowing probabilistic information to flow between the two parts rather than inferring on each part sequentially. This two-way flow of information refines the shape of the components and can thereby improve the accuracy of the decoding. It can also lead to weak components being kept that would otherwise have been eliminated by the VB-PCA, but are preserved and refined because of their value in decoding the behavioural variables. The Fixed PCA keeps all of the original components, but this provides opportunities for overfitting.

Furthermore, the fusion method was shown to combine components in such a way that it makes the decoding part sparser. This means that the information relevant to decoding tends to be stored in fewer components and results in an improvement in prediction accuracy.

When applied to a real data set, it was demonstrated that using a Bayesian PCA to decompose the data provides the flexibility to implement noise models that downweight noisy trials. This was important for analysing the LFP data, however it also required a strong prior to avoid falling into a degenerate solution. A more principled approach would be to use a hierarchical prior to regularize $\boldsymbol{\lambda}$ and avoid having runaway noise precision in a few components.

Using this noise model allowed the component timecourses to be refined based on this noise reweighting and eliminated components that are too noisy to provide useful decoding input. This appears to improve decoding accuracy and using the fusion approach seemed to provide some further improvements compared to the sequential approach.

A more thorough validation of the technique is required to assess if this improvement is significant, and these approaches should also be compared to other methods such as SVMs or nonlinear RVMs that do not rely on decomposing the data into component timecourses first. It would also be valuable to compare the inferred ERPs to those in the literature; however, the estimated ERPs were quite different between the two subjects but quite similar between the two runs on patient 2 (results not shown), suggesting that the observed signal may be highly dependent on the precise location of the recording electrode. It is difficult to assess performance on this limited data set and future research should probably start by adapting this method to work on FMRI spatial maps or extracted EEG/MEG timecourses.

For the LFP data set, it would also be useful to replace the regression model in equation 5.7 with a binary classification model. This could simply be the VB logistic

regression model (Bishop, 2006, Ch. 10.6), which has a rigorous lower bound on the free energy so convergence is guaranteed and model comparison is valid. It would be possible to combine this with other continuous or binary behavioural variables to perform simultaneous inference, by adding each classification as an additional likelihood term (in addition to equations 5.6-5.7), although this may lose some of the benefits of using a correlated noise model across multiple behavioural variables.

The fusion approach bears some strong conceptual similarities to multivariate linear models (Worsley et al., 1997), with the main differences being that the dimensionality is determined automatically for each modality and the inferred posterior is probabilistic. There is also much more flexibility in the type of generative models that can be used, for example allowing separate noise variances on each repeat or by using a classification model.

Increasingly, PCA is giving way to more advanced decompositions like independent components analysis (ICA) in many areas of neuroimaging. In the next chapter, the core VB-PCA equations are replaced by a VB-ICA model using the variational Bayes approach described in Lawrence and Bishop (2000). Furthermore, extensions to multi-dimensional data are explored, using a probabilistic tensor ICA (Beckmann and Smith, 2005) model based on Bayesian parallel factor analysis (Nielsen, 2004). This will be used to explore multimodal MRI data sets across groups of subjects, but could also perform exploratory analysis on spatio-temporal data sets such as MEG, EEG and FMRI.

This framework also has applications outside of supervised learning. Rather than treating this as a decoding method for predicting unknown target vectors, it can also be considered as a fusion method for combining unlike data types into a single unsupervised decomposition; for example, instead of behavioural measures to be predicted they could be additional observations from that trial, such as a reported pain score in a pain experiment. If instead of trials the R dimension represents

different subjects, then the data could be MRI spatial maps and V could be genetic markers or results of blood tests. This would be most powerful if combined with the multi-modal ICA approach discussed in the next chapter.

Chapter 6

Multi-modal data fusion using Linked Bayesian Independent Component Analysis

6.1 Introduction

One of the greatest strengths of MR neuroimaging is its flexibility; by using different pulse sequences in a single scanning session, one can acquire information about the subject's tissue volume and morphology (using high-resolution structural scans), functional activity (using BOLD), white matter integrity (using diffusion-weighted imaging), perfusion (using ASL), to name a few. The result of this is that many recent studies have acquired these multimodal MRI data sets for each subject and analysed them separately to find changes in different aspects of the brain. For example, a recent study by Scholz et al. (2009) showed that learning to juggle induces changes in both grey matter density and in the fractional anisotropy of related white-matter tracts using standard methods. In other cases, the main effects are primarily single-modality: Filippini et al. (2009) found that young people carrying the APOE-ε4 allele (which is associated with increased risk of Alzheimer's later in life) had significant changes in FMRI resting state networks, while no significant changes were observed in grey matter or white matter structure. A major challenge in MRI analysis is to find systematic approaches for fusing data across multiple modalities and automatically

finding patterns of related changes when they exist.

This chapter develops a model based on Bayesian Independent Component Analysis (Bayesian ICA) to extract linked, spatially-independent components from multi-modal data. ICA is a particularly effective framework for finding meaningful components in an unsupervised setting because it searches for non-Gaussian spatial sources that are likely to represent real structured features in the data, while mixing processes and unstructured noise tend to make the observed signal more Gaussian.

The linked PCA decomposition and sparse decoding model presented in the previous chapter involved two Bayesian models linked together probabilistically through a shared per-trial feature vector $\mathbf{h}_r$. This provided a bridge between two different modules, transferring information in both directions. In fact, there is no reason to limit this model to only two parts; the latent space can be thought of as a shared "bus" that any number of self-contained Bayesian modules can be plugged into. This could be used to adaptively weight information from several modalities to jointly inform a single decoding stage. The decoding stage itself can also be removed to convert this into an unsupervised learning model.

The model developed in the previous chapter can be adapted to work with multi-modal group data, but first some of the dimensions take on different meanings. The "repeats" ($r \in 1 \ldots R$) dimension must be the dimension that all of the modalities share; for example, in the case of multi-subject analysis, all of the data sets will share the same subject-course so R is the number of subjects. N represents the number of voxels, and an additional dimension ($t \in 1 \ldots T$) will be used later to stack several modalities into a tensor model.

This novel "Linked ICA" method will be applied to a data set with four different modalities, acquired from 93 subjects (probable-Alzheimer's patients and age-matched controls). One of these modalities is a grey matter partial volume map ("GM") derived from Voxel-Based Morphometry (VBM) methods (Ashburner and

Friston, 2000), and the other three are measures of white matter integrity: Fractional Anisotropy (FA), Mean Diffusivity (MD), and an orthogonal Tensor Mode (MO) described in Ennis and Kindlmann (2006). These last three modalities have been projected onto a two-dimensional white matter surface (the "skeleton") using a Tract-Based Spatial Statistics (TBSS) analysis (Smith et al., 2006).

6.2 Challenges in Decomposing Multi-modal Data

The standard ICA decomposition treats the input data as a rectangular block, typically voxels×timepoints or voxels×subjects. Multi-modal data does not naturally fit into this form and there are a number of different configurations one could consider for doing combined ICA on multi-modal data:

- **Separate ICA analysis** of each modality reveals the salient features for each modality. Since some of these features are caused by distributed neurological variations they would be expected to be visible (to varying degrees) in all modalities, with similar subject-courses.

  Corresponding components can then be matched up using heuristics; but there is no guarantee that components with strongly-correlated timecourses will be extracted, for example if a single component in one modality is explained as a mixture of components in another. When potential matches are found, it can be difficult to determine if they are simply noisy estimates of the same subject-course or whether the underlying subject-courses are different but correlated. A slightly more sophisticated approach to this is the Parallel ICA method described by Liu et al. (2008), which simultaneously decomposes two modalities: subjects' FMRI maps and their genetic profiling data (consisting of 367 single nucleotide polymorphisms). They perform two separate ICA decompositions in parallel, and when matching components (over 30% correlated) are found, an

additional term is added to the overall cost function to encourage this correlation in further iterations.

- **Temporal concatenation** is an approach used for finding resting state networks across multiple subjects. In this case, coregistered FMRI from multiple subjects will not have corresponding time-courses (due to unknown phase of the resting-state oscillations) but one might expect them to have similar spatial patterns. Concatenating the FMRI data in the time dimension makes it possible to find these corresponding patterns by treating the whole multi-subject data set as if it were from one long scan (Beckmann et al., 2005). The equivalent for multi-modal data would be concatenation in the subject direction; however this is not a sensible configuration to use because it would discard the subject correspondence between modalities and treats images from different modalities as being equivalent.

- **Spatial concatenation** is the analogous approach that has been used for analysing multi-modal data, combining all of the data from a single subject into a dataset with more voxels. This "joint ICA" method has been used before for simultaneously analysing functional maps and gray matter maps (Calhoun et al., 2006), and has been used to extract correlations in structural grey-matter/white-matter density data (Xu et al., 2009). Since concatenation is a preprocessing step, the ICA model is completely unaware of which voxels belong to which modality.

  The contribution each source makes to the ICA depends greatly on the scaling of the data from each modality. One of the difficulties of concatenating multi-modal data is that the modalities may have different noise levels and different numbers of voxels. If the scaling is mismatched, unsupervised methods such as PCA and ICA will be dominated by the largest-variance modalities, or those

with the most voxels. It is important to ensure that the scaling to be matched is the noise variance, not signal variance, so robust preprocessing methods must be used to avoid this scaling problem.

There is also an issue of noise co-variance, due to spatial smoothing for example; in particular, adding more smoothing to one modality reduces the noise level but leaves the number of voxels unchanged. Therefore some sort of weighting relative to smoothness would be beneficial, although the proposed method does not deal with this explicitly in its current implementation.

A further problem is that different modalities may have different activation histograms. ICA effectively assumes that each component has a single, non-Gaussian histogram as the prior distribution for all voxels in its spatial map. If this map consists of voxels from several different modalities, the modelled histogram may have to compromise, for example if one modality has a small area of strong activation while the other has a large region of weak activation. This can cause sub-optimal estimates of intensities in spatial maps and will also interact with the relative scaling problem.

We also expect that some of the components modelled by ICA will be structured noise, and this may be completely invisible in some of the other modalities. It would therefore be useful for sources to be "switched off" in the models where they are not needed, just like it is important to eliminate unneeded components in a probabilistic ICA model (Choudrey and Roberts, 2001).

- **Tensor ICA** stacks the modalities to create a 3D block of data. This has been used for multi-subject FMRI analysis, with dimensions of voxels×times×subjects (Beckmann and Smith, 2005). This is related to the PARAFAC model (see Nielsen 2004 for a VB-based implementation) but with the addition of spatial-independence priors. This method assumes that each

component has a single spatial map for all modalities, applied to each modality with different weightings. This can be a beneficial feature because it avoids unnecessary duplication of the spatial maps, but it may be inappropriate (e.g. if the number of voxels is different, or if the spatial maps in different modalities are completely unrelated).

Using a modular Bayesian framework, this chapter develops a novel "Linked ICA" model that allows for either Tensor ICA or Concatenated ICA, or a combination of both at the same time. The same subject weighting matrix $\mathbf{H}$ is shared between all of the modalities, so each component consists of a single subject-course and one spatial map in each of the modalities. The subject weighting matrix automatically balances information from all of the modalities. This model could also be extended in future to incorporate the decoding stage from chapter 5, by having the EVs of interest (such as group membership or a disability score) to be included as another modality in the decomposition.

6.3 Linked ICA Model for Multi-modal Data Sets

It will be assumed that the data set is from a group of R subjects, each scanned using several different modalities. It should be noted that the proposed method has the potential to be applied to any three-dimensional data set (e.g. $space \times time \times trials$ in functional neuroimaging data). Each of the scans is preprocessed, analysed, and coregistered using whatever methods are recommended for single-modality ICA analysis. This produces maps for each modality, which can have different spatial masks and different numbers of voxels. In this model, "modality" is defined as referring to a single output image (per subject) that refers to a particular output extracted from the data. Typically, different modalities will have different units, different scalings and different noise levels. In some cases, a single analysis may result in several different contrast images; for example, a diffusion tensor imaging (DTI)

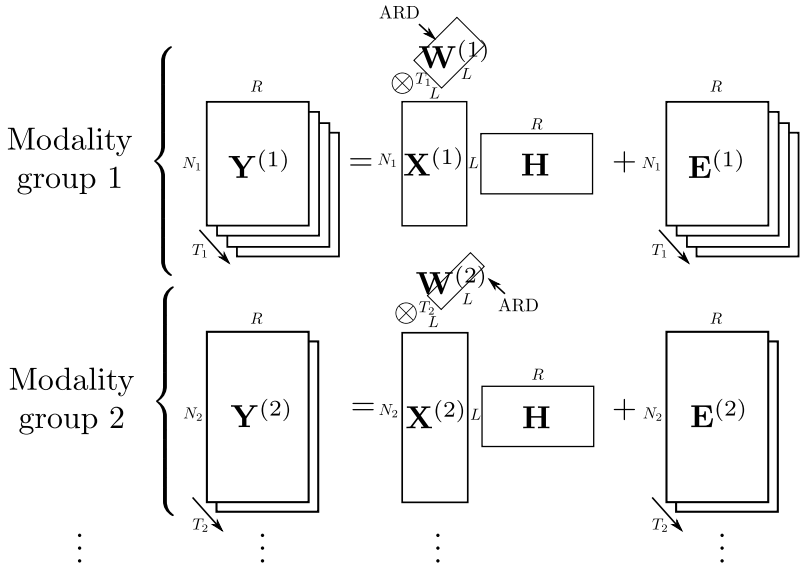


Figure 6.1: The main matrices of the Linked ICA that models multi-modal data $\mathbf{Y}$. Note that the same $\mathbf{H}$ matrix is used for all of the modality groups, but otherwise they are K separate Tensor ICAs, each with separate data dimensions $N_k \times T_k \times R$ (voxels×modalities×subjects). Each of the modality groups contains one or more modalities stacked together, expressed in terms of spatial maps $\mathbf{X}^{(k)}$, modality weights $\mathbf{W}^{(k)}$, a shared subject-weighting matrix $\mathbf{H}$, and additive noise $\mathbf{E}$.

analysis can produce maps of FA (fractional anisotropy), MD (mean diffusivity) and MO (tensor mode). These are treated as separate modalities.

However, to maintain some of the benefits of tensor ICA, similar modalities are collected into K "modality groups". Modalities in the same group must share the same space; this means the modalities must be aligned to each other and have the same spatial mask, and should also have similar spatial properties (for example, the same amount of smoothing). A good example of this are multiple diffusion-derived measures projected onto a white matter skeleton using TBSS. The data can then be packed into a set of 3D arrays $\mathbf{Y}^{(k)} \in \mathscr{R}^{N_k \times T_k \times R}$, where N_k is the number of voxels in the shared spatial map and $T_k \geq 1$ is the number of modalities in the k^{th} modality group. Each modality group is modelled using a Bayesian tensor ICA model. This general configuration is shown in figure 6.1.

6.3.1 Bayesian Tensor ICA Model

Within each modality-group k the data is modelled as a sum of components using a tensor decomposition. Each component $i = 1 \dots L$ is multilinear, which means it can be expressed as the tensor product of one spatial map, one subject-course, and one modality-course (i.e. it is linear in each dimension). These model the data in voxel n, subject r, modality group k and modality t as

$$\mathbf{Y}^{(k)}_{n,t,r} \quad = \quad \sum_{i=1}^{L} \mathbf{X}^{(k)}_{n,i} \mathbf{W}^{(k)}_{t,i} \mathbf{H}_{i,r} + \mathbf{E}^{(k)}_{n,t,r} \tag{6.1}$$

where $\mathbf{X}^{(k)}_{n,i}$ are the spatial maps for component i in modality group k, $\mathbf{W}^{(k)}_{t,i}$ are the modality weightings for component i in modality t (of modality group k), and $\mathbf{H}_{i,r}$ are the weights for component i in subject r. For simplicity this model is used even when $T_k = 1$, so that $\vec{W}^{(k)}_{\cdot,i}$ is just a scalar. Crucially, the same $\mathbf{H}$ matrix is shared between all of the modality-groups; this forms the only probabilistic link between the different modality groups. The i^{th} component has the same subject weightings across modality groups but each group has its own spatial map. Thus the number of repeats R and the maximum number of components L must be the same everywhere, because these dimensions are shared, while N_k and T_k are not. White-noise residuals are assumed, with the modality-dependent noise precision $\boldsymbol{\lambda}^{(k)}_t$:

$$\mathbf{E}^{(k)}_{n,t,r} \quad \sim \quad N(0, 1/\boldsymbol{\lambda}^{(k)}_t). \tag{6.2}$$

The sketch of the Linked ICA matrices is shown in figure 6.1, and figure 6.2 shows how these variables fit into the full Linked ICA graphical model; this also includes the hyperparameters explained in the next two sections. Aside from the shared $\mathbf{H}$, linked ICA operates identically to separate tensor ICA analysis: each modality k has its own separate source mixture model, as well as having its own noise model and separate ARD priors to drive different patterns of sparsity. Note that r indexes the "repeats" dimension and is the dimension that is shared across modality groups, for example r indexes subjects in the multi-subject application.

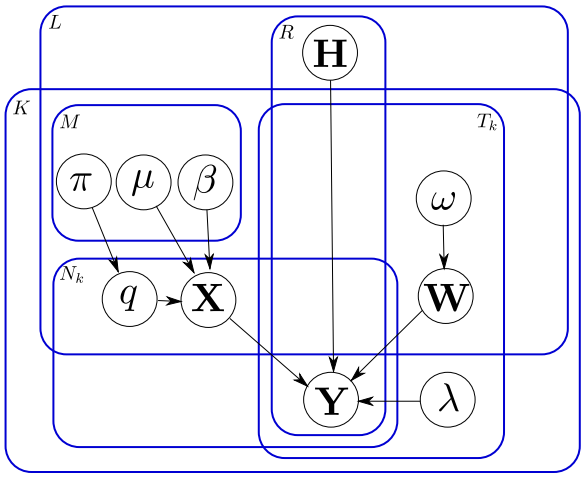


Figure 6.2: The graphical model showing the relationships between all parameters and hyperparameters. Plates represent replicated variables or matrix sizes, with dimension in a top corner; for example, $\mathbf{H}$ is $L \times R$, and $\mathbf{Y}^{(k)}$ consists of K arrays of dimension $N_k \times T_k \times R$. Fixed hyperpriors have been omitted.

6.3.2 Adaptive Modality-weighting

The tensor model (equation 6.1) implies that the same spatial sources $\mathbf{X}^{(k)}_{\cdot,i}$ are used for all of the different maps $t \in 1..T_k$, with weightings given by $\mathbf{W}^{(k)}_{t,i}$. In previous tensor ICA applications (Beckmann and Smith, 2005), this t dimension would correspond to repeats of the same scan, such as in a multi-subject FMRI data from a study with identical stimulus timings. In that case it would make sense to assume the same noise level for all timepoints, i.e. use only a scalar $\boldsymbol{\lambda}^{(k)}$. Instead, the tensor model is being used here to consolidate several different maps produced by the analysis of a single scan, and since in this case each t refers to different underlying modalities of data, it needs to model differences in scaling and noise levels, hence the modality-specific $\boldsymbol{\lambda}^{(k)}_t$ is used.

To adapt to different scalings of the "signal" in each modality, an ARD prior is used on the modality-courses ($\mathbf{W}$). Just like the noise level, the relative scaling of the data in each t needs to be determined independently; thus independent ARDs are

placed on each *element* of $\mathbf{W}$.

$$\mathrm{P}\big(\mathbf{W}^{(k)}\big) \;=\; \prod_{t=1}^{T_k}\prod_{i=1}^{L} N(\mathbf{W}_{t,i}^{(k)}|0,(\boldsymbol{\omega}_{t,i}^{(k)})^{-1}) \tag{6.3}$$

with an approximately scale-free prior on $\boldsymbol{\omega}_i$; as $\boldsymbol{\omega}_i \to \infty$ that will effectively eliminate a source from that timepoint by forcing that element of $\mathbf{W}$ to zero. In this approach, the number of sources is not explicitly chosen, but the method automatically determines the number of sources that are needed to optimally describe the data. Typically, it is best to start with too many sources and allow the model to gradually downweight and eliminate sources that are too weak.

This means that it is now possible to eliminate a source from some modalities while keeping it in others, so it is possible to model effects like single-modality structured noise. This means that the subject-course no longer needs an ARD, so it has a simple fixed prior:

$$\mathrm{P}(\mathbf{H}) \;=\; \prod_{i=1}^{L}\prod_{r=1}^{R} N(\mathbf{H}_{i,r}|0,1). \tag{6.4}$$

This dimensionless prior on $\mathbf{H}$ is a useful feature and is analogous to the fixed unit variance priors used in variational PCA (see chapter 5 or Bishop 1999). When a source has not been eliminated, the ARD priors on $\mathbf{W}$ will tend to balance with this fixed-scale prior to keep the rows of $\mathbf{H}$ close to a (root-mean-squared) amplitude of one. This is useful for linked ICA because it means that each column of $\mathbf{H}$ is a short dimensionless vector summarising everything that varies between different subjects' scans (apart from residual noise). This modality-independent hidden state forms a probabilistic link between separate ICAs.

There are situations in which one modality would itself consist of several distinct timepoints; for example, multi-subject whole FMRI scans with synchronised stimuli (Beckmann and Smith, 2005) or identical structural scans acquired several years apart to image neurodegeneration. These could easily be modelled in this framework by returning to a single $\boldsymbol{\omega}^{(k)}$ and $\boldsymbol{\lambda}^{(k)}$ for all timepoints of that component.

6.3.3 Independent Spatial Sources

The driving force behind an ICA decomposition is that the data is derived from a number of statistically independent spatial sources; these are the spatial maps ($\mathbf{X}^{(k)}_{\cdot,i}$ for $i = 1 \dots L$). By the central limit theorem, linear mixing of independent sources will produce output that is more Gaussian than the sources. FastICA is a blind source separation technique that attempts to reverse this by maximizing explicit measures of non-Gaussianity, such as estimates of kurtosis or negentropy (Hyvärinen and Oja, 2000). Another approach for finding this non-Gaussianity is to explicitly fit a non-Gaussian distribution to each source by assuming a particular functional form. This is the approach taken here, using an M-component Gaussian mixture model as proposed for independent factor analysis by Attias (1998).

