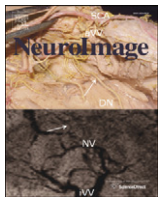
NeuroImage 62 (2012) 801–810

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Review

Bayesian inference in FMRI

Mark W. Woolrich [a,b,*]

[a] OHBA (Oxford Centre for Human Brain Activity), University of Oxford, UK
[b] FMRIB Centre, University of Oxford, UK

ARTICLE INFO

Article history:
Accepted 12 October 2011
Available online 20 October 2011

Keywords:
fMRI
Bayes
Inference
Statistics
Probability
Priors

ABSTRACT

Bayesian inference has taken FMRI methods research into areas that frequentist statistics have struggled to reach. In this article we will consider some of the early forays into Bayes and what motivated its use. We shall see the impact that Bayes has had on haemodynamic modelling, spatial modelling, group analysis, model selection and brain connectivity analysis; and consider how these advancements have spun-off into related areas of neuroscience and some of the challenges that remain. Bayes has brought to the table inference flexibility, incorporation of prior information, adaptive regularisation and model selection. But perhaps more important than these things, is the ability of Bayes to empower the methods researcher with a mathematically principled framework for inferring on any model.

© 2011 Elsevier Inc. All rights reserved.

Contents

Introduction

When I joined the field of fMRI analysis methods as a doctoral student at the FMRIB centre in Oxford in 1999, SPM was established as the most widely used fMRI analysis tool, primarily though classical or frequentist statistical approaches to inferring on the General Linear

* Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK. Fax: +44 1865 222717.
E-mail address: woolrich@fmrib.ox.ac.uk.

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.10.047

Model (GLM). At that time a striking aspect of the papers underpinning SPM, and of the field in general, was the difficulties encountered in tackling haemodynamic variability (Josephs et al., 1997) and the spatio-temporal autocorrelations present in fMRI (Worsley and Friston, 1995; Worsley et al., 1996b). These problems were important as they affected a neuroimager's ability to find anything but the strongest activations, or to report accurate statistics.

I spent the first part of my doctorate on the temporal autocorrelation problem. This led to two things. First, a tool that formed the basis of FEAT, the fMRI GLM tool in the first version of FSL. Second, a realisation that working solely within the confines of parametric frequentist statistics was not for me. The main reason for this were the apparent restrictions on the forms of the model that could be used, and on the nature of the statistical questions that could be asked. Only in certain special cases was it possible to derive analytical forms for the null distributions required by frequentist statistics. For example, while it was possible to dream up a new model of fMRI data that captured some hitherto unmodelled but important phenomenon, it was far from trivial to use classical approaches to derive sensitive and unbiased statistics on the model parameters of interest.

Pressing issues were proving difficult to address. For example, how do we infer on models that properly account for our expectations about the haemodynamics, including their nonlinearities and variability over different brain areas? How do we infer on models that properly incorporate modelling of the temporal and spatial noise correlations, alongside our prior expectation that brain "activity" occurs in contiguous clusters? There was no clear route to answering these questions within a classical statistical framework.

Putting aside the possibility that this may have been a symptom of my limited frequentist statistics expertise, I was compelled to search for alternative solutions. Unsurprisingly, I was not alone in that enterprise. One emerging area that looked to overcome some of the restrictions of parametric frequentist statistics was the use of non-parametric, permutation testing approaches (Nichols and Holmes, 2002). This is indeed a compelling way to handle the problem of deriving unbiased statistics for a given estimator. However, it does not provide a systematic way to derive the most sensitive estimators of model parameters given a new model. Another emerging area in fMRI that could potentially do this was Bayesian statistics.