This models the weightings in each source ($\mathbf{X}^{(k)}_{\cdot,i}$) as being drawn from an M-component Gaussian mixture model with means $\boldsymbol{\mu}_{i,m}$, precisions $\boldsymbol{\beta}_{i,m}$, and component proportions $\boldsymbol{\pi}_{i,m}$. This is a good approximate model for variety of underlying distributions (for examples, see Choudrey and Roberts 2001). This mixture model prior on the spatial maps can be expressed as

$$\mathrm{P}\Big(\mathbf{X}^{(k)}_{n,i}\Big|\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\pi}\Big) \;=\; \sum_{m=1}^{M} \boldsymbol{\pi}^{(k)}_{i,m}\; \mathrm{N}\Big(\mathbf{X}^{(k)}_{n,i}\Big|\boldsymbol{\mu}^{(k)}_{i,m}, 1/\boldsymbol{\beta}^{(k)}_{i,m}\Big)\,. \tag{6.5}$$

It is also be useful to explicitly model the hidden mixture labels using the categorical variable $q^{(k,i)}_n = m$, which specifies that voxel n (in modality group (k,t) and in ICA component i) is drawn from the m^{th} mixture component. The notation $\mathbf{q}^{(k,i)}_{\cdot,n}$ is also used as the $M{\times}1$ vector of all zeros except for a single one in row $q^{(k,i)}_n$. This is convenient because it means that $\left\langle \mathbf{q}^{(k,i)}_{\cdot,n} \right\rangle$ gives the posterior probability of a voxel being drawn from each mixture component. By conditioning on $q^{(k,i)}_n$ this prior can be rewritten as

$$\Big(\mathbf{X}^{(k)}_{n,i}\Big|q^{(k,i)}_n\Big) \sim \mathrm{N}\Big(\boldsymbol{\mu}^{(k)}_{i,q^{(k,i)}_n}, 1/\boldsymbol{\beta}^{(k)}_{i,q^{(k,i)}_n}\Big)\,, \qquad \mathbf{q}^{(k,i)}_{\cdot,n} \sim \mathrm{Cat}\big(\boldsymbol{\pi}^{(k,i)}\big) \tag{6.6}$$

where the categorical distribution $\mathrm{Cat}(\boldsymbol{\pi})$ is defined by the probability mass distribution

$$\mathrm{Cat}(\mathbf{q}|\boldsymbol{\pi}) \quad = \quad \mathbf{q}^{\mathrm{T}}\boldsymbol{\pi} \tag{6.7}$$

assuming that $\mathbf{q}$ has a single element equal to 1 and the rest of the elements equal 0. In other words, $\mathrm{P}(q = m|\boldsymbol{\pi}) = \boldsymbol{\pi}_m$.

An alternative choice for this prior would be the Gamma-Gaussian mixture model used to model FMRI signals in Probabilistic ICA (Beckmann and Smith, 2004). PICA's 3-component model uses a central Gaussian distribution to model noise, and has two gamma distributions (one positive and one negative) to model activation. There are also hard restrictions on the shape and location of these gammas to ensure they only model the tails. This is a more accurate and robust mixture model for FMRI analysis (Beckmann, 2004), partially because gamma distribution has heavier tails than a Gaussian (decaying as e^{-x} rather than e^{-x^2}). Approximations have already been developed to use this mixture model in the VB-ICA framework (Makni et al., 2006).

For simplicity, the model presented in this chapter uses a plain Gaussian mixture model and fixed $M = 3$ mixture components. In practice, using 2 or 3 mixture components seems to extract the non-Gaussian sources from noisy simulated data reasonably well.

6.4 Variational Bayesian Inference

The model is inferred using Variational Bayes (VB). The full posterior distribution is intractable, so the mean field approximation is used and the posterior distribution is factorized as

$$\begin{aligned} &P(\mathbf{Y}|\mathbf{H}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{X}, \mathbf{q}) \quad \approx \\ &\qquad P'(\mathbf{H})P'(\boldsymbol{\beta})P'(\boldsymbol{\mu})P'(\boldsymbol{\pi})P'(\mathbf{W})P'(\omega)P'(\boldsymbol{\lambda}) \prod_{k=1}^{K}\prod_{i=1}^{L} P'(\mathbf{X}^{(k,i)}, \mathbf{q}^{(k,i)}) \end{aligned} \tag{6.8}$$

Notice that this explicitly factorizes over components i. The solution would still be analytic without this factorization, but the number of components in the joint mixture model grows as M^L (see Attias 1998). Factoring over sources avoids this exponential growth at the cost of ignoring any correlation in the posterior distributions of the spatial maps. Note that although the sources are independent (and therefore should have no correlation), the posterior uncertainty in these spatial maps could easily be correlated (for example, if two components have highly-correlated subject-courses in $\mathbf{H}$).

For all details of these updates, see appendix D. The Gaussian mixture model prior $P'(\mathbf{X}^{(\mathbf{k},\mathbf{i})}, \mathbf{q}^{(\mathbf{k},\mathbf{i})})$ naturally factors itself into two parts: the Gaussian conditional distributions of $P'(\mathbf{X}^{(\mathbf{k})}_{n,i} | \mathbf{q}^{(\mathbf{k})}_{n,i} = m)$ and the label probabilities $\mathrm{P}'\big(\mathbf{q}^{(\mathbf{k},\mathbf{i})}\big)$. However, these remain tightly connected; in particular, the update $P'(\mathbf{q})$ depends on $\mathbf{Y}$, despite $\mathbf{q}$ and $\mathbf{Y}$ not being directly connected in the graphical model. These are therefore treated as a single monolithic update during VB inference; for example, it is possible that updating $P'(\mathbf{X}^{(\mathbf{k})}_{n,i} | \mathbf{q}^{(\mathbf{k})}_{n,i} = m)$ can cause the free energy bound F to *decrease* until the $\mathrm{P}'\big(\mathbf{q}^{(\mathbf{k},\mathbf{i})}\big)$ update is also done.

The free energy F is also calculated in appendix D.2. This was used to validate the VB updates (by ensuring that $\Delta F \geq 0$ after each update) and also to monitor convergence. The analysis software is implemented in MATLAB.

6.4.1 Precision Contributions

When a component represents some real underlying variation between subjects, it can fuse information across several modalities but one would still expect these variations show up more clearly in some modalities than in others. This subsection describes a simple measure for assessing the relative influence of each different modality in defining the subject-course of each component.

The latent space $\mathbf{H}$ is shared between all modalities, so its posterior $\mathrm{P}'(\mathbf{H})$

combines contributions from all of the modalities (as well as the prior). The VB update for P′(**H**) is given by equations D.35 and D.36, duplicated here:

$$\left(\boldsymbol{\Sigma}_{\mathbf{H}}^{-1}\right)_{ij} = (\mathbf{I}_L)_{ij} + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\left\langle \mathbf{X}_{ni}^{(k)}\mathbf{X}_{nj}^{(k)}\right\rangle \left\langle \mathbf{W}_{ti}^{(k)}\mathbf{W}_{tj}^{(k)}\right\rangle \boldsymbol{\lambda}_t^{(k)} \tag{6.9}$$

$$\left(\mathbf{M}_{\mathbf{H}}\boldsymbol{\Sigma}_{\mathbf{H}}^{-1}\right)_{ir} = 0 + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\left\langle \boldsymbol{\lambda}_t^{(k)}\right\rangle \left\langle \mathbf{W}_{ti}^{(k)}\right\rangle \sum_{n=1}^{N_k}\mathbf{Y}_{n,r,t}^{(k)}\left\langle \mathbf{X}_{n,i}^{(k)}\right\rangle \tag{6.10}$$

where $\langle\cdot\rangle$ denotes the expectation over the factorized posterior distribution. Ignoring the off-diagonal elements of $\boldsymbol{\Sigma}_{\mathbf{H}}^{-1}$ (which should be small because the spatial maps are independent), this formulation means that each modality has its own ideal (likelihood-maximizing) estimate of **H** and the posterior $\langle\mathbf{H}\rangle$ is a precision-weighted average of these. Conveniently, this precision is the same for each subject r. To find the dominant modalities in estimating each source's subject-course, it is informative to look at these precisions. The "precision contribution" by modality t in modality group k to each source i is defined by looking at the parts of the sum in equation 6.9:

$$PC(k,t,i) = \sum_{n=1}^{N_k}\left\langle \left(\mathbf{X}_{n,i}^{(k)}\right)^2\right\rangle \left\langle \left(\mathbf{W}_{t,i}^{(k)}\right)^2\right\rangle \left\langle \boldsymbol{\lambda}_i^{(k)}\right\rangle \tag{6.11}$$

So overall the precision of $P'(\mathbf{H}_{i,\cdot})$ is given by $1+\sum_{k=1}^{K}\sum_{t=1}^{T_k} PC(k,t,i)$, because the prior makes a constant precision contribution of 1. This provides a fixed scale against which to measure these contributions; if $PC(k,t,i) \ll 1$ then that modality is considered to have been eliminated from that component.

In the results section the figures will show these PCs normalized by the overall precision, so that the sum of all contributions is 1. This makes it easy to see if a component is dominated by one modality or is informed by a combination of several modalities.

6.4.2 Preprocessing

Each modality's data is de-meaned so that every input voxel has zero mean across repeats. This mean could also be inferred within the Bayesian model by adding

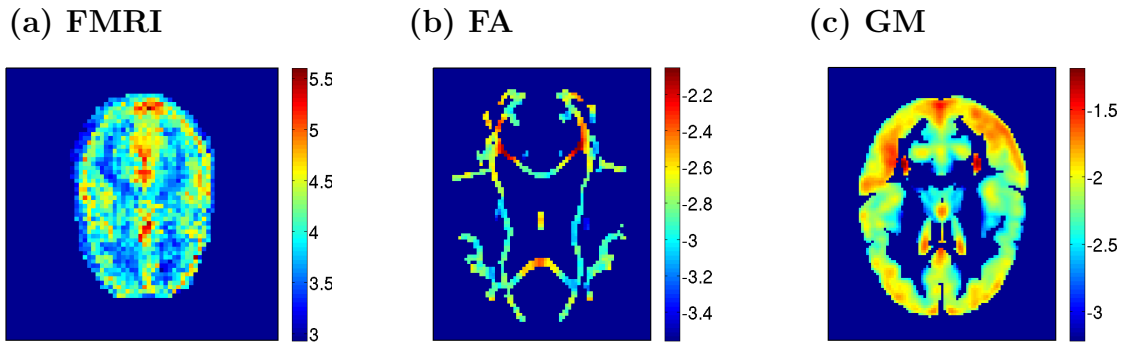


Figure 6.3: Voxelwise $\log_{10}$ standard deviation maps for (a) FMRI, (b) Fractional Anisotropy and (c) Grey Matter density maps. In all of these modalities the noise level can vary by 1.5 to 2.5 orders of magnitude.

appropriate mean variables, but for simplicity this is kept as a preprocessing step. (Note that the trick used in the previous chapter, forcing $\mathbf{H}_{\cdot,1} = \mathbf{1}$, does not work because it is not always reasonable to assume that the mean image has a tensor form.)

A more serious issue is that the variance can vary enormously from voxel to voxel. This is especially true in FMRI data (where there can be a two order-of-magnitude difference in noise variance between CSF and white matter voxels), but it is the case for other modalities as well. For example, the GM image is always between 0 and 1, but some voxels are much more variable between subjects than others. Figure 6.3 shows these log-scaled variance patterns for several modalities. A reliable and robust method is required to whiten this. It would be possible to model different noise levels for each voxel, but this would introduce a large number of new parameters. Instead of implementing this, a well-established empirical method directly from PICA was used (Beckmann, 2004). This attempts to estimate the per-voxel scaling of the underlying white noise by looking only at the centre of the intensity histograms and ignoring the tails, using the following procedure:

1. PCA on the data to get the eigenvectors (spatial maps).

2. In each of these eigenvectors, zero out any voxels beyond 1.6 standard deviations of zero. This is intended to remove most of the high-kurtosis "signal". In the

null model (where the component is Gaussian) this trims all components equally.

3. Reconstruct the image using these truncated spatial maps. This is the first-pass estimate of noise.

4. Calculate the standard deviation of this noise in each voxel, then

5. Use these rescalings to rescale the voxels in the original data set.

This can be iterated as necessary (using the rescaled data to get a better-balanced PCA), but in practice one iteration is normally enough to get a robust estimate. In the null case (where all noise is Gaussian) the thresholding in step 2 will remove an equal amount from each eigenvector, leaving the relative voxel scaling unchanged.

Another problem with multi-subject data is that noise levels will intrinsically vary between subjects, due to different amounts of physical movement during the scan or physiological effects. This could be modelled explicitly by estimating a separate noise precision for each subject in each modality, but this might require strong priors to avoid degenerate solutions (as noted previously in 5.2.2). Instead, the current implementation normalizes across subjects using the same procedure as described above for normalizing between voxels, with the last two steps alternating between normalizing voxels and normalizing the subject-course. Importantly, the same subject-course normalization must be performed across all modalities, because rescaling the data differently for different modalities would mean that the same true subject-course would look different under each modality's rescalings. In future work it would be preferable to incorporate separate per-subject, per-modality noise levels into the Bayesian model; this would avoid this problem by keeping the noise scale separate from data scale. Per-subject noise modelling should be computationally feasible, although it would likely make sense to keep the per-voxel normalization as a preprocessing rather than part of the model step due to the number of parameters involved and the very large variations in noise amplitude.

6.4.3 Initialization

As with most ICA implementations, Linked ICA is initialized from a PCA decomposition. For multi-modal data, the natural method is to concatenate all of the voxels across modalities k, t (to get a $(\sum_{k=1}^{K} N_k T_k) \times R$ matrix) and then do a PCA decomposition on that. To avoid being overly biased by data scaling, the PCA is done after the data has been normalized by the voxel-wise and subject-wise variance normalization step. Taking only the first L components of the decomposition, the loading matrix provides an initial point estimate for the shared subject-course matrix $\mathbf{H}$ and eigenvectors are used to initialize the spatial maps $\mathbf{X}$ as described below.

The initialization of $\mathbf{X}^{(k)}$ depends on whether modality group k is tensor ($T_k > 1$) or non-tensor ($T_k = 1$). If modality group k is non-tensor then this PCA also yields initial point estimates of $\mathbf{X}^{(k)}$ and the weights $\mathbf{W}^{(k)}$ are initialized to 1. For tensor modality groups, the PCA decomposition yields a $N_k T_k \times 1$ vector. Reversing the original concatenation step reshapes this into one $N_k \times T_k$ matrix per component, but in general this will yield a different spatial map for each modality, while the tensor model demands that each component's spatial map be identical for all modalities (aside from scaling). This $N_k \times T_k$ matrix is approximated by the bilinear $\mathbf{X}^{(k)}_{\cdot,i} \mathbf{W}^{(k)\mathrm{T}}_{\cdot,i}$, using another PCA to find the best approximation. This provides a starting point which the Bayesian method can improve upon by looking for independent components.

Once the $\mathbf{X}^{(k)}$ matrices are found, each component's mixture model is initialized by locating means $\boldsymbol{\mu}^{(k)}_{\cdot,i}$ at the 25th, 50th, and 75th percentiles of the spatial map intensities, and setting the standard deviation $\boldsymbol{\beta}^{(k)^{-\frac{1}{2}}}_{\cdot,i}$ equal to half of the spacing between the means.

6.5 Baseline Approaches

Of the data configurations discussed in the introduction, spatially-concatenated ICA represents the only reasonable way to arrange all of the available multi-modal data

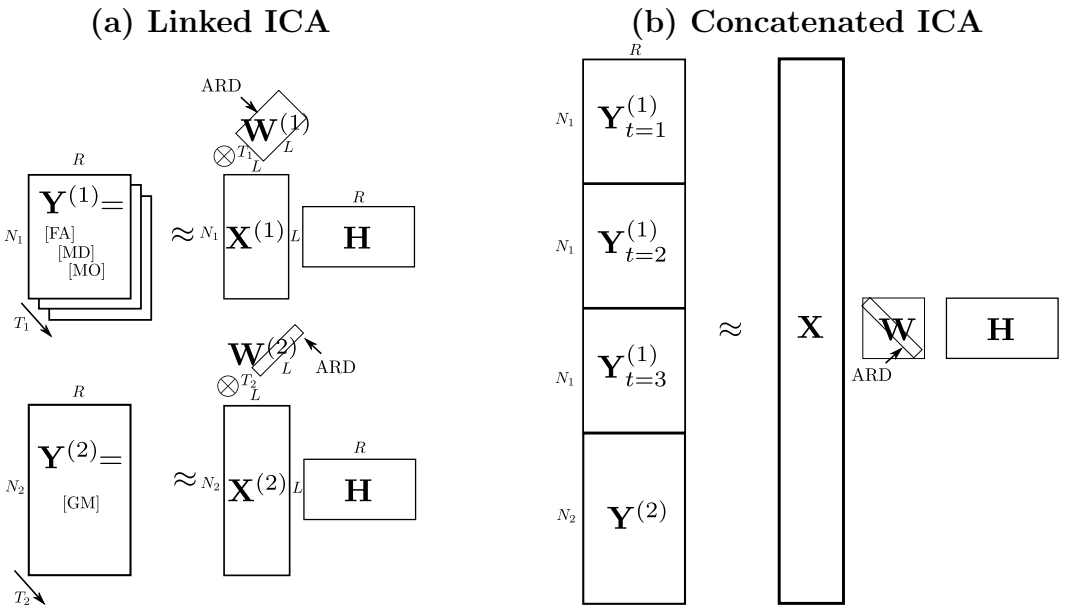


Figure 6.4: Matrix diagrams showing (a) Linked ICA and (b) Concatenated ICA applied to a four-modality data set. Modality group 1 has three modalities, all on the white matter skeleton: FA, MD, and MO. The modality group 2 has only a single modality, the grey matter density (GM) image. In the Linked ICA model the first three modalities share the same spatial maps $\mathbf{X}^{(1)}_{\cdot,i}$, and separate histograms are inferred for the GM spatial maps $\mathbf{X}^{(2)}$. In Concatenated ICA, the all of the voxels are concatenated into a single large spatial map $\mathbf{X}$ with a single histogram. Also, $\mathbf{W}^{(k)}$ makes it possible to switch off individual modalities from a component, while Concatenated ICA can only eliminate components completely.

into a single ICA decomposition. This enables all of the data to be used in inferring subject-courses $\mathbf{H}$, but discards the information about which modality each voxel comes from, and also loses the spatial correspondence between the modalities in any tensor modality groups.

All voxels are concatenated across modalities and a single, large spatial map containing all of the modalities is used, with the same noise-normalizing scaling as used for the initial PCA. (For simulated data, no voxelwise normalization is required; instead, the modalities are rescaled optimally using the true noise amplitude.) This Concatenated ICA configuration can also be inferred using VB and is the baseline model against which Linked ICA will be evaluated.

In this situation, all of the spatial maps are concatenated voxelwise and the same

mixture model hyperparameters are used for the entire concatenated spatial map. Furthermore the tensor model is flattened out so that instead of having t-related weights, the link between corresponding voxels is lost and the spatial map is just replicated once for each modality, as illustrated in figure 6.4. The $\mathbf{W}$ matrix is still present but it can only scale (and eliminate) each component from the entire model. This makes the model almost equivalent to the Bayesian ICA model of Choudrey and Roberts (2001), aside from the fact that the ARD is on the scaling matrix $diag(\mathbf{W})$ rather than on the rows of $\mathbf{H}$.

For comparisons, the precision contributions from individual modalities can also be calculated for the Concatenated ICA results by using only the relevant voxels in the spatial maps:

$$PC(\hat{k}, \hat{t}, i) = \sum_{n \in \mathbf{N}^{(k,t)}} \left\langle \mathbf{X}_{n,i}^2 \right\rangle \left\langle \mathbf{W}_i^2 \right\rangle \langle \lambda \rangle \tag{6.12}$$

where $\mathbf{N}^{(\hat{k},\hat{t})}$ is the set of voxels relating to modality $k = \hat{k}, t = \hat{t}$ in the Linked ICA.

In some cases, comparisons are also made to the Probabilistic ICA (PICA) approach (using the *MELODIC* software) that is the main exploratory analysis tool in the FMRIB Software Library (FSL, `http://www.fmrib.ox.ac.uk/fsl`). PICA differs from the Bayesian ICA in several key respects; one of the main ones is that FastICA is a noise-free model, where the number of sources equals the data dimensionality. Dimensionality reduction is achieved by a temporal PCA preprocessing step, using a dimensionality estimate derived from the eigenspectrum. It also uses a direct measure of independence (estimated negentropy) rather than modelling a non-Gaussian histogram directly.

6.6 Simulated Multi-modal Data

A simulated multi-modal data set was constructed with four modalities in two modality groups. The first group contains three modalities of 1000 voxels each that

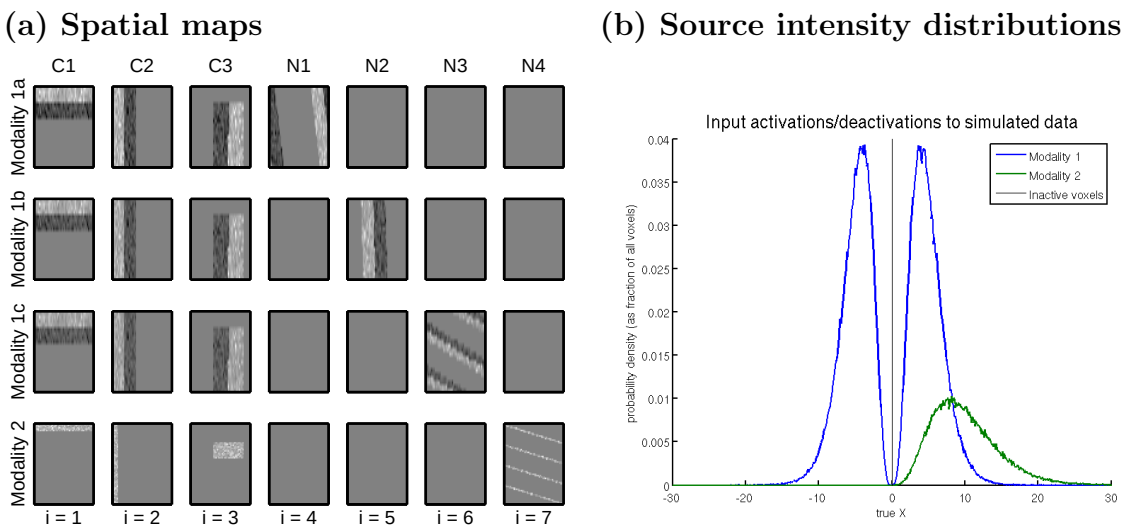


Figure 6.5: The simulated multi-modal data. (a) There are seven components: three shared sources (C1–C3) that appear in all four modalities, and four structured noise sources (N1–N4) each appearing in a single modality. The first three modalities (1a–1c) have 1000 voxels each and are in the same modality group and therefore share the same spatial maps; in this example all of the weights are the same (set to 1) but they have different noise levels. The last modality (labelled modality 2) has 3000 voxels and different spatial maps. (b) The histograms of the two modality groups are very different: group $k = 1$ has 40% of its voxels active (half positive and half negative), while group $k = 2$ has only 10% active (all positive) but the activation is stronger. Note that the remaining voxels in both histograms are completely inactive so they are shown by the peak at exactly zero.

share the same spatial patterns with different weightings, and the second group which has a single, 3000-voxel modality. The spatial maps are shown in figure 6.5. These signals are buried in noise; the images shown are scaled such that full scale represents the range $[-\frac{1}{2}\lambda^{-1/2}, +\frac{1}{2}\lambda^{-1/2}]$, i.e. $\pm\frac{1}{2}$ the standard deviation of the added white noise.

There were three components (labelled C1–C3) that were expressed in each of the modalities, and a structured noise source that was unique to each modality (N1–N4). The subject-courses (not shown, $R = 100$) are white noise, but were generated with somewhat correlated subject-courses (30%) in order to make crosstalk more likely. This means that the initial PCA will mix the signal components together, so the ICA method must move away from its initialization point and seek non-Gaussian

histograms in order to accurately separate components with non-orthogonal subject-courses.

Figure 6.5(b) shows the true activation levels used in the simulation. The spatial maps for modality group 1 have a low level of activation with 40% of voxels active (20% positive, 20% negative), while the maps for modality group 2 have only 10% active voxels (all positive) with a higher intensity. These were chosen so that the two modality groups had very different histograms. Gamma distributions were used because the heavy tails are thought to more accurately reflect the properties of real activation than a Gaussian. The activation distributions shown only account for part of the intensity histogram; the remaining 60%–90% of voxels are inactive and therefore collect in a large point mass at exactly zero. In the high-noise simulation, the white noise added to the four modalities had standard deviations of 15, 20, 25, and 50 respectively. In the low-noise simulation, the same signal was used but with noise scales of 15, 20, 25, and 40.

The number of components was set to $L = 10$ so that with a true dimensionality of 7 there was work for the dimensionality estimation (via the ARD prior on $\mathbf{W}$) to do. The inference ran until it converged ($\Delta F < 0.1$), which took 300–450 iterations. This MATLAB code took 1-2 minutes to run each inference on a single core of a 2.4 GHz Intel Core 2 processor.

6.7 Results on Simulated Multi-modal Data

The precision contributions are shown in figure 6.6. In the Linked ICA model at both noise levels, the three shared sources (1–3) each split their precision fairly equally between the four modalities. The structured noise sources (4–7) are mostly determined by a single modality, and the others (8–10) are eliminated completely (dominated by the prior precision). In the concatenated model at high noise, only

5 sources are inferred and the rest are eliminated. The last two components (4–5) model some combination of the four structured noise sources.

This occurs in high noise (but not low noise) because these structured noise components are only slightly detectable above the high noise level, so the Concatenated model determined that switching off those additional components (by inferring $\mathbf{W}_i = 0$) provides a more concise explanation of the data. In the Linked ICA model it is possible to switch off each modality's contribution to component i separately (by inferring $\mathbf{W}_{ti}^{(k)} = 0$ for only some modalities k, t), so the complexity penalty for keeping each component is reduced.

In the low noise simulation all of the structured-noise components (4–7) are extracted well by both methods, but only Linked ICA correctly infers clean separation of the structured noise by modality. Figure 6.7 shows that the subject-courses are all extracted more accurately by Linked ICA.

Figure 6.8 shows the inferred spatial maps. Linked ICA correctly extracts all seven sources in both the high- and low-noise data, while two of the structured noise

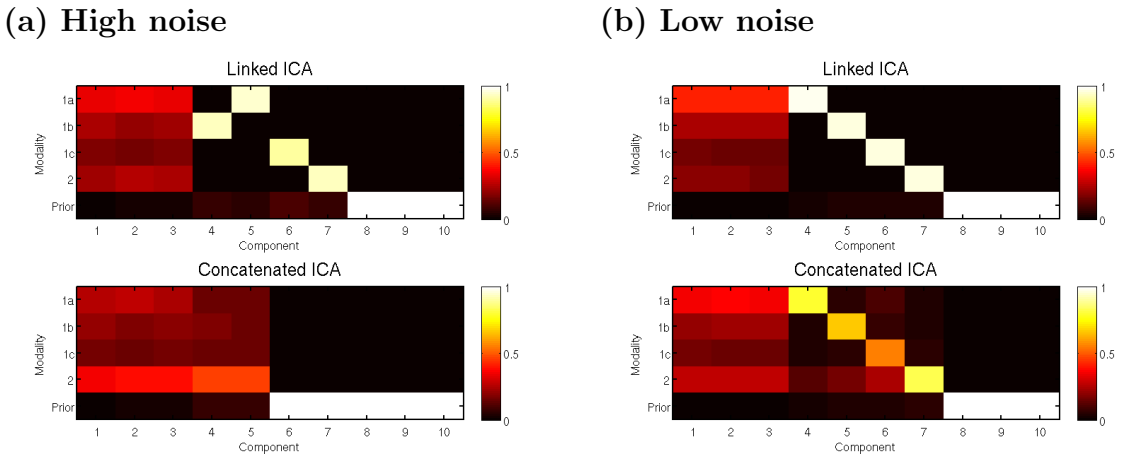


Figure 6.6: Inferred precision contributions for the simulated data set, showing which modalities dominate in determining each source's subject-course. Notice that the Linked ICA method consistently identifies the structured noise modalities (components 4–7) while the Concatenated ICA exhibits mixing (most severely in high noise).

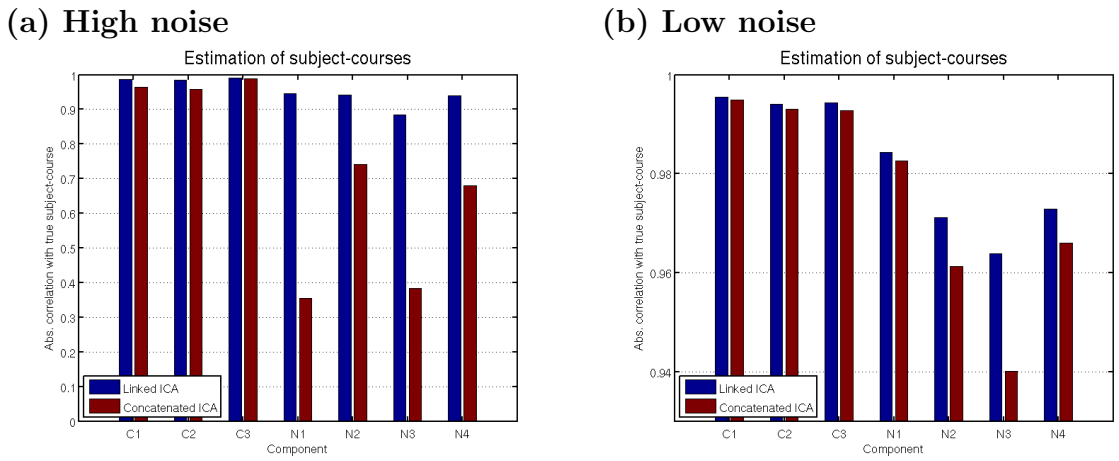


Figure 6.7: Accuracy of the inferred subject-courses for the simulated data set. On the simulated data, the fusion method more accurately recovers the true subject-courses than the concatenation method does. Note the different vertical scales. A slight advantage is shown on the common sources, but there is a large difference in their ability to accurately identify single-modality structured noise.

sources are lost in the high-noise data in the concatenated approach. All of the common-signal spatial maps are recovered well. Some slight crosstalk can be seen in the second column (C2) and is more severe in the concatenated case.

The ROC curves show that the Linked ICA method is better able to discriminate active from non-active voxels more accurately than the concatenation approach, at both noise levels. This advantage is largest in modality group 1 because Linked ICA knows that the three modalities share the same spatial map whereas the concatenated approach does not. Even so, a small improvement is observed in the maps for modality 2 where there is no such benefit. This may be partially due to the improvements in subject-course estimates or may be because it models separate histograms for modality group 1 and modality group 2.

Figure 6.9 shows that the Linked ICA model is able to infer very different histograms for each of the modalities, while the Concatenated ICA approach mixes them together. This occurs because the concatenated approach only learns one histogram per component, and all of the voxels (from all modalities concatenated) are drawn from the same mixture model. The linked approach knows which voxels belong

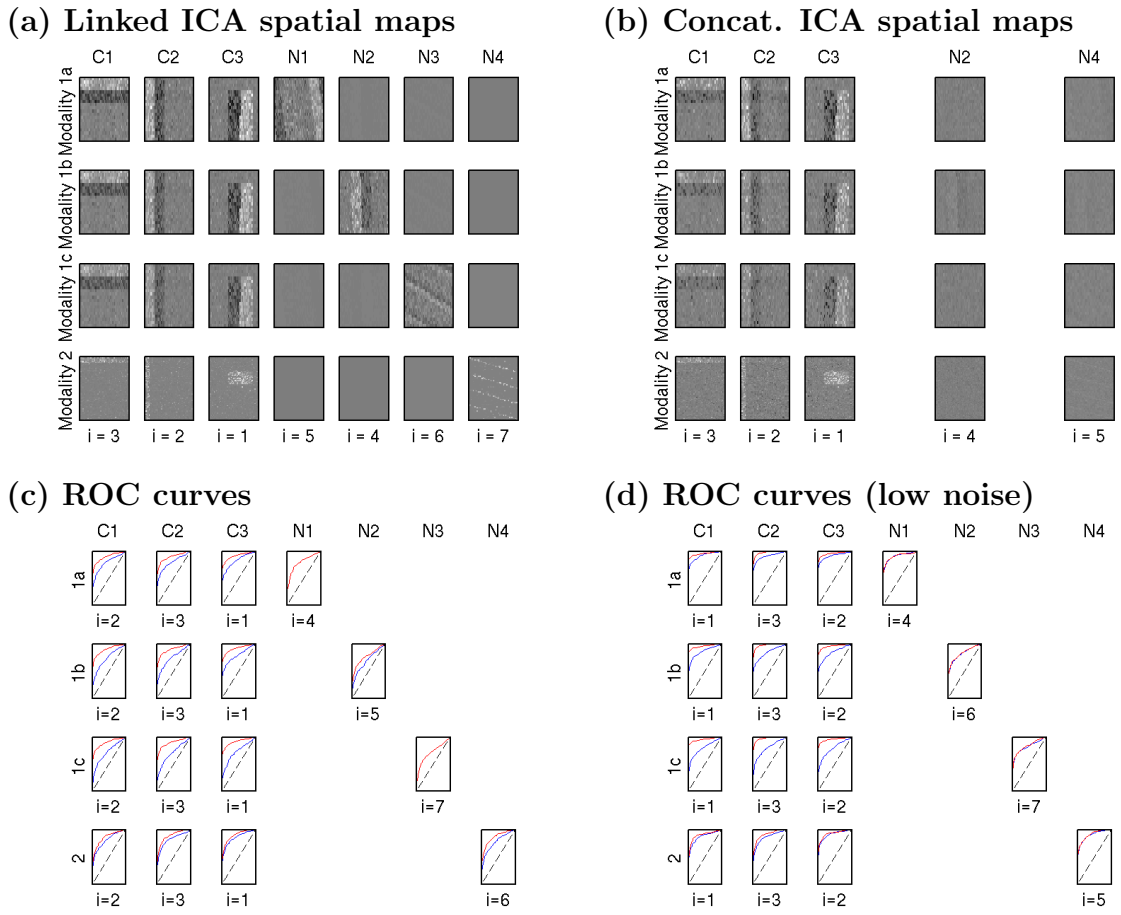


Figure 6.8: Inferred spatial maps on simulated data in high noise (a) using linked ICA and (b) using concatenated ICA. The ROC curves showing how well each method's $|\langle \mathbf{H} \rangle|$ discriminates active voxels (both positive and negative) from inactive voxels. The red lines show Linked ICA results while the blue lines show the concatenated ICA results, in the (c) high noise and (d) low noise conditions. The diagonal dashed line indicates chance.

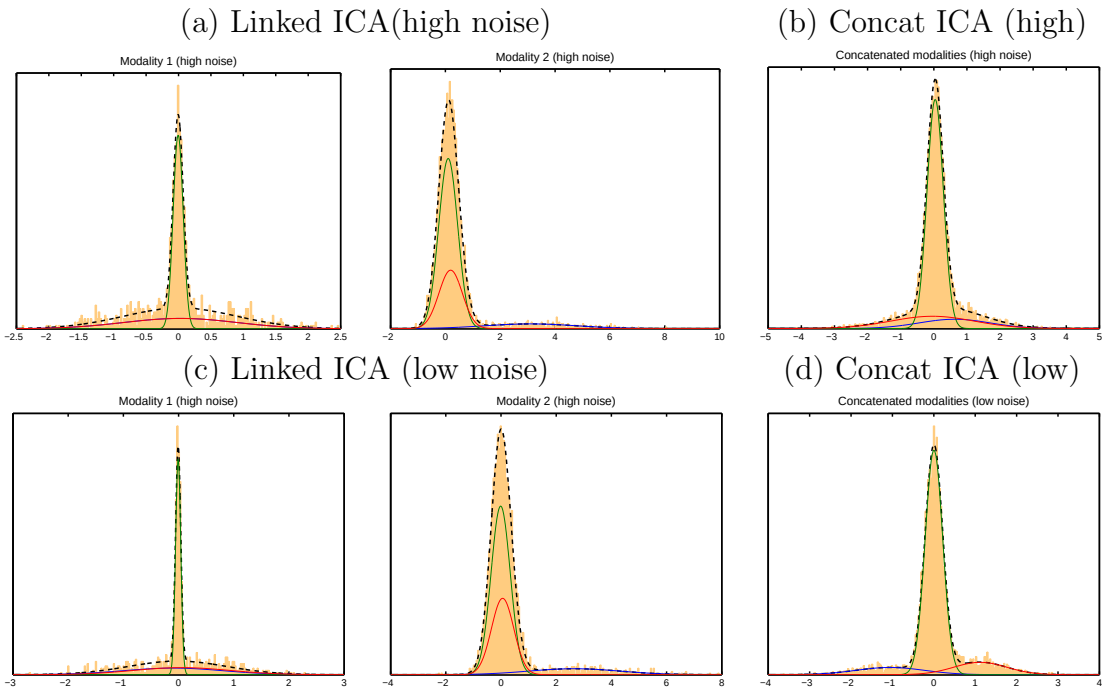


Figure 6.9: The mixture model histograms inferred for the best-estimated shared source, $i = 1$. The Gaussian mixture components are plotted with thin lines and the overall mixture prior is shown as a thick dashed line. The filled area is a histogram of voxel intensities sampled from each voxel's posterior distributions $\mathrm{P}'\left(\mathbf{X}_{n,i}^{(k)}\right)$.

to each modality and learns a separate mixture model for each modality group; this turns out to be a much better model for the simulated data because the modality groups have very different histograms.

This better prior also partially contributes to the more accurate spatial maps inferred by the linked method. Note that the histograms plotted here do not actually show the values of $\langle \mathbf{X} \rangle$ because this would not accurately represent the marginal that the mixture model prior is fitting to. In practice, histograms of $\langle \mathbf{X} \rangle$ make it look like the inferred mixture components are too wide, and the histogram is too sharply pointed. Instead, the mixture model is actually fitting to the full posterior distribution histogram from $\mathrm{P}'(\mathbf{X})$, which includes the posterior uncertainty about these point estimates. The histograms shown here are produced by taking samples from the posterior on $\mathbf{X}$; here only one sample per voxel is used, although a smoother

histogram can easily be produced by taking several samples.

These mixture models should ideally be fitting to the true source distribution that was simulated (before noise was added). In figure 6.9, the central spike is a reasonable approximation to the truth (a point probability mass at zero) but the activation distributions are modelled by a single wide Gaussian rather than separate activating and deactivating classes. This is perhaps because the heavy tails of the gamma source distribution cannot be accurately fitted by the Gaussian mixture model, so a much wider variance is inferred that travels underneath the central noise Gaussian.

6.8 Results on Real FMRI Data

As a preliminary testing step, the Bayesian ICA was tested on a single-modality FMRI data set and compared to the PICA method. Since this is single-modality data, the Linked and Concatenated models would be exactly the same. This is a short 45-timepoint data set featuring both audio and visual stimulus. This short subset of a longer FMRI scan is used in order to make it more difficult to detect these very strong stimuli. The data set was preprocessed with the recommended 3mm FWHM smoothing (on 3mm$\times$3mm$\times$3mm voxels) and high-pass filtered (100 sec). Automatic dimensionality estimation and variance normalization are left turned on.

The GLM timecourses for this experiment were given by the known stimulus timings (block designs) convolved with a canonical HRF. Although this is only an estimate of the true BOLD timecourse (e.g., it assumes a flat neural response to the stimuli and that the haemodynamic response is linear and correctly modelled), it seems reasonable to assume that a higher correlation between the inferred timecourses and the GLM regressors is better.

PICA inferred a dimensionality of 15. As preprocessing for the Bayesian ICA, variance normalization was performed as described in 6.4.2, and the maximum number of components L was set to 40. With the Bayesian ICA, 32 components were inferred

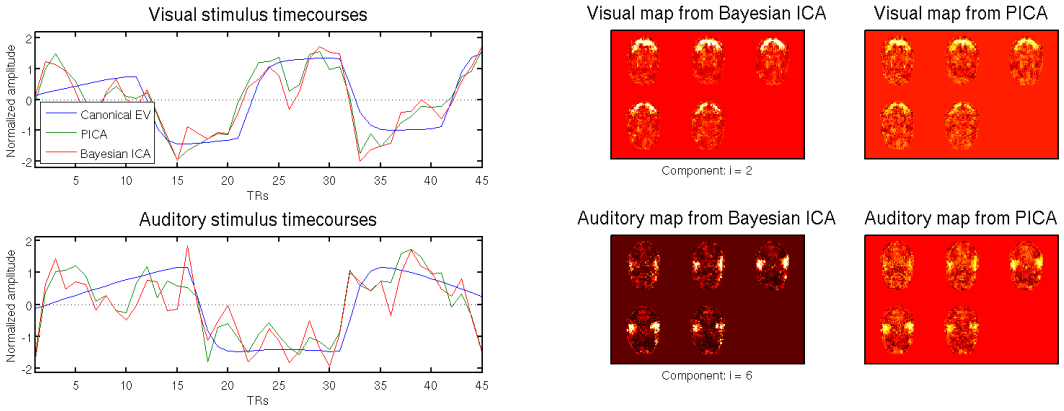


Figure 6.10: Timecourses and spatial maps from the components corresponding to the visual (top) and auditory (bottom) stimuli in the real FMRI data set. Both methods recover the audio and visual components well. The timecourses are both quite similar to the GLM regressor, with PICA matching the regressor somewhat more accurately. The spatial maps also show very similar activation of the visual and auditory cortex, respectively. Note that the intensities are each scaled to the minimum and maximum voxel intensity in each map (hence the different colour of the background, which is set to zero, in each image).

(8 eliminated). In the early components there are clearly some of the same structured sources that PICA pulls out; this suggests that these are sensible components to extract, although of course there is no ground truth. However, the later components extracted by Bayesian ICA have no obvious spatial structure, suggesting that they are probably not interpretable or useful. These extra components could potentially be due to the extra structure in the data introduced by spatial smoothness (since spatial correlation is not explicitly modelled).

A strong visual component and a strong auditory component are extracted cleanly by Bayesian ICA, as shown in figure 6.10. The main visual and auditory components are picked out easily, achieving slightly lower correlation with the canonical timecourses than the PICA approach does. The spatial maps clearly show activation in the visual and auditory cortex.

Two of the components are shown: At the top, with 78% correlation to the visual regressor; and at the bottom, with an 67.6% correlation to the auditory

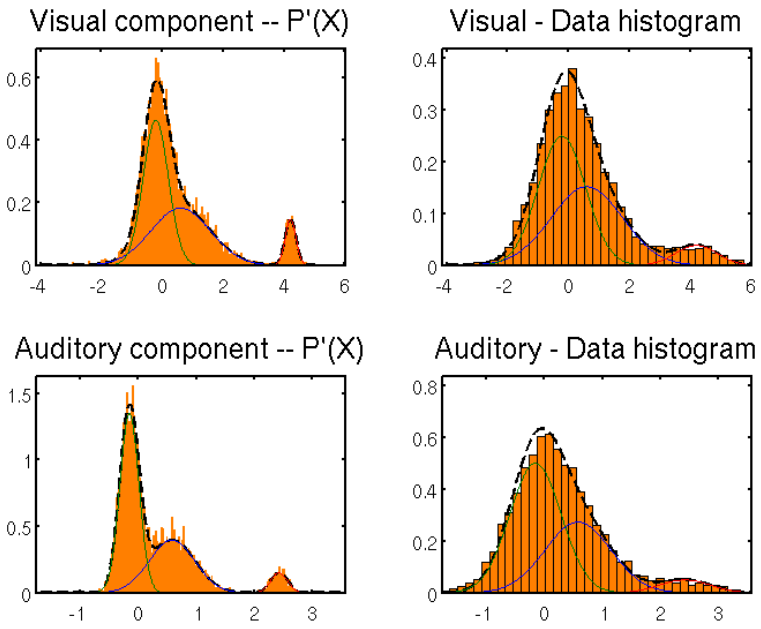


Figure 6.11: Histograms of the spatial maps inferred by Bayesian ICA for the visual and auditory components. The left side shows the estimated noise-free source distribution (samples $\mathrm{P}'(\mathbf{X}_{\cdot,i})$). The right side shows the result of projecting the data in the direction given by the $\mathbf{H}$ – since this includes the noise, it should correspond to a blurred version of the noise-free source histogram. The mixture components are superimposed on this, which have the variance of the original source plus the noise variance that is parallel to the component's timecourse. Notice that the extra bump on the right side of both noise-free source distributions actually corresponds to a heavy tail once the effect of noise is included.

regressor. PICA extracts these components somewhat better, with 83.8% and 75.6% respectively. This cannot be corrected simply by forcing a lower dimensionality; repeating the experiment setting the maximum number of components to match the PICA dimensionality estimate, $L = 15$, only increases the timecourse correlations to the canonical EVs very slightly (by 0–2%) and leave the histograms and spatial maps nearly unchanged.

Figure 6.11(left) show the Bayesian model's posterior prediction of the noise-free source distribution. There is a strange additional bump on the right side of both components' histograms, which is a consequence of trying to fit the tails of a heavy-tailed source distribution obscured by noise. The "data histograms" in figure 6.11(right) shed some light on this. These histograms were obtained by projecting the original data onto the space given by the component timecourses $\mathbf{H}$:

$$\mathbf{Y}/(\langle\mathbf{H}\rangle\, diag(\langle\mathbf{W}\rangle)). \tag{6.13}$$

This is the histogram that would be observed in a noise-free model like PICA. The difference is that this second histogram is derived from the noisy data, and part of this additive white noise will be in the same direction as the signal component. The effect of this is to blur the histogram, so the Bayesian model is effectively trying to deconvolve a sharper source histogram from the original noisy data. (While these equations 6.13 and 6.14 assume that there is only one modality ($K = T_k = 1$), they could potentially be generalized to find the equivalent data histograms for each modality in a multi-modality setting .)

Under the generative model it is possible to predict what should be observed in the data histogram for component i, given the inferred source distribution and noise level. The resulting distribution is also a Gaussian mixture model, with the same component weights $\boldsymbol{\pi}$ and means $\boldsymbol{\mu}$, but broader variances $\mathbf{v}$ given by

$$\mathbf{v}_{m,i} \;=\; {\boldsymbol{\beta}_{m,i}}^{-1} + \langle\lambda\rangle \left\langle\mathbf{W}_i^2\right\rangle \sum_{r=1}^{R} \left\langle\mathbf{H}_{i,r}^2\right\rangle \tag{6.14}$$

where the second term is the noise variance, passed through the same projection given in equation 6.13. These broadened mixture models are a reasonable fit for the data histograms in figure 6.11(right), and look quite sensible even though the inferred source distribution is unrealistic.