The first Bayesian fMRI approaches

While there was a Bayesian inference paper on PET in 1993 by Holmes and Ford (1993), to my knowledge it was not until around the turn of the century that the first Bayesian inference papers for fMRI started to appear. Some of the first Bayesian approaches in fMRI were point estimation Maximum a Posterior Bayesian approaches used to incorporate prior information (Descombes et al., 1998; Hartvig and Jensen, 2000). The first paper implementing a *fully* Bayesian statistics approach, i.e. that considered the full posterior probability distribution rather than just point estimates, appeared in 1998 by Frank et al. (1998). However, to be frank it was, and has been, somewhat overlooked. This is surprising given the novelty and foresight of the paper, and particularly considering the surge of Bayesian approaches that followed. One possible reason for this is that the authors decided to not use the word "Bayes" or "Bayesian" in the title, abstract or even in the introduction — perhaps as a homage to the notion that the adjective "Bayesian" did not come into general use in statistics until the mid-twentieth century (Fienberg, 2006). Nonetheless Bayes rule itself does appear on the second page; and as Fig. **2** shows, they demonstrated the basic principle of posterior probability mapping (PPM) for the first time.

Other fully Bayesian treatments followed including Genovese (2000) and Kershaw et al. (1999). As with Frank et al. (1998) these were voxelwise temporal models, in the same spirit as the existing mass univariate frequentist GLM approaches. So what was the unique

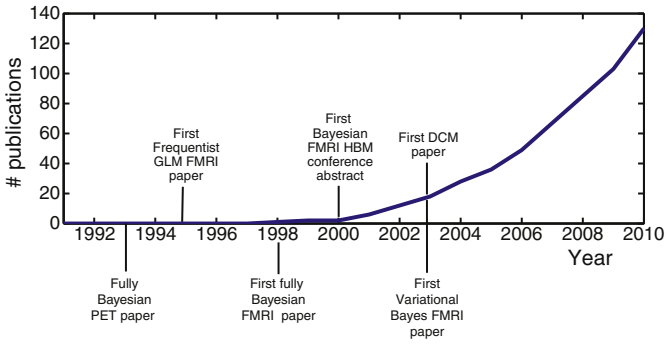


Fig. 1. Plot of cumulative number of publications since 1991 with keywords containing fMRI and Bayes (or Bayesian) [Source: Scopus]. Also shown are the author's best estimate of when notable events occurred in the field of fMRI Bayesian Inference.

perspective that Bayes had to offer compared with the established frequentist GLM approaches? As Kershaw et al. (1999) put it *"With [Bayesian] methodology it is possible to derive a relevant statistical test for activation in an fMRI time series no matter how complicated the parameters of the model are. The derivation is usually quite straightforward and results may be extracted from it without first having to find estimates for all of the parameters."* In other words, Bayes provides you with a mathematically principled framework in which you can probabilistically infer on model parameters no matter what, or how complicated, the model is.

Probabilistic inference on generative models

Generative models are a natural way for us to incorporate our understanding of the brain and of the neuroimaging modality to make predictions about what neuroimaging data looks like (see Fig. **3**). However, in practice we want to do the opposite. We want to be able to take acquired data (plus a generative model) and extract pertinent information about the brain (i.e., "infer" on the model and its parameters). Bayesian statistics offers a solution to this problem, and also provides a framework in which we can do much more besides.