Overall this demonstrates that the Bayesian ICA model is reasonably effective at extracting strong sources in real data, although the Gaussian mixture model is not well tuned to the heavy-tailed activation observed in FMRI data. In future, using a gamma-Gaussian mixture model (like the one in PICA) may help to avoid these unrealistic models of the source distribution.

6.9 Analysis of Structural & Diffusion Data

In the final experiment, Linked ICA was compared to Concatenated ICA and PICA in the task of extracting independent components from a structural and diffusion data set with 47 probable Alzheimer's patients and 46 age-matched controls. Exploratory techniques can be used to find inter-subject variability and identify whether any of these are correlated with regressors of interest; the Linked ICA approach provides a way to perform this across multiple modalities in a data set. Both grey matter density and white matter integrity have previously been used as biomarkers for neurodegeneration. To assess these, structural and diffusion scans (introduced in section 1.2.3) were collected for these subjects. The diffusion scans were preprocessed to extract maps of Fractional Anisotropy (FA), Mean Diffusivity (MD) and Tensor Mode (MO). These were projected onto a white matter skeleton using TBSS, which improves registration and makes sure that observed differences are due to white-matter tract properties and not just movement or misregistration. The grey matter (GM) partial volume maps were extracted using the FSL-VBM tools (including non-linear registration).

Generally, neurodegeneration results in reduced GM density, decreased FA, and increased MD. Tensor mode MO is another measure derived from the diffusion tensor which is mathematically orthogonal to the other two (FA and MD) and is related to whether diffusion is restricted in a line or in a plane (Ennis and Kindlmann, 2006), and therefore may have significance for assessing degeneration in areas where fibre bundles cross.

The TBSS modalities were resampled to $(2\text{mm})^3$ voxels on the skeleton, yielding $N_1 = 28997$ voxels. The grey matter maps were already smoothed by 3mm FWHM to account for anatomical differences, and then nearest-neighbour downsampled by a factor of two yielding $N_2 = 23280$ voxels in an effort to reduce noise correlation. This is because the current implementation of Linked ICA does not model (spatial) correlation in the noise. The data sets were preprocessed by de-meaning across repeats and then rescaling each voxel in proportion to an estimated noise level using the robust noise estimate. In addition, each subject's maps were individually de-meaned, so that overall GM or FA levels are removed. In practice this tends to dominate ICA decompositions and can spread across several components, but it is useful information that could potentially be reintroduced, although that has not been done at present. Furthermore, the results shown were initialized based on the PICA results instead of the PCA method described in section 6.4.3; this produces more easily comparable components.

This concatenated data set was inputted into PICA, with no further preprocessing apart from the variance normalization described in section 6.4.2. Automatic dimensionality estimation (ADE) determined that there were only 6 sources in this data set.

Despite PICA's low dimensionality estimate, both the Linked ICA and Concatenated ICA approaches failed to eliminated any components, even with $L = 90$. Most of these components showed spatial patterns that resembled spatially-smoothed

white noise, indicating that there is smoothness in the noise which is not included in Linked ICA's uncorrelated white noise model. In contrast, PICA estimates the smoothness of the data and uses this in its estimation of the number of components. Instead of allowing Linked ICA to use this many components, the number of modelled dimensions was fixed at this estimate of $L = 6$.

The VB inference was allowed to run for 2000 iterations; the $\Delta F < 0.1$ condition was usually satisfied by this point. This took about an hour for Linked or Concatenated ICA (compared to a few minutes for PICA).

6.9.1 Results on Structural & Diffusion Data

The precision contributions for each model are shown in figure 6.12. In Concatenated ICA, most of the components are spread across all of the modalities. In Linked ICA there is much more separation between the modalities: some components appear to explain variability in the white matter only, while some are shared between white matter and grey matter. It is difficult to assess whether this is separation is realistic. Note that this sparsity in modalities is not just due to ARD eliminating components that are too weak; a number of components make an 8–25% contribution in the Concatenated results but are reduced to $\approx 1\%$ in Linked ICA. This suggests that sparsity in modalities (i.e. allowing whole modalities to be eliminated from a component at once) may be a driving force that affects how the components are unmixed; this may be similar to the rotation of components observed in chapter 5.

To evaluate this, the inferred subject-courses are correlated against the EVs for Alzheimer's status and for age. Figure 6.13 shows a slice from the spatial maps of the components with the highest-correlated matches. Linked ICA shows quite a low correlation in the Alzheimer's status, and completely eliminates GM; however, there are several other components with nearly equal correlation. Linked ICA shows a higher correlation with the age EV than the other methods.

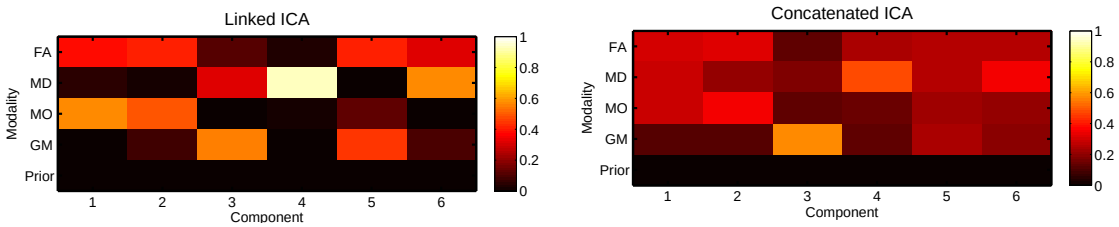


Figure 6.12: Precision contribution plots for the Alzheimer's data set. Linked ICA uses two modality groups: the three DTI modalities (FA, MD, and MO) form one group, while GM is by itself. This method has eliminated some modalities from some of the components, for example MD has been eliminated from components 1, 2, and 5. In Concatenated ICA, all components are shared across the four modalities much more evenly.

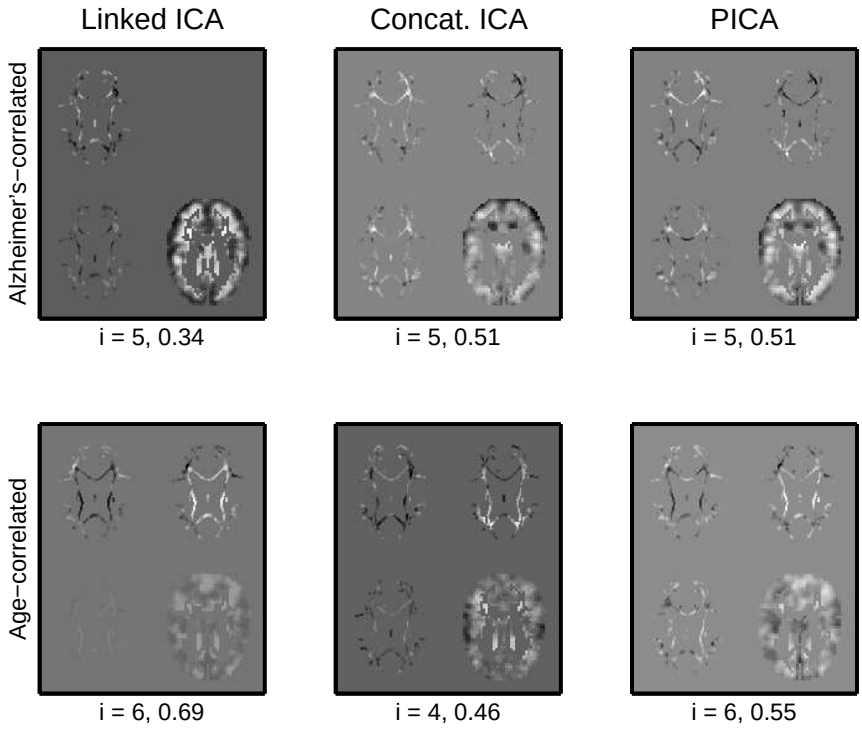


Figure 6.13: Spatial maps from the real multi-modal data analysis, from the components most highly correlated with two EVs: a binary Alzheimer's-vs-controls regressor and the subject's age. Component numbers and correlation coefficients (between the inferred subject-course of that component and the true subject regressor) are shown below each image. Clockwise from top left within each panel, the modalities are FA, MD, GM and MO. Note that the component numbers match those in figure 6.12; for example, in the top-left image MD has been eliminated from i=5, and the corresponding precision contribution of zero. Only the central slice is shown.

	Linked ICA	Concatenated ICA	PICA
Alzheimer's status	66%	72%	73%
Male/female	61%	66%	68%
Age (normalized)	0.722	0.710	0.708

Table 6.1: Sparse classification (and regression) performance using only the subject-courses as inputs (i.e., 6 values per subject). These are leave-one-out cross-validation accuracies. Since age is a continuous variable the RMS estimation error is shown. The established PICA model is the best in all cases, while the concatenated model appears to outperform the flat linked ICA model in all cases. This poor performance may be related to the fact that the number of components (6) has been optimized for PICA.

6.9.2 Evaluation in Terms of Sparse Classification Accuracy

Correlation assesses how well an EV is modelled as a single component, but it is possible that the information could be split over several components but still encoded very accurately. In order to get a more definitive assessment of the methods, the subject-courses were used as inputs into a sparse classification algorithm.

Since the components are independent and many of them may be modelling structured noise rather than useful signal, it makes sense to use a sparse model that avoids overfitting by only selecting those sources that are significantly relevant. It was simple to adapt the decoding module from the previous chapter to this task, just inputting a point estimate $\langle \mathbf{H} \rangle$ as data, and using either the single-regression or single-classification model as appropriate, with leave-one-out cross-validation. In tests on the simulated data, this approach consistently picked out the single highest-correlated regressor and eliminated all of the other components.

Surprisingly, the resulting classification or regression was consistently better with Concatenated ICA than Linked ICA. PICA also beats both methods. Using the PICA results as initialization instead of an initial PCA makes very little difference to the output. In particular, some of the components extracted by PICA have markedly different histograms in the different modality groups; Linked ICA would be able to model this more accurately than the Concatenated method, yet performs worse. It is

possible that the mixed histograms of the Concatenated model reduces the effective degrees of freedom and avoids overfitting to the histogram; perhaps this regularizing effect accounts for the better performance. It is also quite possible that the sparsity-in-modalities that Linked ICA encourages is actually counterproductive and discarding useful information.

6.10 Results on DTI Measures Only

One of the current concerns about the Linked ICA model is that it may be sensitive to differences in the amount of smoothness between modalities. Spatially-correlated noise is not modelled at all, which can cause the smoothest modality to dominate the decomposition. In this section, the same data set is used but the very smooth GM map is removed and only the three diffusion modalities are considered. In addition, the individual subject maps were not de-meaned.

Since all of these modalities exist in the same space, they can be combined in a single modality group to give a fully tensorial model. They can also be linked (with different histograms and spatial maps), or a combination of the two can be used. The concatenation approach is used as the baseline. These configurations are presented in figure 6.14. For simplicity, this section makes no comparisons to PICA, and the dimensionality is also fixed (arbitrarily) at $L = 20$.

The following models were evaluated:

- **A** is the fully tensor model with $K = 1, T_1 = 3$.
- $\mathbf{B_{FA}}$, $\mathbf{B_{MD}}$ and $\mathbf{B_{MO}}$ are three partially tensor models, with $K = 1$ and $T = [2, 1]$. The subscript indicates the modality that is not part of the tensor.
- **C** is the flat linked ICA model, with $K = 3$ and $T = [1, 1, 1]$.
- **D** is the standard spatial-concatenation model. This is different from **C** because the same histogram is used for all modalities and there is no sparseness in

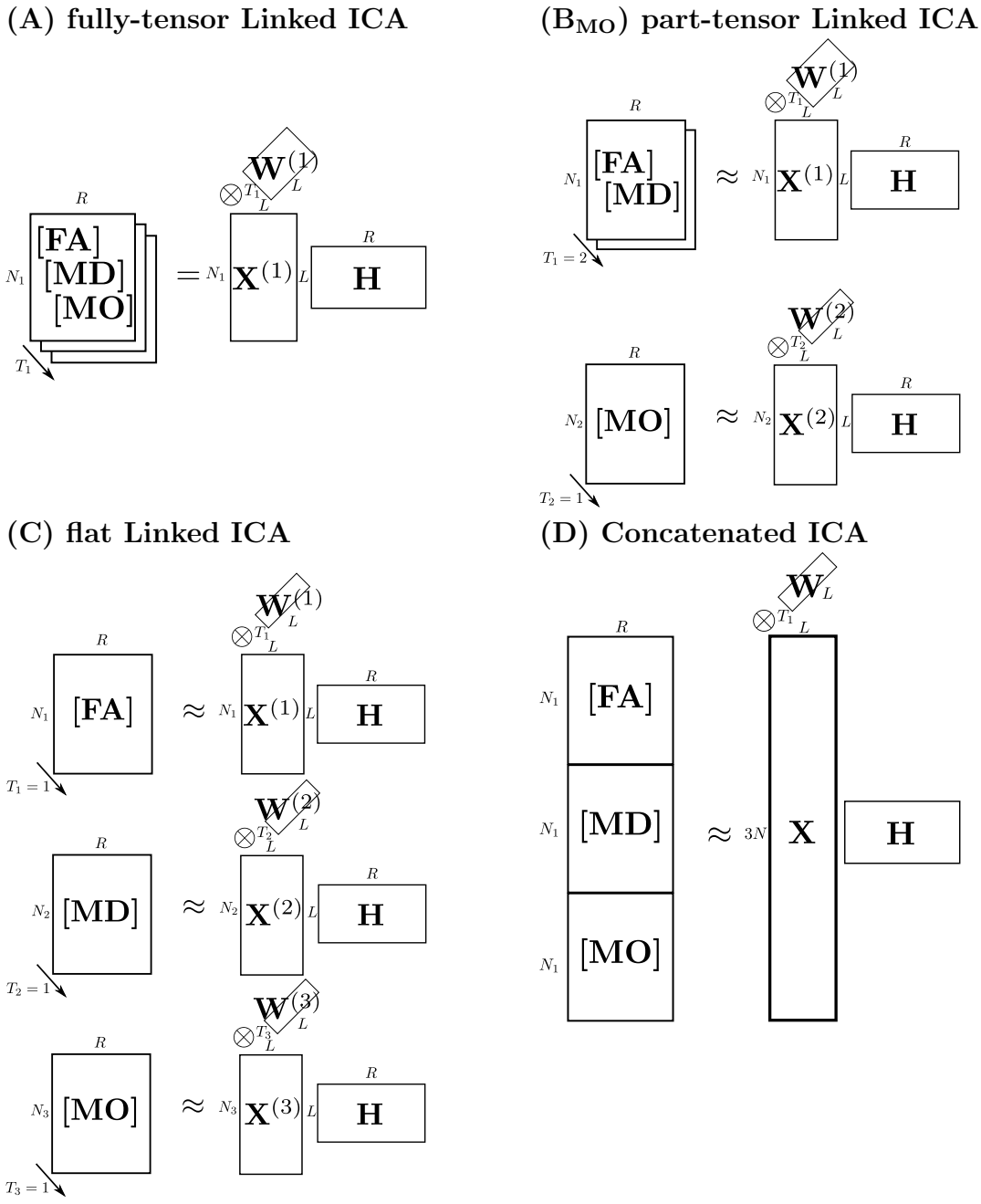


Figure 6.14: Matrix diagrams of the ICA configurations evaluated for the diffusion data: (**A**) fully tensor, ($\mathbf{B_{MO}}$) partially tensor with MO isolated (there are also similar $\mathbf{B_{FA}}$ and $\mathbf{B_{MD}}$ configurations not shown), (**C**) flat linked; (**D**) concatenation. Note that the differences between models **C** and **D** are that the linked model infers a different histogram for each modality and allows sparsity in modalities.

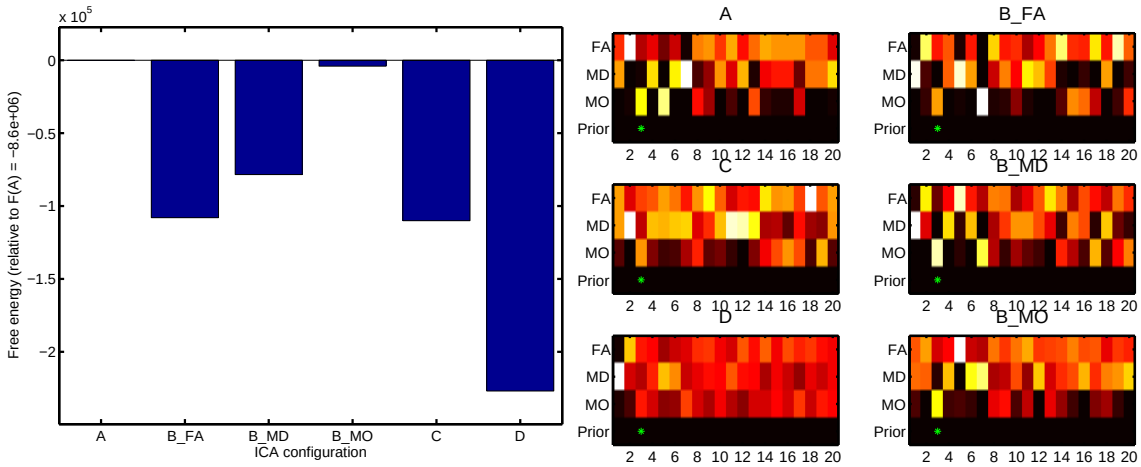


Figure 6.15: (Left) The free energy after inference on each of the proposed configurations for DTI, relative to the Tensor ICA model (**A**) which has the highest evidence. The flattened Linked ICA (model **C**) performs noticeably worse, with the partial-tensor models **B** mostly between these extremes. Interestingly, the concatenated method actually has a much higher evidence than the flat linked ICA method, possibly because the limited number of components $L = 20$ affects them differently. (Right) Precision contributions of the three DTI modalities, under each of the configurations. Note that the components are much more mixed in models **C** and especially **D** than in the tensor-based models. The green dot in the last row indicates one particular component that is highly preserved across all methods ($i = 3$ in all approaches, all $> 93\%$ correlated timecourses to each other). Note that the pattern of precision contributions remains qualitatively similar in all cases (mostly MO with some FA).

modalities. **D** also assumes the same noise variance for all modalities, but this is a reasonable assumption because the voxelwise noise-variance normalization should leave all voxels (in all modalities) with the same noise level.

Each of these six models is allowed to run for 2000 iterations.

First, figure 6.15 shows a comparison of the free energy of each model, The fully tensor model (**A**) performs best, although moving MO out of the tensor only decreases the free energy slightly (by over 4000, so still significant). In contrast to this, isolating either FA or MD decreases the evidence by a large amount. This might imply that the true spatial patterns in FA and MD are very similar, while MO has somewhat different patterns. Similarly, the flat linked ICA (C) is slightly worse than the partial-

	Model A	$\mathbf{B_{FA}}$	$\mathbf{B_{MD}}$	$\mathbf{B_{MO}}$	Model C	Model D
Alzheimer's status	80%	76%	78%	75%	78%	80%
Male/female	72%	80%	74%	73%	75%	68%
Age (normalized)	0.715	0.747	0.727	0.738	0.719	0.723

Table 6.2: Sparse classification (and regression) performance on the DTI-only data set using only the subject-courses as inputs (i.e., 20 values per subject). These are leave-one-out cross-validation accuracies (each of the classifications is approximately balanced). Age is a normalized continuous variable, and the RMS estimation error is shown (so chance = 1.0). These accuracies do not tell a consistent story; for example, the full tensor model **A** is the best for predicting age and Alzheimer's status (tied), but almost the worst at predicting sex.

tensor models (although by 2000 compared to $\mathbf{B_{FA}}$). The concatenated method (D) has much lower evidence than any of the other approaches.

To see if these differences are due to initialization, each method was initialized using three methods: the PCA decomposition, the **A** subject-course, and the **D** subject-course. The result that had the largest free energy was used. In general, the difference in F due to different initialization were quite large (around 50–500 points) but not big enough to change the overall model comparison results. This suggests that the Linked ICA model might have problems with local minima, and a multistart approach (picking the highest-F result) could help with this.

This free-energy comparison is not necessarily fair because of the restricted number of dimensions ($L = 20$ was chosen arbitrarily). It would be preferable to allow each model to select the correct number of components, especially because models **A**–**C** will give sparser solutions (by excluding some modalities) and may choose to have more components than the concatenated model **D**. Unfortunately, re-running the model with a range of values up to $L = 80$ still did not eliminate any components. This is not an inference problem, but rather one with the model: as shown in figure 6.16 the evidence continues to increase as more components are added. Finding the source of this problem and resolving it is a key direction for future work on this model.

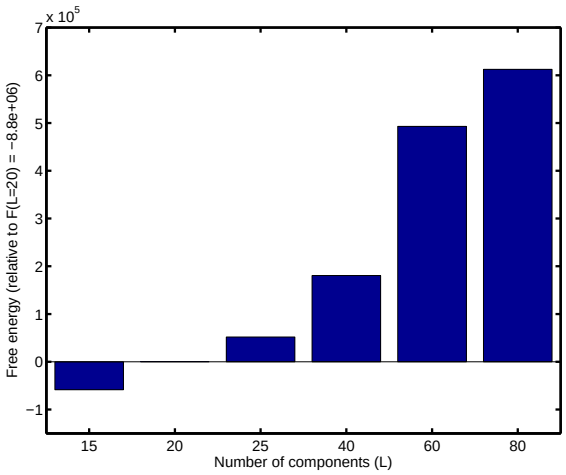


Figure 6.16: The free energy on the tensor model ($\mathbf{A}$) for different settings of the number of initial components L. Even with $L = 80$, none of the components were eliminated (on a group of $R = 93$ subjects). This comparison of free energy shows that these components are indeed modelling something structured in the data; one possible candidate is unmodelled spatial smoothness in the noise.

6.11 Discussion

The Linked ICA method presented in this chapter provides a flexible way to perform ICA on multi-modal data sets, by allowing a combination of spatial concatenation and tensor ICA in the same model. This allows components to be sparse in modalities and allows different noise levels and histograms for each modality group.

Linked ICA performed well in simulated data, and combined information from across modalities more accurately than standard Bayesian ICA with spatial concatenation. In estimating spatial maps, it benefits from the partial tensor configuration and from the more accurate histogram estimates found by separating voxels into modality groups.

On real data with four modalities, it seems to be outperformed by the simpler Concatenated approach, and PICA is better still. The histogram-estimation on the FMRI data suggests that the Gaussian mixture model causes overfitting on real neuroimaging data and it may be that the concatenated ICA's shared histograms provide regularization (and more voxels) to keep the mixture models more sensible.

It is also possible that inferring a separate noise precision for each modality is unnecessary, and it is better to instead rely completely on the robust noise estimation to find the right noise level.

However, removing the GM modality and only considering the DTI modalities made the results more reliable and interpretable. Model comparison demonstrated that the tensor model is preferred for these, but all of the methods tried were ranked higher than the concatenated model. In terms of classification accuracy the results are more mixed, suggesting that this may not be the best way to assess these methods when the differences are this small.

The PICA model has been very successful using a 3-component model with positive and negative gamma distributions (with hard constraints on the mode) to model activation and de-activation and a Gaussian between them to model the noise in the inactive voxels. The Gamma-Gaussian is a preferable model for neuroimaging data because of its heavy tails and the fact it is restricted to only explaining one sign of the data; it is common to have only activations or only deactivations within a single spatial map, and if this is fitted using a Gaussian the opposite tail can cause problems. This is examined in detail on single-subject FMRI analysis in Beckmann (2004). Future work will involve modifying this mixture model to incorporate a Gamma-distributed activation model, although the constraints of VB to having conjugate priors means this will require further approximations. An alternative would be to keep the Gaussian mixture model and try to control the qualitative form of the mixture model using priors.

Further work is needed to bridge the gap between the (rather ideal) simulations and real multi-modal data. One unmodelled aspect of the data is spatial smoothness, which tends to cause Bayesian ICA to overestimate the number of components in simulated data (results not shown). This could be modelled either by spatial autocorrelation or by a Gaussian-process-based noise model.

One future extension of this model would be to take advantage of spatial information by using a spatial mixture model prior on $\mathbf{q}$ to emphasize the fact that activation tends to occur in clusters; Woolrich and Behrens (2006) have previously used this in a VB framework.

There is also the possibility of applying these techniques to non-MRI neuroimaging data. In MEG and EEG, tensorial decompositions like PARAFAC are a natural way to model the space×time×frequency information in single-subject data (Miwakeichi et al., 2004), and in ICA has been used in this context to localise sources. This framework only requires that the modalities share a single dimension (e.g. subjects), so finding covarying patterns in modalities as different as EEG and FMRI may still be possible. The major challenge would be in finding appropriate preprocessing methods to keep the data size down and avoid needing too many components just for structured noise.

Chapter 7

Conclusion

7.1 Summary of Contributions

This thesis presents several novel Bayesian approaches for univariate and multivariate modelling of neuroimaging data. These are evaluated using appropriate data sets for each method, including dual-echo ASL, BOLD FMRI, multi-inversion ASL, LFP recordings, and multimodal structural/diffusion group data. The major contributions arising from this thesis work are outlined below:

Correlated noise modelling for multimodal MRI time-series

Chapter 3 looked at noise correlation issues in modelling a multimodal data set using a single nonlinear generative model, rather than modelling each modality separately and linearly. Simultaneously-acquired multimodal data can contain between-modality noise correlations which can bias results, and using a Bayesian framework it is possible to infer these correlations as part of the model inference. This chapter developed a way of modelling noise correlation as part of a standard autoregressive noise modelling approach. For dual-echo ASL, modelling the noise correlation makes a clear difference in the uncertainty inferred on the CBF and BOLD changes, in a way that is consistent with the type of noise correlation expected in the biophysical model.

Gaussian process prior for combining biophysical priors with spatial regularization

A novel Gaussian process prior was developed which makes it possible to provide spatial regularization on a signal parameter while still correctly enforcing an informative biophysical prior on parameter values. This is applied to two different applications in which the informative prior has previously been shown to be valuable: constrained basis set models of FMRI HRFs and arrival time estimation in multi-inversion ASL data. The prior's behaviour is explored across a range of possibilities using simulated FMRI data and was consistently found to balance the informative prior information and spatial smoothness in a sensible way. On multi-inversion ASL data, more accurate arrival time maps were found from cut-down data sets. This work also demonstrated the ease with which different priors can be used for each signal parameter in the non-linear signal model (and by extension, the linear Bayesian GLM).

Integrated Decomposition and Decoding

The novel data fusion model proposed in chapter 5 allows several different types of data to be linked together probabilistically using a shared, low-dimensional latent space. This is interpreted differently by the generative models for each data type, providing a great deal of flexibility. The way that the neuroimaging data and the behavioural variables are jointly decomposed into components can provide insight into the decoding being used and what structure exists within the data.

The specific supervised learning application of this framework is an integrated decomposition and decoding model that links a generative PCA model of the neuroimaging data to an ARD-based sparse decoding model, inferring on both simultaneously. This allows the components to be refined based on feedback from the decoding stage, which in simulations results in more accurate reconstruction of the data features and improved decoding accuracy.

This framework also provides a natural way to decode multiple behavioural variables at once, which means that all variables are decoded in terms of the same data components. A general positive-definite Gaussian noise model allows for noise correlation between the different behavioural variables.

Linked ICA approach for joint analysis of multimodal data

The same basic data fusion model was applied to an unsupervised learning problem, using ICA to find the interesting patterns across several modalities of structural and diffusion data. This is a more powerful model than simple spatial concatenation ICA because it knows which voxels belong to each modality and can customise the model histograms and component elimination to each modality's needs. On simulated data, it was shown that this causes cleaner isolation of single-modality structured noise and more reliable estimation of the true model order. By inferring different histograms on each modality, more accurate spatial maps can be inferred.

The model also permits new combinations of tensor and concatenated ICA by stacking only those modalities that are expected to share spatial maps. Model comparison on the DTI modalities demonstrated that free energy can be used to choose between the various possible arrangements. In this case, the tensor arrangement of the modalities was preferable.

7.2 Problems to Overcome and Future Directions

Modelling noise correlation structure

The work on dual-echo ASL data demonstrated the importance of modelling noise correlations to avoid overconfidence in parameter estimates, but the correlated noise model developed in chapter 3 is not particularly principled. The general positive-definite noise model used in the decoding stage in chapter 5's model would also be an excellent candidate for modelling inter-modality correlated noise. Validation on

simulations could assess how well each model fits noise generated by the other, and model comparison would be the best way to assess which noise model is a better fit for the dual-echo ASL data. Although the free energy calculated by the linearized VB model is not actually a lower bound on evidence, the comparisons should be valid as long as the linearizations are similar.

Temporal autocorrelation could also be a valuable addition to the PCA decomposition. In the LFP time-series analysis these were heavily downsampled, but at higher temporal resolutions this non-white noise could lead to extra PCA components being preserved. It would be very valuable to incorporate an adaptive model of AR noise on time-series decompositions since this will also be important for analysis of EEG, MEG, and spatio-temporal FMRI data.

The Bayesian ICA approaches kept far more components than PICA, in all of the real MRI data sets examined. However, it did correctly estimate the dimensionality in the simulations. This suggests that there is some sort of modelling error that causes these extra components to be a more parsimonious explanation of the noise than white noise is. One potential culprit is spatial smoothing; in additional simulations (not shown) with spatial smoothing applied to the noise, many extra components were observed, and these spatial maps looked like smoothed white noise. Within the Bayesian framework there are many ways to model noise correlation, and one of the simplest initial steps in this would be to model it as *spatial* autoregressive noise, which uses very few parameters and may help to ensure that the Linked ICA components are modelling spatial structure rather than just explaining simple spatially-smooth noise.

Application of the fusion approach to functional modalities

The fusion methods developed in chapters 5 and 6 should be applied to FMRI data. The combined decomposition/decoding approach, especially with the inclusion of ICA priors on the decoding, could offer an exciting way to potentially decode FMRI from

data-derived components, while simultaneously learning the remaining structure in the data that is not used in the decoding. Since the rows of the latent space are not exactly determined by the behavioural variables, but linked to them through a sparse decoding and a noise model, there would be more flexibility in the way that informative components are extracted from data sets.

In particular, the Linked ICA approach could easily use activation maps from FMRI studies as input, or maps extracted from group resting state network analysis (Beckmann et al., 2005). These can be combined with stimulus-related functional maps from other scans of the same patients, or behavioural/genetic variables.

This framework's separate generative models make it very flexible for dealing with assorted types of multimodal data, and the potential of this has not yet been exploited by the relatively simple models used. One particularly suitable application of this will be tensor decoding of MEG data. ICA is already an accepted method for extracting sources from MEG, and it would be interesting to see how the selected components change when the trial-courses are informed by behavioural variables. There is also potential to use a wavelet transform (or short-time Fourier transform) to convert this into a 4-dimensional space of sensors×frequency×time×trials.

A key part of this future work will be direct comparisons between this decoding approach and existing supervised learning methods. It is important to look at more than just classification performance; the decomposition provides interesting information about data structure by itself, and it would be interesting to see if this generative model of the neuroimaging data helps to regularize the model to make the inferred decoding vectors any more stable or consistent.

Alternative models in the data fusion framework

One of the major benefits of the fusion framework is that each modality can be described using its own generative model, which allows different noise models, priors, and even data arrangements (e.g. as a tensor). This can also be extended

to incorporate binary rather than continuous data, by using a variational logistic regression model with an accompanying lower bound on the free energy to ensure that model comparison is still valid (Bishop, 2006, Ch. 10.6).

One interesting idea would be to use a non-negative matrix factorization (NMF) for one part of the model, which would be useful if one of the modalities is restricted to be positive (e.g. a set of genetic data). NMF is beginning to see some use for meta-analysis of neuroimaging data (Nielsen et al., 2004) and has been formulated as a VB model (Cemgil, 2009).

Since the positivity constraint fundamentally changes the nature of the decomposition, it is likely that all modalities will have to be positive. This might be the most sensible way to represent frequency-decomposed MEG data, representing the signal as power rather than amplitude in order to ignore phase.

Machine learning on multimodal data

Multimodal data still represents a major challenge for machine learning techniques. One of the paradoxes in supervised learning is that in many cases adding additional data to the feature vector (e.g. adding another modality to a discrimination task) often reduces classification performance significantly. That can happen for a number of reasons, most obviously if the new modality is pure noise but has many more voxels than the modalities that are useful for classifying. Using an adaptive Bayesian approach that is explicitly aware of the divisions between modalities should at least not be harmed by the addition of another modality. Even if adding modalities does continue to degrade performance, the fusion framework provides more insight into its internal operations, making it easier to compare the decompositions and decodings that are inferred in both cases and determine *why* the changes happened, and suggest ways to avoid these problems in order to make optimal use of all of the information provided.

MEG data is temporally well-localized and the challenge lies in *spatial* decoding to find the location of sources. ICA's ability to specify spatial sparsity (where sparsity is defined by the form of these priors) means that different priors and basis sets can be used tune the method to the problem at hand. Furthermore the model can be constructed to allow simultaneous signal modelling and artefact removal, for example by explaining the data using both time-locked ICA to extract components and additional ICA components on the whole original time-series to model artefacts.

7.3 Final Conclusions

Bayesian modelling is a powerful way to extract signals and patterns from neuroimaging data through a variety of generative models. One of the greatest strengths is the ability to directly express our understanding of brain function and signal generation in the generative model, so that results can be more interpretable. This interpretability is particularly powerful because of Bayesian modelling's uncanny ability to outwit its inventor and detect patterns and structure in the data that were previously overlooked. Supervised learning models are becoming increasingly interpretable under Bayesian modelling, and this trend represents the best way forward in understanding how information is encoded in the brain and discovering the underlying mechanisms of thought.

Appendix A

Generative model for dual-echo ASL

The total magnetization for timepoint t is given by

$$S(t) = M(t) + B(t) \tag{A.1}$$

where $M(t)$ and $B(t)$ are the time-varying magnetization due to static tissue and blood, respectively. Note that the static magnetization changes are not well understood but are likely to be a combination of outflow effects and activation-related changes in cerebral blood volume. At a given timepoint t, the magnetization of the CBF (in scanner units) is given by $Q(t)$. The blood magnetization is defined by the flow rate, multiplied by the magnetization of each piece of the flow times that piece's duration:

$$B(t) = Q(t)\ (pretag\ \Delta t + bolus\ TI_1 + posttag\ (TI_2 - TI_1 - \Delta t)) \tag{A.2}$$

$$pretag = 1 \tag{A.3}$$

$$bolus = \begin{cases} 1, & \text{for control timepoints} \\ 1 - 2\alpha \mathrm{e}^{-TI_2/T_{1b}}, & \text{for tag timepoints} \end{cases} \tag{A.4}$$

$$posttag = 1 - \mathrm{e}^{-(TI_2 - TI_1)/T_{1b}} \tag{A.5}$$

where *pretag*, *bolus*, and *posttag* give the relative magnetization of blood compared to the initial magnetization of 1. Only the *bolus* component is affected by the tagging,

which has a fixed length but unknown arrival time Δt. The *posttag* piece is saturated (magnetization destroyed) at time TI_1.

This magnetization is converted into a measurable signal by an excitation pulse at TI_2, and the signal decays at a rate of $R_2^*(t)$ (in 1/sec) until the readouts occur at TE_1 and TE_2. The predicted signal for echo time i and timepoint t is given by

$$p_{i,t} \quad = \quad S(t)\ \mathrm{e}^{-TE_i R_2^*(t)}. \tag{A.6}$$

The nonlinear signal model $\mathbf{g}$ is simply this predicted signal with the two echo times interleaved.

Furthermore, a particular parametric form was chosen for the time-varying quantities $R_2^*(t)$, $Q(t)$ and $M(t)$. Each is expressed as a fixed baseline plus an activation-related signal, where activation time-course shapes are found from the known stimulus timings and fixed Haemodynamic Response Functions (HRFs). The HRFs generated by Woolrich et al. (2006) were reused; these were derived empirically from data averaged across an ROI, and separate HRFs were found for each of the three quantities.

Appendix B

Detailed Derivations of Updates for Spatial Gaussian Process Priors

B.1 Estimating δ_k by Free-energy Maximization

In mean-field VB, the value of a parameter that maximizes the free energy F is found analytically for each factor. The combined prior does not allow the smoothness parameter to be factored out of the covariance matrix (equation 4.13), so there is no analytic solution for the optimal δ_k. It is still possible to find the F-maximizing value of δ_k numerically. This approach was tried first but suffered from extremely slow convergence in high-noise situations. Possible causes are discussed at the end of this appendix.

The free energy is calculated as the expected value of the log-posterior, minus the KL-divergence of the factorized posterior terms:

$$\begin{aligned} F &= \left\langle \log p(\mathbf{y}_*|\mathbf{w}_*, \boldsymbol{\phi}, \mathbf{a})\ p(\mathbf{w}_*|\mathbf{C}_*)\ p(\boldsymbol{\phi})\ p(\mathbf{a}) \right\rangle_{\prod_v (q(\mathbf{w}_v) q(\phi_v) q(a_v))} \\ &\quad - \mathrm{KL}(q(\mathbf{w}_*)) - \mathrm{KL}(\text{noise terms}) - \mathrm{KL}(\text{hyperparameters}) \end{aligned} \tag{B.1}$$

Most of these terms are constant. In particular, the distribution of $q(\mathbf{w}_*) = N(\mathbf{m}_*, \boldsymbol{\Sigma}_*)$ is taken as constant. This is in contrast to the EO method which integrates out the dependencies of $\mathbf{w}_*$ on the priors $\mathbf{C}_*$. The KL(hyperparameters) can also be

dropped only a point estimate of δ_k is needed. This leaves us with

$$F = \left\langle -\frac{1}{2}\left(\mathbf{w}_* - \boldsymbol{\mu}_*\right)^{\mathrm{T}} \mathbf{C}_*^{-1}\left(\mathbf{w}_* - \boldsymbol{\mu}_*\right) - \frac{1}{2}\log|\mathbf{C}_*| \right\rangle_{\prod_v q(\mathbf{w}_v)} \quad \text{(B.2)}$$

$$= -\frac{1}{2}\left(\mathbf{m}_* - \boldsymbol{\mu}_*\right)^{\mathrm{T}} \mathbf{C}_*^{-1}\left(\mathbf{m}_* - \boldsymbol{\mu}_*\right) - \frac{1}{2}\mathrm{Tr}\left[\mathbf{C}_*^{-1}\boldsymbol{\Sigma}_*\right] - \frac{1}{2}\log|\mathbf{C}_*| \quad \text{(B.3)}$$

Note that in this case, $\boldsymbol{\Sigma}_*$ has been reassembled from the individual $\boldsymbol{\Sigma}_v$ estimates, and therefore lacks the inter-voxel correlation. Taking the derivative of this with respect to δ_k (again, remembering that $\boldsymbol{\Sigma}_*$ and $\mathbf{m}_*$ are constant) leads to

$$\frac{\partial \log F}{\partial \delta_k} = \frac{1}{2}\mathrm{Tr}\left[(\mathbf{C}_* - \boldsymbol{\Sigma}_* - (\mathbf{m}_* - \boldsymbol{\mu}_*)(\mathbf{m}_* - \boldsymbol{\mu}_*)^{\mathrm{T}})\frac{\partial(\mathbf{C}_*^{-1})}{\partial \delta_k}\right]. \quad \text{(B.4)}$$

which is an identical expression to equation 4.15, but reached by different means. The difference in practice is due to the fact that here $\boldsymbol{\Sigma}_*$ is treated as a constant because it describes a different factor, $\mathrm{P}'(\mathbf{w})$. This can lead to extremely slow convergence. In the preferred EO-based method, $\boldsymbol{\Sigma}_*$ is a function of the data $\mathbf{y}_*$ and $\mathbf{C}_*$, since the signal parameters $\mathbf{w}$ have been integrated out. In both inference methods, seeking a descending zero-crossing of this function yields a (potentially local) maximum of F.

In low-noise situations, inference behaves well and produces estimates that are reasonably similar to the EO results. The ROC curve in figure 4.2(b) illustrates this (the free energy maximization is called GPP+VB). Note that the discrimination is not quite as good as when using EO. This may a consequence of factorizing $\prod_v q(\mathbf{w}_v)$ rather than keeping $q(\mathbf{w}_*)$. The loss of the covariance terms would tend to mean a lower smoothness is inferred. There is also an issue of slower convergence due to each $q(\mathbf{w}_v)$ being held back by old values for its neighbours rather than all of the parameter values being inferred simultaneously.

However, in high-noise situations the F-maximization algorithm does not consistently converge. It becomes very sensitive to initial conditions – in particular, the initial spatial smoothness of $\mathbf{w}_*$. This is not a simple problem of local minima in $F(\delta_k)$; starting the search from a different value of δ has not effect on the outcome.

It is not entirely clear why this happens and why it does not appear to affect the Laplacian spatial prior. The most plausible explanation is that the individual $q(\mathbf{w}_v)$ terms become stuck on the values of their neighbours. Conditional on the mean fields of all other voxels, the VB update on $q(\mathbf{w}_v)$ gets a large fraction of its precision from the prior: 97% in this problematic case and 83% in the working case, while in the case of the Laplacian spatial prior this was around 80%. This work has not yet found an initialization method to reliably avoid this situation. However, this is only a problem if the values must step individually rather than moving simultaneously; a fully-VB method might not suffer from this problem if there were no factorization of $\mathbf{w}_*$ between voxels voxels. Another approach to consider would be to modify the prior to allow more freedom in the conditional, by adding additional variation on the diagonal; for example $\mathbf{C}_k = \sigma^2(\lambda \exp(-\boldsymbol{\Delta}/\delta) + (1-\lambda)\mathbf{I})$ with $\lambda \sim \mathbf{U}[0,1]$.

However, there would be little computation advantage in switching from the hybrid VB-EO approach to one of these full-VB approaches, with or without factorization over voxels; all are limited by $O(V^3)$ cost of trying a single δ_k value in the maximization step.

B.2 Details of EO Derivation

The conditional evidence, $\mathscr{E}(\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a}) = P(\mathbf{y}_*|\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a})$, is found by integrating out $\mathbf{w}_*$ from the model (equation 4.8):

$$\mathscr{E}(\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a}) = \int P(\mathbf{y}_*, \mathbf{w}_*|\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a}) d\mathbf{w}_* = \int P(\mathbf{y}_*|\mathbf{w}_*, \boldsymbol{\phi}, \mathbf{a}) P(\mathbf{w}_*|\mathbf{C}_*) d\mathbf{w}_* \quad \text{(B.5)}$$

$$= \int N(\mathbf{R}_*\mathbf{y}_*; \mathbf{R}_*\mathbf{g}(\mathbf{w}_*), \boldsymbol{\Phi}_*^{-1})\ N(\mathbf{w}_*; \boldsymbol{\mu}_*, \mathbf{C}_*) d\mathbf{w}_* \quad \text{(B.6)}$$

$$= \int \frac{\mathrm{e}^{-\frac{1}{2}\left[(\mathbf{y}_* - \mathbf{g}(\mathbf{w}_*))^{\mathrm{T}} \mathbf{R}_*^{\mathrm{T}} \boldsymbol{\Phi}_* \mathbf{R}_* (\mathbf{y}_* - \mathbf{g}(\mathbf{w}_*)) + (\mathbf{w}_* - \boldsymbol{\mu}_*)^{\mathrm{T}} \mathbf{C}_*^{-1} (\mathbf{w}_* - \boldsymbol{\mu}_*)\right]}}{|2\pi \boldsymbol{\Phi}_*^{-1}|^{-1/2} \cdot |2\pi \mathbf{C}_*|^{-1/2}} d\mathbf{w}_* \quad \text{(B.7)}$$

Substituting in the approximate forward model $\overline{\mathbf{g}}$ from equation 3.3, the term in the exponential can be rewritten as:

$$-\frac{1}{2}\Bigg[(\mathbf{w}_* - \boldsymbol{\mu}_*)^{\mathrm{T}} \left(\mathbf{J}_*^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*\mathbf{J}_* + \mathbf{C}_*^{-1}\right)(\mathbf{w}_* - \boldsymbol{\mu}_*)$$

$$-2(\mathbf{w}_* - \boldsymbol{\mu}_*)^{\mathrm{T}}\mathbf{J}_*^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*(\mathbf{y}_* - \overline{\mathbf{g}}(\boldsymbol{\mu}_*)) + (\mathbf{y}_* - \overline{\mathbf{g}}(\boldsymbol{\mu}_*))^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*(\mathbf{y}_* - \overline{\mathbf{g}}(\boldsymbol{\mu}_*))\Bigg]$$

This implies a normal distribution on the forward model parameters $\mathbf{w}_* \sim N(\mathbf{m}_*, \boldsymbol{\Sigma}_*)$, where the covariance $\boldsymbol{\Sigma}_*$ and mean $\mathbf{m}_*$ are defined by

$$\boldsymbol{\Sigma}_*^{-1} = \mathbf{J}_*^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*\mathbf{J}_* + \mathbf{C}_*^{-1} \tag{B.8}$$

$$\boldsymbol{\Sigma}_*^{-1}(\mathbf{m}_* - \boldsymbol{\mu}_*) = \mathbf{J}_*^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*(\mathbf{y}_* - \overline{\mathbf{g}}(\boldsymbol{\mu}_*)). \tag{B.9}$$