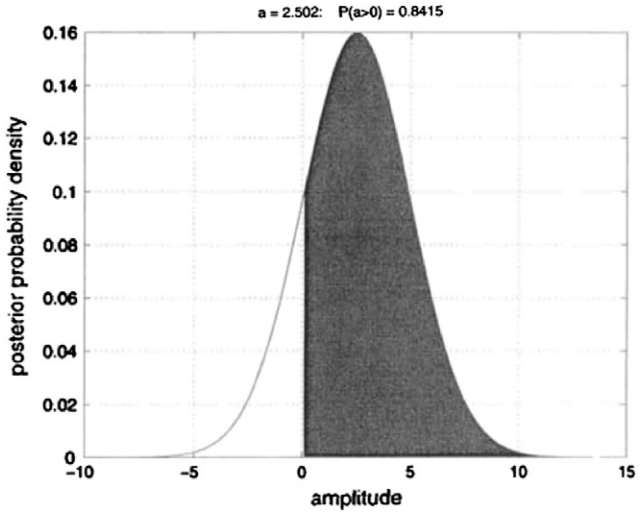


Fig. 2. Caption and figure reproduced from one of the first Bayesian fMRI papers by Frank et al. (1998): "An example of the posterior distribution of the amplitude [of the effect size]. The amplitude estimate d is the value at the peak in the posterior distribution. The probability of the amplitude being within a certain range is obtained from the area contained within the distribution between those limits (i.e.,the integral of the distribution over that range). For instance, the shaded area shown is the probability that the activation amplitude is greater than zero, $P(a>0)=0.8415$. The peak in the distribution yields the amplitude estimate $d=2.502$.".

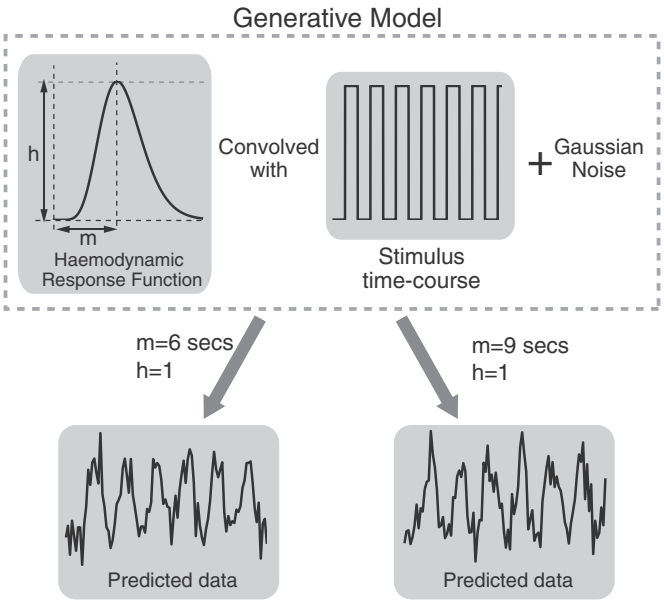


Fig. 3. Generative models are at the core of all Bayesian neuroimaging analysis techniques. This figure shows an example of a simple voxelwise generative model for predicting fMRI data. It consists of a parameterised haemodynamic response function (HRF) that is convolved with a time-course of the known experimental stimulus (in this case a boxcar stimulus) and then added to Gaussian noise. For the sake of simplicity we are assuming that the variance of the Gaussian noise is known. The only unknown parameters in the model are the time-to-peak, m, and the height, h, of the HRF. For different values for m and h, we can predict what the fMRI data looks like in a voxel.

Bayes' rule tells us how (for a model $\mathcal{M}$) we should use the data, $\mathbf{Y}$, to update our *prior* belief in the values of the parameters θ, $p(\theta|\mathcal{M})$ to a *posterior* distribution of the parameter values $p(\theta|Y,\mathcal{M})$:

$$p(\theta|\mathbf{Y},\mathcal{M}) = \frac{p(\mathbf{Y}|\theta,\mathcal{M})p(\theta|\mathcal{M})}{p(\mathbf{Y}|\mathcal{M})} \quad (1)$$

The term $p(\mathbf{Y}|\theta,\mathcal{M})$ is the *likelihood* and typically corresponds to the generative model. Fig. **4** illustrates just such a case for the application of Bayesian inference on fMRI data from Fig. **3**.

Computational complexity

Given the apparent benefits of Bayesian inference, why did Bayes not take off until the turn of the century, as Fig. **1** demonstrates? Besides the need for the field of fMRI analysis to mature, one reason is perhaps that calculating the *posterior* probability distribution function (PDF) in Eq. (1) is seldom straightforward. The denominator in Eq. (1) is:

$$p(\mathbf{Y}|\mathcal{M}) = \int_\theta p(\mathbf{Y}|\theta,\mathcal{M})p(\theta|\mathcal{M})d\theta \quad (2)$$

and unfortunately this integral can not be often solved analytically.

We will not go into the technical details of how to solve this and related integrals in this article. However, Fig. **5** summarises some of the options available. Many of the early Bayesian fMRI papers relied on numerical integration approaches, such as Markov Chain Monte Carlo (MCMC) sampling (Genovese, 2000; Woolrich et al., 2001). These were particularly time consuming. For example, a fully Bayesian inference approach for spatio-temporal modelling of fMRI data that we developed took some six hours to infer on a single slice (Woolrich et al., 2001, 2004c). This was clearly prohibitive as a "goto" tool for the neuroscience community. Computers have, of course, since speeded up, allowing techniques like MCMC to become viable approaches in certain applications (Behrens et al., 2007a; Woolrich et al., 2004a).

Variational Bayes

A key innovation in bringing Bayesian inference and its advantages to the neuroimaging masses, was to develop techniques for solving Bayes equation that did not rely on prohibitively time-consuming numerical integration. A general framework for doing this is *Variational Bayes*. The first "pure" Variational Bayes (VB) fMRI paper was Penny et al. (2003), in which it was used to tackle the temporal autocorrelation problem in fMRI time series. At roughly the same time, the related approach of Expectation Maximisation (EM) appeared in Friston (2002).

The cunning plan in Variational Bayes is to approximate the true posterior distribution by estimating it using a posterior factorised over subsets of the model parameters. This results in update equations that provide us with the desired (approximate) posterior distributions in a fraction of the time needed with approaches such as MCMC. However, models do not often meet the required conditions for even VB to be tractable, for example, when we have nonlinear generative models (as in Fig. **3**). An important solution to this problem was first introduced by Friston (2002) and is what we refer to here as *Approximate Variational Bayes*. This deploys first or second order Taylor series approximations of the generative model, allowing the problem to be solved using VB in the normal way (Chappell, 2009; Friston, 2002; Friston et al., 2007; Woolrich and Behrens, 2006). This particular approach, and variations of it, has proved to be the workhorse of the more advanced Bayesian methods such as Dynamic Causal Modelling (DCM) (Friston et al., 2003).

Bayesian versus frequentist inference

With the arrival of Bayes on the scene, there were some inevitable debates about the meaning and utility of Bayes when compared with the established frequentist method. On the one hand, classical frequentist statistics P-values refer to the frequency of outcomes, whereas Bayesians use probabilities to express degrees of belief. Debates comparing Bayes to frequentist have raged in the applied statistics community for sometime (Fienberg, 2006), but one of the first papers to expose the neuroimaging community to these issues was Friston et al. (2002). There were two particular points of comparison that stood out for me.

First, Friston et al. (2002) pointed out that a problem with frequentist inference is that, with sufficient data or sensitivity, trivial departures from the null hypothesis can be declared as significant. In other words with enough fMRI data every voxel in the brain will reject the null hypothesis. This is because in practice *no* voxels will show exactly zero response to the stimulus. Although in practise, for typical group study sample sizes, this is not a dominant problem in the inference of fMRI group "activation" maps.

Second, Friston et al. (2002) suggested that Bayes eschews the multiple comparisons problem. Does Bayes indeed avoid the need for the multiple comparison corrections when we look to threshold marginal PPMs of GLM regression parameters? In my opinion, the answer is the same as it is in the frequentist setting: it depends on what you want to control when you threshold.

For example, if you want to label voxels as "active" by thresholding marginal PPMs at a probability of 95% that the regression parameter is greater than zero, then in a completely non-activating brain, on average 5% of the voxels will be labelled as active. In this particular case, Bayes is no different to frequentist null hypothesis testing. That is, if you want to control for wrongly labelling voxels across multiple tests, then by definition it is still necessary to apply multiple comparison corrections on Bayesian PPMs as well. Nonetheless, as we will consider in Other approaches section, Bayes does allow for a wider range of inference questions to be asked.