Note that both $\boldsymbol{\Sigma}_*$ and $\mathbf{m}_*$ depend on the hyperparameters of $\mathbf{C}_*$ (but also that $\boldsymbol{\Sigma}_*^{-1}(\mathbf{m}_* - \boldsymbol{\mu}_*)$ is a constant with respect to $\mathbf{C}_*$). The evidence expression is now given by

$$\begin{aligned}\mathscr{E}(\mathbf{C}_*, \boldsymbol{\phi}, \mathbf{a}) &= \left(\int N(\mathbf{w}_*; \mathbf{m}_*, \boldsymbol{\Sigma}_*)d\mathbf{w}_*\right)\sqrt{\frac{|2\pi\boldsymbol{\Sigma}_*|}{|2\pi\mathbf{C}_*||2\pi\boldsymbol{\Phi}_*^{-1}|}} \\ &\quad \cdot \exp\Big[-\frac{1}{2}(\mathbf{y}_* - \overline{\mathbf{g}}(\boldsymbol{\mu}_*))^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*(\mathbf{y}_* - \overline{\mathbf{g}}(\boldsymbol{\mu}_*)) \\ &\quad +\frac{1}{2}(\mathbf{m}_* - \boldsymbol{\mu}_*)^{\mathrm{T}}\boldsymbol{\Sigma}_*^{-1}((\mathbf{m}_* - \boldsymbol{\mu}_*)\Big] \qquad \text{(B.10)} \\ &\propto \sqrt{\frac{|\boldsymbol{\Sigma}_*|}{|\mathbf{C}_*|}} \cdot \exp\left[+\frac{1}{2}(\mathbf{m}_* - \boldsymbol{\mu}_*)^{\mathrm{T}}\boldsymbol{\Sigma}_*^{-1}(\mathbf{m}_* - \boldsymbol{\mu}_*)\right]. \qquad \text{(B.11)}\end{aligned}$$

Therefore,

$$\log\mathscr{E} = \frac{1}{2}|\boldsymbol{\Sigma}_*| - \frac{1}{2}|\mathbf{C}_*| + \frac{1}{2}(\mathbf{m}_* - \boldsymbol{\mu}_*)^{\mathrm{T}}\boldsymbol{\Sigma}_*^{-1}(\mathbf{m}_* - \boldsymbol{\mu}_*) + \text{consts.} \tag{B.12}$$

Next, take the derivative of $\log\mathscr{E}$ with respect to the GPP parameter δ_k. Considering each term separately,

$$\begin{aligned}\frac{\partial}{\partial\delta_k}\log|\boldsymbol{\Sigma}_*| &= -\frac{\partial}{\partial\delta_k}\log|\boldsymbol{\Sigma}_*^{-1}| = \frac{-1}{|\boldsymbol{\Sigma}_*^{-1}|}\frac{\partial|\boldsymbol{\Sigma}_*^{-1}|}{\partial\delta_k} \\ &= -\mathrm{Tr}\left[\boldsymbol{\Sigma}_*\frac{\partial(\boldsymbol{\Sigma}_*^{-1})}{\partial\delta_k}\right] = -\mathrm{Tr}\left[\boldsymbol{\Sigma}_*\frac{\partial(\mathbf{C}_*^{-1})}{\partial\delta_k}\right] \qquad \text{(B.13)}\end{aligned}$$

Similarly,

$$\frac{\partial}{\partial \delta_k} \log|\mathbf{C}_*| = -\mathrm{Tr}\left[\mathbf{C}_* \frac{\partial(\mathbf{C}_*^{-1})}{\partial \delta_k}\right] \tag{B.14}$$

and recalling that $\mathbf{\Sigma}_*^{-1}(\mathbf{m}_* - \boldsymbol{\mu}_*)$ is a constant with respect to δ_k,

$$\begin{aligned}
\frac{\partial}{\partial \delta_k} \left[((\mathbf{m}_* - \boldsymbol{\mu}*))^{\mathrm{T}} \mathbf{\Sigma}_*^{-1} \mathbf{\Sigma}_* \mathbf{\Sigma}_*^{-1} ((\mathbf{m}_* - \boldsymbol{\mu}*))\right] & \\
&= ((\mathbf{m}_* - \boldsymbol{\mu}*))^{\mathrm{T}} \mathbf{\Sigma}_*^{-1} \frac{\partial \mathbf{\Sigma}_*}{\partial \delta_k} \mathbf{\Sigma}_*^{-1} ((\mathbf{m}_* - \boldsymbol{\mu}*)) \\
&= ((\mathbf{m}_* - \boldsymbol{\mu}*))^{\mathrm{T}} \mathbf{\Sigma}_*^{-1} \left[-\mathbf{\Sigma}_* \frac{\partial \mathbf{\Sigma}_*^{-1}}{\partial \delta_k} \mathbf{\Sigma}_*\right] \mathbf{\Sigma}_*^{-1} ((\mathbf{m}_* - \boldsymbol{\mu}*)) \\
&= -((\mathbf{m}_* - \boldsymbol{\mu}*))^{\mathrm{T}} \frac{\partial \mathbf{\Sigma}_*^{-1}}{\partial \delta_k} ((\mathbf{m}_* - \boldsymbol{\mu}*)) .
\end{aligned} \tag{B.15}$$

These yield

$$\frac{\partial \log \mathscr{E}}{\partial \delta_k} = \frac{1}{2}\mathrm{Tr}\left[(\mathbf{C}_* - \mathbf{\Sigma}_* - (\mathbf{m}_* - \boldsymbol{\mu}_*)(\mathbf{m}_* - \boldsymbol{\mu}_*)^{\mathrm{T}}) \frac{\partial(\mathbf{C}_*^{-1})}{\partial \delta_k}\right]. \tag{B.16}$$

Also note that $\mathbf{C}_*$ is assembled from individual $\mathbf{C}_j$ matrices. From equation 4.13, it follows that the derivative $\partial \mathbf{C}_k / \partial \delta_k = (-\mathbf{\Delta}/\delta_k^2) \circ \mathbf{C}_k$, so $\partial(\mathbf{C}_k^{-1})/\partial \delta_k = \mathbf{C}_k^{-1}(\mathbf{C}_k \circ \mathbf{\Delta}/\delta_k^2)\mathbf{C}_k^{-1}$, where $\circ$ is the Hadamard (elementwise) product. For all other $\mathbf{C}_j$ $(j \neq k)$ the derivative is zero.

A simple bisection algorithm is used to find a descending zero-crossing of this expression; this is the value of δ_k that maximizes the conditional evidence. Each EO "step" involves many evaluations of this expression; in practice, a coarse estimate is used in early iterations and more accurate estimate are required in later iterations as the model converges.

Even with $K = 2$ is large, factorizing the EO step over parameters considerably speeds up the calculation by reducing the $KV \times KV$ matrix inversion a single $V \times V$ inversion. This approach is described here.

First, define a set of $V \times KV$ matrix $\mathbf{H}_k$, which extracts only those elements relating to parameter k. In particular, the i, j^{th} element is 1 if $j = k + (i-1)K$ and 0 otherwise. Then isolate individual parameter matrices: $\mathbf{C}_k = \mathbf{H}_k \mathbf{C}_* \mathbf{H}_k^{\mathrm{T}}$, and $\mathbf{C}_* = \sum_k \mathbf{H}_k^{\mathrm{T}} \mathbf{C}_k \mathbf{H}_k$.

Note that the global prior covariance matrix $\mathbf{C}_*$ is already factorized over parameters, and δ_k only appears in rows and columns relating to parameter k. Therefore $\partial\mathbf{C}_*/\partial\delta_k = \mathbf{H}_k^{\mathrm{T}}\left(\partial\mathbf{C}_k/\partial\delta_k\right)\mathbf{H}_k$ and $\partial(\mathbf{C}_*^{-1})/\partial\delta_k = \mathbf{H}_k^{\mathrm{T}}\left(\partial(\mathbf{C}_k^{-1})/\partial\delta_k\right)\mathbf{H}_k$.

Continuing from equation 4.15 to obtain

$$
\begin{aligned}
0 = \frac{\partial \log\varepsilon}{\partial\delta_k} &= \frac{1}{2}\mathrm{Tr}\left[(\mathbf{C}_* - \boldsymbol{\Sigma}_* - (\mathbf{m}_* - \boldsymbol{\mu}_*)(\mathbf{m}_* - \boldsymbol{\mu}_*)^{\mathrm{T}})\mathbf{H}_k^{\mathrm{T}}\frac{\partial(\mathbf{C}_k^{-1})}{\partial\delta_k}\mathbf{H}_k\right] && \text{(B.17)}\\
&= \frac{1}{2}\mathrm{Tr}\left[\mathbf{H}_k(\mathbf{C}_* - \boldsymbol{\Sigma}_* - (\mathbf{m}_* - \boldsymbol{\mu}_*)(\mathbf{m}_* - \boldsymbol{\mu}_*)^{\mathrm{T}})\mathbf{H}_k^{\mathrm{T}}\frac{\partial(\mathbf{C}_k^{-1})}{\partial\delta_k}\right] && \text{(B.18)}\\
&= \frac{1}{2}\mathrm{Tr}\left[\left(\mathbf{C}_k - (\mathbf{m}_k - \boldsymbol{\mu}_k)(\mathbf{m}_k - \boldsymbol{\mu}_k)^{\mathrm{T}} - \mathbf{H}_k\boldsymbol{\Sigma}_*\mathbf{H}_k^{\mathrm{T}}\right)\frac{\partial(\mathbf{C}_k^{-1})}{\partial\delta_k}\right] && \text{(B.19)}
\end{aligned}
$$

leaving only the troublesome $KV \times KV$ term

$$
\mathbf{H}_k\boldsymbol{\Sigma}_*\mathbf{H}_k^{\mathrm{T}} = \mathbf{H}_k(\mathbf{C}_*^{-1} + \mathbf{J}_*^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*\mathbf{J}_*)^{-1}\mathbf{H}_k^{\mathrm{T}}. \qquad \text{(B.20)}
$$

which does not factor over parameters cleanly.

One option is simply to approximates the necessary block of $\boldsymbol{\Sigma}_*$ by the inverse of the same block of $\boldsymbol{\Sigma}_*^{-1}$. This approximation ignores the correlation between signal parameters, which is unlikely to change the estimated δ_k value noticeably because inter-voxel correlations are far more relevant. This is the approach taken in the current implementation.

However, there is an exact method using the Woodbury matrix identity (Woodbury, 1950). First, define

$$
\mathbf{A} = \mathbf{C}_{*-k} + \mathbf{J}_*^{\mathrm{T}}\mathbf{R}_*^{\mathrm{T}}\boldsymbol{\Phi}_*\mathbf{R}_*\mathbf{J}_* \qquad \text{(B.21)}
$$

where $\mathbf{C}_{*-k}$ is $\mathbf{C}_*$ with the k^{th} block zeroed out. The $KN \times KN$ matrix $\mathbf{A}$ is therefore constant with respect to δ_k, so precalculate $\mathbf{A}^{-1}$ only once per iteration. Now, apply the identity:

$$
\begin{aligned}
\boldsymbol{\Sigma}_* &= (\mathbf{H}_k^{\mathrm{T}}\mathbf{C}_k\mathbf{H} + \mathbf{A})^{-1}\\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{H}_k^{\mathrm{T}}(\mathbf{C}_k + \mathbf{H}_k\mathbf{A}^{-1}\mathbf{H}_k^{\mathrm{T}})^{-1}\mathbf{H}_k\mathbf{A}^{-1} && \text{(B.22)}
\end{aligned}
$$

so

$$\begin{aligned}\mathbf{H}_k\boldsymbol{\Sigma}_*\mathbf{H}_k^{\mathrm{T}} &= (\mathbf{H}_k\mathbf{A}^{-1}\mathbf{H}_k^{\mathrm{T}}) \\ &\quad -(\mathbf{H}_k\mathbf{A}^{-1}\mathbf{H}_k^{\mathrm{T}})(\mathbf{C}_k+(\mathbf{H}_k\mathbf{A}^{-1}\mathbf{H}_k^{\mathrm{T}}))^{-1}(\mathbf{H}_k\mathbf{A}^{-1}\mathbf{H}_k^{\mathrm{T}}) \qquad \text{(B.23)}\end{aligned}$$

where all of the terms in parentheses are $K \times K$.

Appendix C

Integrated Decomposition and Decoding Calculations

C.1 VB Updates

PCA timecourse update: W

The posterior distribution on $\mathbf{W}$ is a matrix normal distribution.

$$\mathrm{P}'(\mathbf{W}) \quad = \quad \mathrm{MN}(\mathbf{W}|\mathbf{M}_{\mathbf{W}}, \mathbf{\Omega}_{\mathbf{W}}, \mathbf{\Sigma}_{\mathbf{W}}) \tag{C.1}$$

The matrix normal on the $T \times L$ matrix $\mathbf{W}$ is defined in terms of a $T \times L$ mean $\mathbf{M}_{\mathbf{W}}$, an $L \times L$ row covariance matrix $\mathbf{\Omega}_{\mathbf{W}}$ and a $T \times T$ column covariance matrix $\mathbf{\Sigma}_{\mathbf{W}}$. This is a generalization of the multivariate normal to matrices, defined by

$$\mathrm{MN}(\mathbf{W}|\mathbf{M}, \mathbf{\Omega}, \mathbf{\Sigma}) \quad = \quad \frac{\exp\left(-\frac{1}{2}\mathrm{Tr}\left[\mathbf{\Omega}^{-1}\left(\mathbf{W}-\mathbf{M}\right)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\left(\mathbf{W}-\mathbf{M}\right)\right]\right)}{|\mathbf{\Omega}|^{T/2}|\mathbf{\Sigma}|^{L/2}(2\pi)^{TL/2}} \tag{C.2}$$

which can be defined in terms of the multivariate normal itself:

$$\mathrm{MN}(\mathbf{W}|\mathbf{M}, \mathbf{\Omega}, \mathbf{\Sigma}) \quad = \quad \mathrm{N}(\mathrm{vec}\ \mathbf{W}|\mathrm{vec}\ \mathbf{M}, \mathbf{\Omega} \otimes \mathbf{\Sigma}) \tag{C.3}$$

where $\otimes$ denotes the Kronecker product and vec$\mathbf{W}$ is the vectorization of $\mathbf{W}$ given by stacking its columns. Note that the matrix normal can only represent a limited subset of possible covariance matrices because the combined matrix must be able to be decomposed into this Kronecker-product form.

The updates are given by

$$\begin{aligned}
\boldsymbol{\Sigma}_W &= \mathbf{I}_T \qquad \text{(because all timepoints are equivalent)} && \text{(C.4)}\\
{\boldsymbol{\Omega}_W}^{-1} &= \left\langle \mathbf{H} \left\langle \operatorname{diag} \boldsymbol{\lambda} \right\rangle \mathbf{H}^{\mathrm{T}} \right\rangle + \operatorname{diag} \langle \boldsymbol{\omega} \rangle && \text{(C.5)}\\
\mathbf{M_W}\boldsymbol{\Omega}^{-1} &= \mathbf{Y} \left\langle \operatorname{diag} \boldsymbol{\lambda} \right\rangle \langle \mathbf{H} \rangle^{\mathrm{T}} && \text{(C.6)}
\end{aligned}$$

The matrix normal has some convenient properties in finding expectations involving it. In particular, for $\mathbf{W} \sim \mathrm{MN}(\mathbf{M_W}, \boldsymbol{\Omega_W}, \boldsymbol{\Sigma_W})$, some of the moments are given by

$$\begin{aligned}
\langle \mathbf{W} \rangle &= \mathbf{M_W} && \text{(C.7)}\\
\left\langle \mathbf{W}\mathbf{C}\mathbf{W}^{\mathrm{T}} \right\rangle &= \mathbf{M_W}\mathbf{C}{\mathbf{M_W}}^{\mathrm{T}} + \boldsymbol{\Sigma_W} \mathrm{Tr}\left[\boldsymbol{\Omega_W}\mathbf{C}\right] && \text{(C.8)}\\
\left\langle \mathbf{W}^{\mathrm{T}}\mathbf{C}\mathbf{W} \right\rangle &= {\mathbf{M_W}}^{\mathrm{T}}\mathbf{C}\mathbf{M_W} + \boldsymbol{\Omega_W} \mathrm{Tr}\left[\boldsymbol{\Sigma_W}\mathbf{C}\right] && \text{(C.9)}
\end{aligned}$$

where $\mathbf{C}$ denotes any arbitrary fixed matrix of the appropriate size for each equation.

Regression model update: A

The posterior on the $B \times L$ regression matrix $\mathbf{A}$ is somewhat more complicated because the correlation structure of the noise model and the prior cannot be summarized in a matrix normal distribution.

$$\mathrm{P}'(\mathbf{A}) = \mathrm{N}(\mathrm{vec}\ \mathbf{A} | \mathrm{vec}\ \mathbf{M_A}, \boldsymbol{\Sigma_A}) \qquad \text{(C.10)}$$

where $\boldsymbol{\Sigma_A} \in \mathscr{R}^{BL \times BL}$.

The reason for this is evident when examining the updates:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{A}}^{-1} &= \operatorname{diag}\left(\mathrm{vec}\ \langle \boldsymbol{\alpha} \rangle\right) + \left\langle \mathbf{H}\mathbf{H}^{\mathrm{T}} \right\rangle \otimes \boldsymbol{\Gamma} && \text{(C.11)}\\
\mathbf{M_A}\boldsymbol{\Sigma}_{\mathbf{A}}^{-1} &= 0 + \boldsymbol{\Gamma}\mathbf{V} \langle \mathbf{H} \rangle^{\mathrm{T}} && \text{(C.12)}
\end{aligned}$$

This full expansion is necessary in because of the ARD prior on each element of $\mathbf{A}$. An alternative to this arrangement is to place the ARD only on each column of $\mathbf{A}$

(considered in equation 5.18), which would produce a precision term of $\mathrm{diag}\left(\langle\boldsymbol{\alpha}\rangle\right) \otimes \mathbf{I}_B$. However, these would still have different row precisions and different column precisions, so it would still not be possible to condense the two Kronecker products into a single one; this requires that either the row precisions be equal or the column precisions be equal (up to a constant).

Since the $B \times L$ matrix $\mathbf{A}$ does not have a matrix normal form the needed expectations are somewhat more complicated:

$$\langle \mathbf{A} \rangle = \mathbf{M}_{\mathbf{A}} \tag{C.13}$$

$$\left\langle \mathbf{A}\mathbf{C}\mathbf{A}^{\mathrm{T}} \right\rangle = \mathbf{M}_{\mathbf{A}}\mathbf{C}{\mathbf{M}_{\mathbf{A}}}^{\mathrm{T}} + \sum_{j=1}^{L}\sum_{k=1}^{L} \mathbf{C}\left(\boldsymbol{\Sigma}_{\mathbf{A}}\right)_{(jB+1:jB+B),(kB+1:kB+B)} \tag{C.14}$$

$$\tag{C.15}$$

where the $\left(\boldsymbol{\Sigma}_{\mathbf{A}}\right)_{(jB+1:jB+B),(kB+1:kB+B)}$ notation picks out the separate blocks within the full covariance matrix, giving the covariance structure between $\mathbf{A}_{\cdot,j}$ and $\mathbf{A}_{\cdot,k}$. The summations given above are a generalization of equation C.8 that works when for general covariance matrices that cannot be separated into separate row and column covariances.

ARD sparsity hyperparameter updates

The priors on $\boldsymbol{\Sigma}_{\mathbf{W}}$ are given by the ARD hyperparameters $\boldsymbol{\omega}$:

$$\mathrm{P}'(\boldsymbol{\omega}) = \prod_{i=1}^{L} \mathrm{Ga}(\boldsymbol{\omega}_i | b_i, c_i) \tag{C.16}$$

$$c_i = c_0 + T/2 \tag{C.17}$$

$$b_i^{-1} = \left\langle \mathbf{W}_{\cdot,i}^{\mathrm{T}} \mathbf{W}_{\cdot,i} \right\rangle /2 + b_0^{-1} \tag{C.18}$$

Similarly, the prior on each element $\mathbf{A}_{b,i}$ is given by:

$$\begin{aligned}
\mathrm{P}(\mathbf{A}) &= \prod_{b=1}^{B}\prod_{i=1}^{L} \mathrm{N}\left(\mathbf{A}_{b,i}|0, \boldsymbol{\alpha}_{b,i}^{-1}\right) && \text{(C.19)}\\
\mathrm{P}'(\boldsymbol{\alpha}) &= \prod_{b=1}^{B}\prod_{i=1}^{L} \mathrm{Ga}(\alpha_{b,i}|b_{\alpha bi}, c_{\alpha bi}) && \text{(C.20)}\\
c_{\alpha bi} &= 1/2 + c_{\alpha 0} && \text{(C.21)}\\
b_{\alpha bi}^{-1} &= \left\langle \mathbf{A}_{b,i}^2 \right\rangle + b_{\alpha 0}^{-1}. && \text{(C.22)}
\end{aligned}$$

Noise updates: $\boldsymbol{\lambda}_r$ and $\boldsymbol{\Gamma}$

For the simulated data, a single noise precision is used for all trials so $\boldsymbol{\Lambda} = \lambda\mathbf{I}$ with:

$$\begin{aligned}
\mathrm{P}'(\lambda) &= \mathrm{Ga}(\lambda|b_\lambda, c_\lambda) && \text{(C.23)}\\
c_\lambda &= c_{\lambda 0} + RT/2 && \text{(C.24)}\\
b_\lambda^{-1} &= b_{\lambda 0}^{-1} + \frac{1}{2}\mathrm{Tr}\left[\left\langle (\mathbf{Y} - \mathbf{XH})^{\mathrm{T}} (\mathbf{Y} - \mathbf{XH}) \right\rangle\right] \\
&= b_{\lambda 0}^{-1} + \frac{1}{2}\mathrm{Tr}\left[\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right] - \mathrm{Tr}\left[\mathbf{Y}^{\mathrm{T}} \langle \mathbf{XH} \rangle\right] + \frac{1}{2}\mathrm{Tr}\left[\left\langle \mathbf{H}^{\mathrm{T}} \left\langle \mathbf{X}^{\mathrm{T}}\mathbf{X} \right\rangle \mathbf{H} \right\rangle\right] && \text{(C.25)}
\end{aligned}$$

The noise on the neuroimaging data $\mathbf{Y}$ is different for each repeat but the same for each timepoint. The posterior noise precisions $\boldsymbol{\lambda} = \mathrm{diag}\left[\lambda_1 \ldots \lambda_R\right]$ are given by splitting up the elements of the Trace functions above:

$$\begin{aligned}
\mathrm{P}'(\boldsymbol{\lambda}_r) &= \mathrm{Ga}(\boldsymbol{\lambda}_r|b_{\lambda r}, c_{\lambda r}) && \text{(C.26)}\\
c_{\lambda r} &= c_{\lambda 0} + T/2 && \text{(C.27)}\\
b_{\lambda r}^{-1} &= b_{\lambda 0}^{-1} + \frac{1}{2}\left[\left\langle (\mathbf{Y} - \mathbf{XH})^{\mathrm{T}} (\mathbf{Y} - \mathbf{XH}) \right\rangle\right]_{rr} \\
&= b_{\lambda 0}^{-1} + \frac{1}{2}\left[\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right]_{rr} - \left[\mathbf{Y}^{\mathrm{T}} \langle \mathbf{XH} \rangle\right]_{rr} + \frac{1}{2}\left[\left\langle \mathbf{H}^{\mathrm{T}} \left\langle \mathbf{X}^{\mathrm{T}}\mathbf{X} \right\rangle \mathbf{H} \right\rangle\right]_{rr} && \text{(C.28)}
\end{aligned}$$