Priors

Bayesian statistics requires that we specify our prior probabilistic belief about the model parameters. This requirement has sometimes

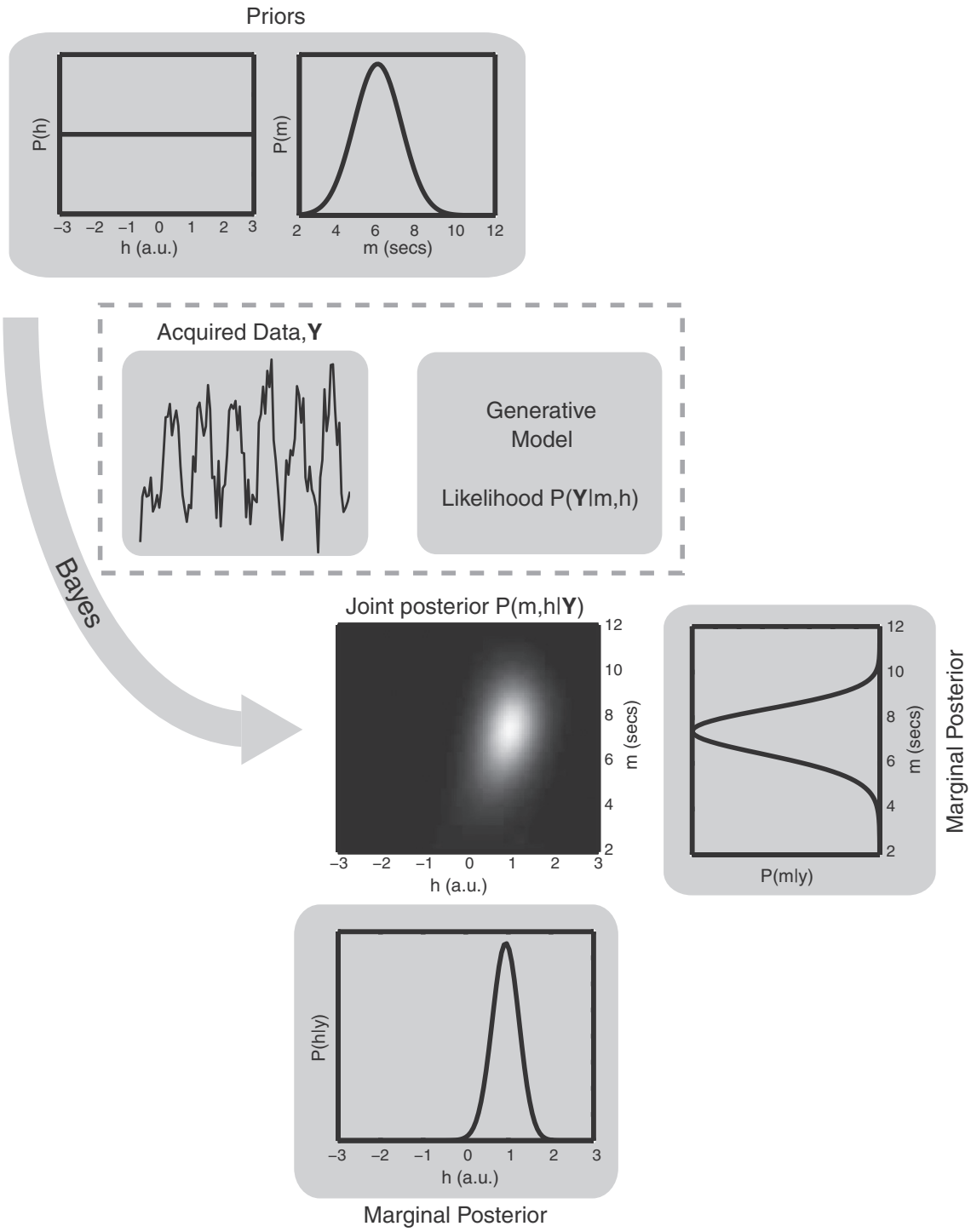


Fig. 4. Schematic of the process of Bayesian inference on the generative model in Fig. 3. The ingredients consist of the acquired fMRI data, the generative model, and the priors. The generative model provides us with the likelihood term in Bayes' rule. Bayes combines these ingredients to give us probabilistic inference on the model parameters (Eq. (1)) in the form of the joint posterior distribution across all parameters in the model. However, we are often interested in the posterior distribution on a single parameter, or a subset of parameters. This can be obtained by integrating (i.e., averaging) over parameters to obtain the "marginal" distribution.

been a source of criticism of the Bayesian approach over the years: how do we sensibly and appropriately specify the prior distributions? However, Bayesians support the view that we cannot infer from data without making assumptions. Indeed, the act of choosing a generative model itself constitutes an assumption; and a decision to ignore prior information, as we implicitly do in a frequentist approach, is a subjective decision in itself. Furthermore, it turns out that having a framework within which we can specify prior assumptions can be a big advantage.

HRF models

Biophysical priors