For the behavioural noise the updates appear similar but without taking the trace at all, resulting in a Wishart distribution. Note that instead summarizing across timepoints as above to get a repeat-by-repeat noise estimate, with the behavioural

noise the information is summarized across repeats to produce a $B \times B$ noise matrix.

$$\begin{aligned} \mathrm{P}'(\boldsymbol{\Gamma}) &= \mathrm{Wishart}(\boldsymbol{\Gamma}|\mathbf{Z}_{\boldsymbol{\Gamma}}, n_{\boldsymbol{\Gamma}}) && \text{(C.29)} \\ n_{\boldsymbol{\Gamma}} &= n_{\boldsymbol{\Gamma}0} + R && \text{(C.30)} \\ \mathbf{Z}_{\boldsymbol{\Gamma}} &= \mathbf{Z}_{\boldsymbol{\Gamma}0} + \left\langle (\mathbf{V} - \mathbf{AB})(\mathbf{V} - \mathbf{AB})^{\mathrm{T}} \right\rangle \\ &= \mathbf{Z}_{\boldsymbol{\Gamma}0} + \mathbf{V}\mathbf{V}^{\mathrm{T}} - \langle \mathbf{A} \rangle \langle \mathbf{B} \rangle \mathbf{V}^{\mathrm{T}} - \mathbf{V} \langle \mathbf{B} \rangle^{\mathrm{T}} \langle \mathbf{A} \rangle^{\mathrm{T}} + \left\langle \mathbf{A} \left\langle \mathbf{B}^{\mathrm{T}}\mathbf{B} \right\rangle \mathbf{A}^{\mathrm{T}} \right\rangle && \text{(C.31)} \end{aligned}$$

where the Wishart is an extension of the Gamma distribution to the space of arbitrary positive-definite matrices. The necessary moments for the other VB updates and the free energy are given by

$$\begin{aligned} \langle \boldsymbol{\Gamma} \rangle &= \mathbf{Z}_{\boldsymbol{\Gamma}} n_{\boldsymbol{\Gamma}} && \text{(C.32)} \\ \langle \log|\boldsymbol{\Gamma}| \rangle &= \log|\mathbf{Z}_{\boldsymbol{\Gamma}}| + \psi\left(n_{\boldsymbol{\Gamma}}\right) && \text{(C.33)} \end{aligned}$$

Mean updates: $\boldsymbol{\mu}^{(Y)}$ and $\boldsymbol{\mu}^{(V)}$

For simplicity of implementation, the means are implemented by concatenating an additional trial-course of all ones to $\mathbf{H}$. Instead of having a $\mathrm{N}(0, 1)$ prior, elements in the first row are fixed at 1 (which is achieved in practice by using a very strong $\mathrm{N}(1, 10^{-12})$ prior on these elements). This leads the first column of $\mathbf{W}$ and $\mathbf{A}$ representing the means $\boldsymbol{\mu}^{(Y)}$ and $\boldsymbol{\mu}^{(V)}$ respectively. To allow unbiased estimation of the mean, these first columns have an uninformative $\mathrm{N}(0, 10^{12})$ prior instead of an ARD.

Hidden state updates: H

The hidden state matrix is where the data fusion occurs, since $\mathbf{H}$ is informed by both data neuroimaging data and behavioural data (as well as a strong prior). There is no correlation between repeats to it naturally factorizes into $\mathrm{P}'(\mathbf{H}) = \prod_{r=1}^{R} \mathrm{P}'(\mathbf{H}_{\cdot,r})$,

each given by:

$$\begin{aligned} \mathrm{P}'(\mathbf{H}_{\cdot,r}) &= \mathrm{N}(\mathbf{H}_{\cdot,r}|\mathbf{M}_{\mathbf{H}r}, \mathbf{\Sigma}_{\mathbf{H}r}) && \text{(C.34)} \\ \mathbf{\Sigma}_{\mathbf{H}r}^{-1} &= \mathbf{\Sigma}_{\mathbf{H}0}^{-1} + \langle \boldsymbol{\lambda}_r \rangle \left\langle \mathbf{W}^{\mathrm{T}}\mathbf{W} \right\rangle + \left\langle \mathbf{A}^{\mathrm{T}} \langle \mathbf{\Gamma} \rangle \mathbf{A} \right\rangle && \text{(C.35)} \\ \mathbf{\Sigma}_{\mathbf{H}r}^{-1}\mathbf{M}_{\mathbf{H}r} &= \mathbf{\Sigma}_{\mathbf{H}0}^{-1}\mathbf{M}_{\mathbf{H}0} + \langle \boldsymbol{\lambda}_r \rangle \langle \mathbf{W} \rangle^{\mathrm{T}} \mathbf{Y}_{\cdot,r} + \langle \mathbf{A} \rangle^{\mathrm{T}} \langle \mathbf{\Gamma} \rangle \mathbf{V}_{\cdot,r} && \text{(C.36)} \end{aligned}$$

The priors are given by $\mathbf{\Sigma}_{\mathbf{H}0} = \mathbf{I}$ and $\mathbf{M}_{\mathbf{H}0} = \mathbf{0}$, but if the means are included as the first columns of $\mathbf{W}$ and $\mathbf{A}$ then $(\mathbf{M}_{\mathbf{H}0})_1 = 1$ and $(\mathbf{\Sigma}_{\mathbf{H}0})_{1,1} = 10^{-12}$, effectively defining the first component as being 1 in all repeats.

Note that in the absence of a trial-by-trial noise model, this calculation is simplified because the same posterior covariance $\mathbf{\Sigma}_{\mathbf{H}}$ can be used for all R trials.

C.2 Equivalent Linear Decoding Matrix

Starting from the joint distribution of the neuroimaging and behavioural data (equation 5.21)

$$\begin{bmatrix} \mathbf{y}_r \\ \mathbf{v}_r \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}^{(\mathbf{Y})} \\ \boldsymbol{\mu}^{(\mathbf{V})} \end{bmatrix}, \begin{bmatrix} \mathbf{W}\mathbf{W}^{\mathrm{T}} + \lambda_r^{-1}\mathbf{I} & \mathbf{W}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\mathbf{W}^{\mathrm{T}} & \mathbf{A}\mathbf{A}^{\mathrm{T}} + \mathbf{\Gamma}^{-1} \end{bmatrix} \right) \qquad \text{(C.37)}$$

the probability distribution can be written as

$$\begin{aligned} &\mathrm{P}\big(\mathbf{y}_r, \mathbf{v}_r | \mathbf{W}, \mathbf{A}, \lambda_r, \mathbf{\Gamma}, \boldsymbol{\mu}^{(Y)}, \boldsymbol{\mu}^{(V)}\big) \\ &\quad \propto -\frac{1}{2} \begin{bmatrix} \mathbf{y}_r - \boldsymbol{\mu}^{(Y)} \\ \mathbf{v}_r - \boldsymbol{\mu}^{(V)} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{W}\mathbf{W}^{\mathrm{T}} + \lambda_r^{-1}\mathbf{I} & \mathbf{W}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\mathbf{W}^{\mathrm{T}} & \mathbf{A}\mathbf{A}^{\mathrm{T}} + \mathbf{\Gamma}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_r - \boldsymbol{\mu}^{(Y)} \\ \mathbf{v}_r - \boldsymbol{\mu}^{(V)} \end{bmatrix} \\ &\quad \propto -\frac{1}{2} \begin{bmatrix} \mathbf{y}_r - \boldsymbol{\mu}^{(Y)} \\ \mathbf{v}_r - \boldsymbol{\mu}^{(V)} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{P_1} & \mathbf{P_2} \\ \mathbf{P_2}^{\mathrm{T}} & \mathbf{P_3} \end{bmatrix} \begin{bmatrix} \mathbf{y}_r - \boldsymbol{\mu}^{(Y)} \\ \mathbf{v}_r - \boldsymbol{\mu}^{(V)} \end{bmatrix} \end{aligned} \qquad \text{(C.38)}$$

where precision matrix $\begin{bmatrix} \mathbf{P_1} & \mathbf{P_2} \\ \mathbf{P_2}^{\mathrm{T}} & \mathbf{P_3} \end{bmatrix}$ is the inverse of the covariance in equation 5.21, found by the blockwise matrix inversion identity for any $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$:

$$\begin{aligned} \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} &= \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{E}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{E} \\ -\mathbf{E}\mathbf{C}\mathbf{A}^{-1} & \mathbf{E} \end{bmatrix} && \text{(C.39)} \\ \text{where } \mathbf{E} &= (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} && \text{(C.40)} \end{aligned}$$

yielding the precision matrix in parts:

$$\mathbf{P_1} \quad \text{is not needed} \tag{C.41}$$

$$\mathbf{P_2} = -\mathbf{P_3}\mathbf{A}\mathbf{W}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \lambda_r^{-1}\mathbf{I}\right)^{-1} \tag{C.42}$$

$$\mathbf{P_3} = \left(\mathbf{A}\mathbf{A}^{\mathrm{T}} + \mathbf{\Gamma}^{-1} - \mathbf{A}\mathbf{W}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \lambda_r^{-1}\mathbf{I}\right)^{-1}\mathbf{W}\mathbf{A}^{\mathrm{T}}\right)^{-1} \tag{C.43}$$

The conditional probability is found by expanding equation C.38 and completing the square, yielding

$$\mathrm{P}\left(\mathbf{v}_r|\mathbf{y}_r, \mathbf{W}, \mathbf{A}, \lambda_r, \mathbf{\Gamma}, \boldsymbol{\mu}^{(Y)}, \boldsymbol{\mu}^{(V)}\right) \propto -\frac{1}{2}\mathbf{v}_r^{\mathrm{T}}\mathbf{S}_r^{-1}\mathbf{v}_r + \mathbf{v}_r\mathbf{S}_r^{-1}\mathbf{m}_r \tag{C.44}$$

with

$$\mathbf{S}_r^{-1} = \mathbf{P}_3 = \mathbf{A}\mathbf{A}^{\mathrm{T}} + \mathbf{\Gamma}^{-1} - \mathbf{A}\mathbf{W}^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \gamma_r^{-1}\mathbf{I})^{-1}\mathbf{W}\mathbf{A}^{\mathrm{T}} \tag{C.45}$$

$$\mathbf{m}_r = -\mathbf{S}_r^{-1}\mathbf{P}_2^{\mathrm{T}}\mathbf{y}_r = \mathbf{A}\mathbf{W}^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \lambda_r^{-1}\mathbf{I})^{-1}\mathbf{y}_r. \tag{C.46}$$

This result is valid when $\mathbf{A}$ and $\mathbf{W}$ are point estimates. When they are given by probabilistic posterior distributions, the equivalent linear decoding matrix is instead calculated by finding the posterior estimate of the behavioural variables in terms of the latent space:

$$\mathrm{P}'(\hat{\mathbf{v}}) = \mathrm{N}\left(\hat{\mathbf{m}}, \hat{\mathbf{S}}\right) \tag{C.47}$$

$$\hat{\mathbf{m}} = \left\langle \boldsymbol{\mu}^{(V)} \right\rangle + \langle \mathbf{A} \rangle \left\langle \hat{\mathbf{h}} \right\rangle \tag{C.48}$$

$$\hat{\mathbf{S}} = \langle \mathbf{\Gamma} \rangle^{-1} + \left\langle \mathbf{A} \left\langle \hat{\mathbf{h}}\hat{\mathbf{h}}^{\mathrm{T}} \right\rangle \mathbf{A}^{\mathrm{T}} \right\rangle - \hat{\mathbf{m}}\hat{\mathbf{m}}^{\mathrm{T}} \tag{C.49}$$

and then inserting the expected value of the latent state in terms of that trial's data:

$$\left\langle \hat{\mathbf{h}} \right\rangle = \left(\mathbf{I} + \hat{\lambda}\left\langle \mathbf{W}^{\mathrm{T}}\mathbf{W} \right\rangle\right)^{-1} \hat{\lambda}\left\langle \mathbf{W} \right\rangle^{\mathrm{T}} \hat{\mathbf{y}}. \tag{C.50}$$

Thus there is a linear decoding matrix and offset given by

$$\mathbf{D} = \langle \mathbf{A} \rangle \left(\mathbf{I} + \hat{\lambda}\left\langle \mathbf{W}^{\mathrm{T}}\mathbf{W} \right\rangle\right)^{-1} \hat{\lambda}\left\langle \mathbf{W} \right\rangle^{\mathrm{T}} \tag{C.51}$$

$$\mathbf{d}_0 = \boldsymbol{\mu}^{(\mathbf{V})} - \mathbf{D}\boldsymbol{\mu}^{(\mathbf{Y})}. \tag{C.52}$$

On models with per-trial noise, this simply assumes a point estimate of $\hat{\lambda}$ using the median of the $\boldsymbol{\lambda}$s of the training data. It would be more correct to estimate the noise level from the residuals in explaining $\hat{\mathbf{y}}$, but in practice the shape of the decoding is not very sensitive to reasonable variations in noise precision.

Appendix D

Updates for Linked ICA

D.1 Priors, VB updates and Moments

Mixture component means $\boldsymbol{\mu}$ and precisions $\boldsymbol{\beta}$

The priors are given by

$$\mathrm{P}\big(\boldsymbol{\mu}^{(k,i)}\big) = \prod_{m=1}^{M_{k,i}} \mathrm{N}\big(\boldsymbol{\mu}_m^{(k,i)}|u_0, v_0\big) \tag{D.1}$$

$$\mathrm{P}\big(\boldsymbol{\mu}^{(k,i)}\big) = \prod_{m=1}^{M_{k,i}} \mathrm{Ga}\big(\boldsymbol{\mu}_m^{(k,i)}|b_0, c_0\big) \tag{D.2}$$

with the relatively uninformative priors $u_0 = 0$, $v_0 = 10^6$, $b_0 = 10^3$, $c_0 = 10^{-6}$, and the Gamma distribution in terms of the Gamma function:

$$\mathrm{Ga}(x|b, c) = \frac{x^{(c-1)}\mathrm{e}^{-x/b}}{\Gamma(c)b^c} \tag{D.3}$$

The posterior forms are given by

$$\begin{aligned}
\mathrm{P}'\big(\boldsymbol{\mu}_m^{(k,i)}\big) &= \mathrm{N}(u,v) && \text{(D.4)}\\
v^{-1} &= v_0^{-1} + \big\langle \boldsymbol{\beta}_m^{(k,i)} \big\rangle \sum_{n=1}^{N_k} \big\langle \mathbf{q}_{m,n}^{(k,i)} \big\rangle && \text{(D.5)}\\
uv^{-1} &= u_0 v_0^{-1} + \big\langle \boldsymbol{\beta}_m^{(k,i)} \big\rangle \sum_{n=1}^{N_k} \big\langle \mathbf{X}_n^{(k,i)} \big| \mathbf{q}_{m,n}^{(k,i)} = 1 \big\rangle \big\langle \mathbf{q}_{m,n}^{(k,i)} \big\rangle && \text{(D.6)}\\
\mathrm{P}'\big(\boldsymbol{\beta}_m^{(k,i)}\big) &= \mathrm{Ga}(b,c) && \text{(D.7)}\\
b^{-1} &= b_0^{-1} + \frac{1}{2} \sum_n \big\langle \mathbf{q}_{m,n}^{(k,i)} \big\rangle \Big(\big\langle (\mathbf{X}_{m,n}^{(k,i)})^2 | \mathbf{q}_{m,n}^{(k,i)} = 1 \big\rangle \\
&\quad -2 \big\langle \mathbf{X}_{m,n}^{(k,i)} \big| \mathbf{q}_{m,n}^{(k,i)} = 1 \big\rangle \big\langle \boldsymbol{\mu}_m^{(k,i)} \big\rangle + \big\langle (\boldsymbol{\mu}_m^{(k,i)})^2 \big\rangle \Big) && \text{(D.8)}\\
c &= c_0 + \frac{1}{2} \sum_n \big\langle \mathbf{q}_{m,n}^{(k,i)} \big\rangle && \text{(D.9)}
\end{aligned}$$

And the necessary moments are

$$\big\langle \boldsymbol{\beta}_m^{(k,i)} \big\rangle = bc \qquad \big\langle \log \boldsymbol{\beta}_m^{(k,i)} \big\rangle = \log(b) + \psi(c) \qquad \text{(D.10)}$$

$$\big\langle \boldsymbol{\mu}_m^{(k,i)} \big\rangle = u \qquad \big\langle (\boldsymbol{\mu}_m^{(k,i)})^2 \big\rangle = u^2 + v \qquad \text{(D.11)}$$

Mixture model weights π

The priors are given as follows:

$$\mathrm{P}(\boldsymbol{\pi}) = \prod_{k=1}^{K} \prod_{i=1}^{L} \mathrm{Dir}\big(\boldsymbol{\pi}^{(k,i)} \big| \boldsymbol{\ell}_0\big) \qquad \text{(D.12)}$$

where the uniform prior ($\boldsymbol{\pi}_0 \in \mathscr{R}^M$ is a vector of all ones) was used. The Dirichlet distribution is defined as

$$\mathrm{Dir}(\boldsymbol{Q}|\boldsymbol{\pi}) \propto \sum_{m=1}^{M} (Q_m)^{\pi_m - 1} \qquad \text{(D.13)}$$

so the needed moments are given by

$$\langle \mathrm{Dir}(\boldsymbol{\pi}) \rangle = \boldsymbol{\pi} \Big/ \sum_m \boldsymbol{\pi}_m \qquad \text{(D.14)}$$

$$\langle \log \mathrm{Dir}(\boldsymbol{\pi}) \rangle = \psi(\boldsymbol{\pi}) - \psi\left(\sum_m \boldsymbol{\pi}_m \right) \qquad \text{(D.15)}$$

where $\psi(\cdot)$ is the digamma function.

The VB updates are given by:

$$\mathrm{P}'(\boldsymbol{\pi}) = \prod_{k=1}^{K}\prod_{i=1}^{L}\mathrm{Dir}\big(\boldsymbol{\pi}^{(k,i)}\big|\boldsymbol{\ell}^{(k,i)}\big) \tag{D.16}$$

$$\boldsymbol{\ell}^{(k,i)} = \sum_{n}\big\langle \boldsymbol{q}_n^{(k,i)}\big\rangle + \boldsymbol{\pi}_0 \tag{D.17}$$

Spatial sources X and hidden mixture memberships q

The posterior distribution is also a Gaussian mixture model, factorized over components:

$$\mathrm{P}'(\mathbf{X}) = \prod_{k=1}^{K}\prod_{i=1}^{L}\mathrm{P}'\big(\mathbf{X}^{(k,i)}\big) \tag{D.18}$$

$$\mathrm{P}'\big(\mathbf{X}^{(k,i)}\big) = \sum_{m=1}^{M}\mathrm{P}'\big(\mathbf{X}^{(k,i)}|q_n^{(k,i)}=m\big)\,\mathrm{P}'\big(q_n^{(k,i)}=m\big) \tag{D.19}$$

$$\tag{D.20}$$

The individual mixture components have the distribution:

$$\mathrm{P}'\big(\mathbf{X}^{(k,i)}|q_n^{(k,i)}=m\big) = \mathrm{N}\big(\mathbf{X}_n^{(k,i)}|\mathbf{M}_{\mathbf{X},n,m},\boldsymbol{\Sigma}_{\mathbf{X},n,m}\big) \tag{D.21}$$

$$\boldsymbol{\Sigma}_{\mathbf{X},n,m}^{-1} = \Big\langle \boldsymbol{\beta}_{i,m}^{(k)}\Big\rangle + \left(\Big\langle\sum_{r=1}^{R}\mathbf{H}_{ir}^2\Big\rangle\sum_{t}\langle\lambda_t\rangle\big\langle\mathbf{W}_{ti}^2\big\rangle\right) \tag{D.22}$$

$$\begin{aligned}\mathbf{M}_{\mathbf{X},n,m}\boldsymbol{\Sigma}_{\mathbf{X},n,m}^{-1} = {}& \Big\langle\boldsymbol{\mu}_{i,m}^{(k)}\Big\rangle\Big\langle\boldsymbol{\beta}_{i,m}^{(k)}\Big\rangle \\ & + \sum_{t}\left(\Big\langle\boldsymbol{\lambda}_t^{(k)}\Big\rangle\langle\mathbf{W}_{t,i}\rangle\sum_{r=1}^{R}\mathbf{Y}_{n,t,r}\langle\mathbf{H}_{i,r}\rangle\right. \\ & \left. - \Big\langle\boldsymbol{\lambda}_t^{(k)}\Big\rangle\sum_{j\neq i}\langle\mathbf{X}_{nj}\rangle\langle\mathbf{W}_{ti}\mathbf{W}_{tj}\rangle\langle\mathbf{H}_{ir}\mathbf{H}_{jr}\rangle\right)\end{aligned} \tag{D.23}$$

While the mixture weights are distributed as

$$\mathrm{P}'\big(q_n^{(k,i)}\big) = \mathrm{Cat}\left(q_n^{(k,i)}\middle|\mathbf{Q}/\sum_{m}\mathbf{Q}_m\right) \tag{D.24}$$

$$\begin{aligned}\log(\mathbf{Q}_m) = {}& \Big\langle\log\,\boldsymbol{\pi}_{i,m}^{(k)}\Big\rangle + \frac{1}{2}\Big\langle\log\,\boldsymbol{\beta}_{i,m}^{(k)}\Big\rangle - \frac{1}{2}\Big\langle\boldsymbol{\beta}_{i,m}^{(k)}\Big\rangle\Big\langle\boldsymbol{\mu}_{i,m}^{(k)\,2}\Big\rangle \\ & -\frac{1}{2}\big\langle\log\,\boldsymbol{\Sigma}_{\mathbf{X},n,m}^{-1}\big\rangle + \frac{1}{2}(\mathbf{M}_{\mathbf{X},n,m})^2\;\boldsymbol{\Sigma}_{\mathbf{X},n,m}^{-1}\end{aligned} \tag{D.25}$$

Temporal decomposition matrix W and ARD prior $\boldsymbol{\omega}$

The posterior on $\mathbf{W}$ naturally factors across modalities:

$$\mathrm{P}'(\mathbf{W}) = \prod_{k=1}^{K}\prod_{t=1}^{T_k}\mathrm{P}'(\mathbf{W}_{t,\cdot}) \tag{D.26}$$

each of which is a normal distribution given by

$$\mathrm{P}'(\mathbf{W}_{t,\cdot}) = \mathrm{N}(\mathbf{W}_{t,\cdot}|\mathbf{m},\mathbf{V}) \tag{D.27}$$

$$\mathbf{V}^{-1} = \langle\boldsymbol{\omega}_t\rangle\,\mathbf{I} + \langle\lambda_t\rangle\left\langle\mathbf{X}^{\mathrm{T}}\mathbf{X}\right\rangle \circ \left\langle\mathbf{H}\mathbf{H}^{\mathrm{T}}\right\rangle \tag{D.28}$$

$$(\mathbf{m}\mathbf{V}^{-1})_i = \langle\lambda_t\rangle\sum_{r=1}^{R}\langle\mathbf{H}_{ir}\rangle\sum_{n=1}^{N_k}\mathbf{Y}_{nrt}\,\langle\mathbf{X}_{ni}\rangle\,. \tag{D.29}$$

The posterior on the ARD parameter $\boldsymbol{\omega}$ naturally factorizes as

$$\mathrm{P}'(\boldsymbol{\omega}) = \prod_{k=1}^{K}\prod_{t=1}^{T_k}\prod_{i=1}^{L}\mathrm{P}'\left(\boldsymbol{\omega}_{ti}^{(k)}\right) \tag{D.30}$$

with each element of $\boldsymbol{\omega}$ distributed as

$$\mathrm{P}'\left(\boldsymbol{\omega}_{ti}^{(k)}\right) = \mathrm{Ga}(()\,\boldsymbol{\omega}_t^{(k)}|b,c) \tag{D.31}$$

$$b^{-1} = b_0^{-1} + \left\langle(\mathbf{W}_{ti}^{(k)})^2\right\rangle \tag{D.32}$$

$$c = c_0 + 1/2. \tag{D.33}$$

Shared latent space (subject-course) matrix H

The posterior on $\mathbf{H}$ is a matrix normal distribution:

$$\mathrm{P}'(\mathbf{H}) = \mathrm{MN}(\mathbf{H}|\mathbf{M}_{\mathbf{H}},\boldsymbol{\Omega}_{\mathbf{H}},\boldsymbol{\Sigma}_{\mathbf{H}})\,. \tag{D.34}$$

Since all subjects r have the same noise levels and the same prior, the posterior row covariance $\boldsymbol{\Omega}_{\mathbf{H}} = \mathbf{I}_R$. The $L \times L$ column covariance is given by

$$\left(\boldsymbol{\Sigma}_{\mathbf{H}}^{-1}\right)_{ij} = \left(\boldsymbol{\Sigma}_{\mathbf{H},0}^{-1}\right)_{ij} + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\left\langle\mathbf{X}_{ni}^{(k)}\mathbf{X}_{nj}^{(k)}\right\rangle\left\langle\mathbf{W}_{ti}^{(k)}\mathbf{W}_{tj}^{(k)}\right\rangle\boldsymbol{\lambda}_t^{(k)} \tag{D.35}$$

while the posterior mean is

$$\begin{aligned}\left(\mathbf{M}_{\mathbf{H}}\boldsymbol{\Sigma}_{\mathbf{H}}^{-1}\right)_{ir} &= \left(\mathbf{M}_{\mathbf{H},0}\boldsymbol{\Sigma}_{\mathbf{H},0}^{-1}\right)_{ir} + \\ &\sum_{k=1}^{K}\sum_{t=1}^{T_k}\left\langle \boldsymbol{\lambda}_t^{(k)}\right\rangle \left\langle \mathbf{W}_{ti}^{(k)}\right\rangle \sum_{n=1}^{N_k}\mathbf{Y}_{n,r,t}^{(k)}\left\langle \mathbf{X}_{n,i}^{(k)}\right\rangle . \qquad \text{(D.36)}\end{aligned}$$

The priors are simply $\mathrm{N}(0,1)$ on each element, i.e. $\boldsymbol{\Sigma}_{\mathbf{H},0} = \mathbf{I}_L$, $\boldsymbol{\Omega}_{\mathbf{H},0} = \mathbf{I}_R$ and $\mathbf{M}_{\mathbf{H},0} = \mathbf{0}$. For MATLAB implementation, these updates are more efficiently expressed in matrix form:

$$\boldsymbol{\Sigma}_{\mathbf{H}}^{-1} = \mathbf{I}_L + \sum_{k=1}^{K}\left\langle \mathbf{X}^{(k)\mathrm{T}}\mathbf{X}^{(k)}\right\rangle \circ \left\langle \mathbf{W}^{(k)\mathrm{T}}\mathrm{diag}\left\langle \boldsymbol{\lambda}^{(k)}\right\rangle \mathbf{W}^{(k)}\right\rangle \qquad \text{(D.37)}$$

$$\mathbf{M}_{\mathbf{H}}\boldsymbol{\Sigma}_{\mathbf{H}}^{-1} = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\mathbf{Y}_{\cdot,\cdot,t}^{(k)\mathrm{T}}\left\langle \mathbf{X}^{(k)}\right\rangle \mathrm{diag}\left(\boldsymbol{\lambda}_t^{(k)}\mathbf{W}_{t,\cdot}^{(k)\mathrm{T}}\right) \qquad \text{(D.38)}$$

where $\circ$ represents the elementwise (Schur) product of two matrices.

Noise precision $\boldsymbol{\lambda}$

When separate noise is estimated for each modality (k,t), the posterior distribution of the noise precision is given by

$$\mathrm{P}'(\boldsymbol{\lambda}) = \prod_{k=1}^{K}\prod_{t=1}^{T}\mathrm{P}'\left(\boldsymbol{\lambda}_t^{(k)}\right) \qquad \text{(D.39)}$$

$$\mathrm{P}'\left(\boldsymbol{\lambda}_t^{(k)}\right) = \mathrm{Ga}\left(\boldsymbol{\lambda}_t^{(k)}\middle| b_{kt}, c_{kt}\right) \qquad \text{(D.40)}$$

$$c_{kt} = c_0 + N_k R/2 \qquad \text{(D.41)}$$

$$\begin{aligned}b_{kt}^{-1} &= b_0^{-1} + \frac{1}{2}\sum_{n=1}^{N_k}\sum_{r=1}^{R}\mathbf{Y}_{ntr}^{(k)\,2} - \sum_{n=1}^{N_k}\sum_{r=1}^{R}\left(\mathbf{Y}_{ntr}^{(k)}\sum_{i=1}^{L}\left\langle \mathbf{X}_{ni}^{(k)}\right\rangle \left\langle \mathbf{W}_{ti}^{(k)}\right\rangle \left\langle \mathbf{H}_{ir}\right\rangle\right) \\ &+\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\left[\left(\sum_{n=1}^{N_k}\left\langle \mathbf{X}_{ni}^{(k)}\mathbf{X}_{nj}^{(k)}\right\rangle\right)\left\langle \mathbf{W}_{ti}^{(k)}\mathbf{W}_{tj}^{(k)}\right\rangle\left(\sum_{r=1}^{R}\left\langle \mathbf{H}_{ir}\mathbf{H}_{jr}\right\rangle\right)\right] \qquad \text{(D.42)}\end{aligned}$$

D.2 Free Energy

Recall the notation that

$$\mathrm{KL}(z) = \mathrm{KL}\Big(\mathrm{P}'(z)\,\Big\|\,\mathrm{P}(z)\Big) = \Big\langle \log\,\mathrm{P}'(z)\,\Big\rangle_{\mathrm{P}'(z)} - \Big\langle \log\,\mathrm{P}(z)\,\Big\rangle_{\mathrm{P}'(z)} \qquad \text{(D.43)}$$

The free energy is given by

$$F = \sum_k \left[\left\langle \mathrm{P}\left(\mathbf{Y}^{(k)}|\mathbf{X}^{(k)},\mathbf{W}^{(k)},\mathbf{H},\boldsymbol{\lambda}^{(k)}\right)\right\rangle - \mathrm{KL}\left(\boldsymbol{\lambda}^{(k)}\right) - \mathrm{KL}\left(\boldsymbol{\omega}^{(k)}\right) - \mathrm{KL}\left(\mathbf{W}^{(k)}\right)\right] - \mathrm{KL}(\mathbf{H}) + \sum_k \sum_i \left[-\mathrm{KL}\left(\boldsymbol{\mu}^{(k,i)}\right) - \mathrm{KL}\left(\boldsymbol{\beta}^{(k,i)}\right) - \mathrm{KL}\left(\boldsymbol{\pi}^{(k,i)}\right) - \mathrm{KL}\left(\mathbf{q}^{(k,i)}\right) - \mathrm{KL}\left(\mathbf{X}^{(k,i)}\right)\right] \tag{D.44}$$

with the k^{th} likelihood term expanding to

$$\begin{aligned}\left\langle \mathrm{P}\left(\mathbf{Y}^{(k)}|\mathbf{X}^{(k)},\mathbf{W}^{(k)},\mathbf{H},\boldsymbol{\lambda}^{(k)}\right)\right\rangle &= \sum_{t=1}^{T_k}\left[\frac{N_k R}{2}\left\langle \log\left(\frac{\lambda_t}{2\pi}\right)\right\rangle \right. \\ &\quad -\frac{\langle\lambda_t\rangle}{2}\sum_{n=1}^{N_k}\sum_{r=1}^{R}\mathbf{Y}_{ntr}^2 + \langle\lambda_t\rangle\sum_{n=1}^{N_k}\sum_{r=1}^{R}\left(\sum_{i=1}^{L}\left\langle\mathbf{X}_n^{(k,i)}\right\rangle\left\langle\mathbf{W}_{t,i}^{(k)}\right\rangle\langle\mathbf{H}_{ir}\rangle\right)\mathbf{Y}_{ntr} \\ &\quad \left. -\frac{\langle\lambda_t\rangle}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\left\langle\mathbf{X}^{\mathrm{T}}\mathbf{X}\right\rangle_{ij}\langle\mathbf{W}_{ti}\mathbf{W}_{tj}\rangle\left\langle\mathbf{H}\mathbf{H}^{\mathrm{T}}\right\rangle_{ij}\right]\end{aligned} \tag{D.45}$$

and the KL divergence on the Gaussian mixture model expressed as

$$\begin{aligned}\mathrm{KL}\left(\mathbf{X}^{(k,i)}\right) &= \\ &\sum_{n=1}^{N_k}\sum_{m=1}^{M}\mathrm{P}'\left(q_n^{(k,i)}{=}m\right)\mathrm{KL}\left(\mathrm{P}'\left(\mathbf{X}_n^{(k,i)}|q_n^{(k,i)}{=}m\right)\middle\|\mathrm{P}\left(\mathbf{X}_n^{(k,i)}|q_n^{(k,i)}{=}m\right)\right)\end{aligned} \tag{D.46}$$

where both the conditional prior and conditional posterior are normal distributions.

Bibliography

G. Aguirre, E. Zarahn, and M. D'Esposito. Empirical analyses of BOLD fMRI statistics. *NeuroImage*, 5:199–212, 1997.

J. Ashburner and K. J. Friston. Voxel-based morphometry - the methods. *NeuroImage*, 11(6):805–821, 2000.

Y. Assaf, T. Blumenfeld-Katzir, Y. Yovel, and P. J. Basser. AxCaliber: A method for measuring axon diameter distribution from diffusion MRI. *Magnetic Resonance in Medicine*, 59:1347–1354, 2008.

H. Attias. A variational Bayesian framework for graphical models. *Advances in Neural Information Processing Systems*, 12(1–2):209–215, 2000.

H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1998.

T. Bayes and R. Price. An Essay towards Solving a Problem in the Doctrine of Chances. *Phil. Trans. R. Soc.*, 53:370–418, 1763.

C. F. Beckmann. *Independent component analysis for functional magnetic resonance imaging*. D.Phil. in information engineering, Image Analysis Group, FMRIB Centre and Robotics Research Group, University of Oxford, UK, 2004.

C. F. Beckmann and S. M. Smith. Probabilistic independent components analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23:137–152, 2004.

C. F. Beckmann and S. M. Smith. Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, 25(1):294–311, 2005.

C. F. Beckmann, M. DeLuca, J. T. Devlin, and S. M. Smith. Investigations into resting-state connectivity using independent component analysis. *Phil. Trans. R. Soc. B*, 360:1001–1013, 2005.

T. E. Behrens, M. W. Woolrich, M. E. Walton, and M. F. Rushworth. Learning the value of information in an uncertain world. *Nat Neurosci*, 10(9):1214–21, 2007.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

C. M. Bishop. Variational principal components. *Artificial Neural Networks*, 7–10 September 1999(Conference Publication No. 470):509–514, 1999.

F. D. Bowman. Spatiotemporal models for region of interest analyses of functional neuroimaging data. *Journal of the American Statistical Association*, 102(478):442–453, 2007.

R. B. Buxton, L. R. Frank, E. C. Wong, B. Siewert, S. Warach, and R. R. Edelman. A general kinetic model for quantitative perfusion imaging with arterial spin labeling. *Magn Reson Med*, 40(3):383–96, 1998.

V. Calhoun, T. Adali, N. Giuliani, J. Pekar, and G. Pearlson. Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data. *Hum Brain Mapp*, 27(1):47–62, Jan 2006.

A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, Article ID 785152:1–17, 2009.

M. A. Chappell, A. R. Groves, B. Whitcher, and M. W. Woolrich. Variational Bayesian inference for a non-linear forward model. *IEEE Trans Sig Proc*, 57(1): 223–236, 2009.

R. Choudrey and S. Roberts. Flexible Bayesian independent component analysis for blind source separation. *Proc. Int. Conf. on Independent Component Analysis*, 2001.

D. B. Ennis and G. Kindlmann. Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images. *Magnetic Resonance in Medicine*, 55: 136–146, 2006.

P. M. Figueiredo, S. Clare, and P. Jezzard. Quantitative perfusion measurements using pulsed arterial spin labeling: effects of large region-of-interest analysis. *J Magn Reson Imaging*, 21(6):676–82, 2005.

N. Filippini, B. J. MacIntosh, M. G. Hough, G. M. Goodwin, G. B. Frisoni, S. M. Smith, P. M. Matthews, C. F. Beckmann, and C. E. MacKay. Distinct patterns of brain activity in young carriers of the APOE-ε4 allele. *PNAS*, 106(17):7209–7214, 2009.

G. Flandin, F. Kherif, X. Pennec, D. Riviere, N. Ayache, and J.-B. Poline. Parcellation of brain images with anatomical and functional constraints for fMRI data analysis. In *IEEE International Symposium on Biomedical Imaging*, 2002.

K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny. Variational free energy and the Laplace approximation. *Neuroimage*, 34(1):220–34, 2007.

K. Friston, C. Chu, J. Mourao-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner. Bayesian decoding of brain images. *Neuroimage*, 39(1):181–205, 2008a.

K. J. Friston. Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage*, 16:513–530, 2002.

K. J. Friston, N. Trujillo-Barreto, and J. Daunizeau. DEM: A variational treatment of dynamic systems. *NeuroImage*, 41:849–885, 2008b.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Boca Raton: CRC Press, 2004.

L. M. Harrison, W. Penny, J. Ashburner, N. Trujillo-Barreto, and K. J. Friston. Diffusion-based spatial priors for imaging. *NeuroImage*, 38(4):677–95, 2007.

L. M. Harrison, W. Penny, G. Flandin, C. C. Ruff, N. Weiskopf, and K. J. Friston. Graph-partitioned spatial priors for functional magnetic resonance images. *NeuroImage*, 43:694–707, 2008.

J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425–2430, 2001.

J.-D. Haynes and G. Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8(5):686–691, 2005.

J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523–534, 2006.

R. D. Hoge and G. B. Pike. Quantitative measurement using fMRI. In P. Jezzard, P. M. Matthews, and S. M. Smith, editors, *Functional MRI: An introduction to methods*, pages 159–174. Oxford University Press, 2001.

R. D. Hoge, J. Atkinson, B. Gill, G. R. Crelier, S. Marrett, and G. B. Pike. Investigation of BOLD signal dependence on cerebral blood flow and oxygen

consumption: the deoxyhemoglobin dilution model. *Magn Reson Med*, 42(5):849–63, 1999.

L. Hunt, A. Groves, and T. Behrens. Investigating the role of ventromedial prefrontal cortex in decision-making under risk using magnetoencephalography. *NeuroImage*, 47(Supplement 1):S95, 2009. Organization for Human Brain Mapping 2009 Annual Meeting.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

P. Jezzard, P. M. Matthews, and S. M. Smith. *Functional Magnetic Resonance Imaging: An Introduction to Methods*. Oxford University Press, 2001.

I. Jolliffe. *Principal Component Analysis*. New York: Springer, 2002.

K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(20):352–356, 2008.

S. J. Kiebel, J. Daunizeau, C. Phillips, and K. J. Friston. Variational Bayesian inversion of the equivalent current dipole model in EEG/MEG. *NeuroImage*, 39: 728–741, 2008.

N. D. Lawrence and C. M. Bishop. Variational Bayesian independent component analysis. Technical report, 2000.

D. Le Bihan, J.-F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, and H. Chabriat. Diffusion Tensor Imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, 13:534–546, 2001.

J. Liu, O. Demirci, and V. D. Calhoun. A Parallel Independent Component Analaysis approach to investigate gemonic influence on brain function. *IEEE Sig Proc Letters*, 15:413–416, 2008.

W.-M. Luh, E. C. Wong, P. A. Bandettini, and J. S. Hyde. QUIPSS II with thin-slice TI1 periodic saturation: a method for improving accuracy of quantitative perfusion imaging using pulsed arterial spin labelling. *Magnetic Resonance in Medicine*, 41: 1246–1254, 1999.

W. M. Luh, E. C. Wong, P. A. Bandettini, B. D. Ward, and J. S. Hyde. Comparison of simultaneously measured perfusion and BOLD signal increases during brain activation with T(1)-based tissue identification. *Magn Reson Med*, 44(1):137–43, 2000.

D. J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge university press, 2003.

S. Makni, P. Ciuciu, J. Idier, and J.-B. Poline. Bayesian joint detection-estimation of brain activity using MCMC with a gamma-gaussian mixture prior model. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2006*, volume 5, 2006.

S. Makni, J. Idier, T. Vincent, B. Thirion, G. Dehaene-Lambertz, and P. Ciuciu. A fully Bayesian approach to the parcel-based detection-estimation of brain activity in fMRI. *Neuroimage*, 2008.

J. Mattout, C. Phillips, W. D. Penny, M. D. Rugg, and K. J. Friston. MEG source localization under multiple constraints: an extended Bayesian framework. *Neuroimage*, 30(3):753–67, 2006.

F. Miwakeichi, E. Martinez-Montes, P. A. Valdes-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi. Decomposing EEG data into space-time-frequency components using Parallel Factor Analysis. *NeuroImage*, 22:1035–1045, 2004.

J. Mourao-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter. Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *NeuroImage*, 28(4):980–995, 2005.

J. Mourao-Miranda, K. J. Friston, and M. Brammer. Dynamic discrimination analysis: A spatio-temporal SVM. *NeuroImage*, 36:88–99, 2007.

J. A. Mumford, L. Hernandez-Garcia, G. R. Lee, and T. E. Nichols. Estimation efficiency and statistical power in arterial spin labeling fMRI. *NeuroImage*, 33(1): 103–114, 2006.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *J. R. Statist. Soc A*, 135(Part 3):370–384, 1972.

F. Nielsen, L. Hansen, and D. Balslev. Mining for associations between text and brain activation in a functional neuroimaging database. 2(4):369–380, 2004.

F. B. Nielsen. *Variational Approach to Factor Analysis and Related Models.* Master of Science in Engineering, Intelligent Signal Processing group, Institute of Informatics and Mathematical Modelling – Technical University of Denmark, Anker Engelundsvej 1, Building 101A, 2800 Kgs. Lyngby, Denmark, 2004.

K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*, 10(9):424–30, 2006.

W. Penny, S. Kiebel, and K. Friston. Variational Bayesian inference for fMRI time series. *Neuroimage*, 19(3):727–41, 2003.

W. Penny, G. Flandin, and N. Trujillo-Barretto. Bayesian comparison of spatially regularized general linear models. *Human Brain Mapping*, 28:275–293, 2007.

W. D. Penny, N. J. Trujillo-Barreto, and K. J. Friston. Bayesian fMRI time series analysis with spatial priors. *Neuroimage*, 24(2):350–62, 2005.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

S. J. Roberts and W. D. Penny. Variational Bayes for generalized autoregressive models. *IEEE Trans Sig Proc*, 50(9):2245–2257, Sept 2002.

M. Sahani and J. F. Linden. Evidence optimization techniques for estimating stimulus-response functions. *Advances in Neural Information Processing Systems*, 15:317–324, 2003.

M. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato. Hierarchical Bayesian estimation for MEG inverse problem. *NeuroImage*, 23:806–826, 2004.

J. Scholz, M. C. Klein, T. E. Behrens, and H. Johansen-Berg. Training induces changes in white-matter architecture. *Nat Neurosci*, 12(11):1367–1368, 2009.

S. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. Nichols, C. Mackay, K. Watkins, O. Ciccarelli, M. Cader, and T. Behrens. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage*, 31:1487–1505, 2006.

M. E. Tipping. The Relevance Vector Machine. *Proceedings of NIPS*, 2000.

N. Trujillo-Barreto, E. Aubert-Vázquez, and P. A. Valdés-Sosa. Bayesian model averaging in EEG/MEG imaging. *NeuroImage*, 21:1300–1319, 2004.

N. Trujillo-Barreto, E. Aubert-Vázquez, and W. D. Penny. Bayesian M/EEG source reconstruction with spatio-temporal priors. *NeuroImage*, 39:318–335, 2008.

E. C. Wong, R. B. Buxton, and L. R. Frank. Quantitative imaging of perfusion using a single subtraction (QUIPSS and QUIPSS II). *Magnetic Resonance in Medicine*, 39(5):702–708, May 1998.

M. A. Woodbury. Inverting modified matrices. *Memorandum Rept. 42, Statistical Research Group, Princeton University, Princeton, NJ*, MR38136. 4pp, 1950.

M. W. Woolrich and T. E. Behrens. Variational Bayes inference of spatial mixture models for segmentation. *IEEE Trans Med Imaging*, 25(10):1380–91, 2006.

M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith. Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, 14:1370–1386, 2001.

M. W. Woolrich, T. E. J. Behrens, and S. M. Smith. Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, 21:1748–1761, 2004.

M. W. Woolrich, P. Chiarelli, D. Gallichan, J. Perthen, and T. T. Liu. Bayesian inference of hemodynamic changes in functional arterial spin labeling data. *Magn Reson Med*, 56(4):891–906, 2006.

K. J. Worsley, J. B. Poline, K. J. Friston, and A. C. Evans. Characterizing the response of pet and fmri data using multivariate linear models. *Neuroimage*, 6(4): 305–319, Nov 1997.

L. Xu, G. Pearlson, and V. D. Calhoun. Joint source based morphometry identifies linked gray and white matter group differences. *NeuroImage*, 44(3):777 – 789, 2009.

X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-supervised learning: From gaussian fields to gaussian processes. *Carnegie Mellon University technical report*, CMU-CS-03-175, 2003.