By the end of the nineties it was firmly established that fMRI analysis required flexible hemodynamic response function (HRF) modelling, both across the brain and between subjects. As shown in Fig. **6**, HRF flexibility could be achieved within the frequentist GLM framework by using basis functions (Josephs et al., 1997). However, it is

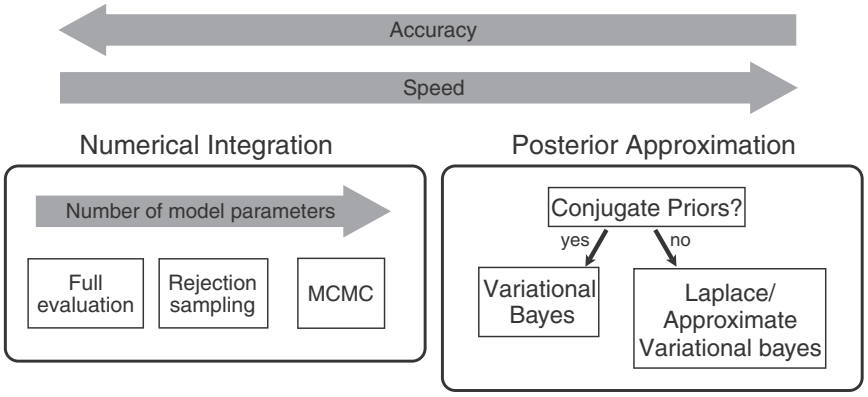


Fig. 5. The use of Bayes requires integrations to be performed that are seldom tractable, in which case there are broadly speaking two separate approaches used: (1) solve the integrals numerically, or (2) make approximations to the posterior distribution. Note that "full evaluation" refers to exhaustive evaluation of the posterior distribution over a grid of parameter values.

possible for a basis set to produce nonsensical HRFs, as no constraints are placed on the possible linear combinations (Fig. **6**c). Subsequently, Genovese (2000), Gössl et al. (2001b), Woolrich et al. (2001, 2004c) took similar Bayesian approaches to using parameterised HRFs with parameters describing characteristics such as the time-to-peak and size of the undershoot. Priors could be placed on these HRF model parameters to ensure biological plausibility. As illustrated in Fig. **6**d the inclusion of such prior information can result in increased sensitivity (Woolrich et al., 2004b).

More advanced HRF modelling approaches followed, made possible by the mathematical framework provided by Bayesian inference. In particular, Friston (2002) was the first to use Bayes to infer on a fully Bayesian biologically informed generative model.

Regularisation priors

An early example of the use of regularisation priors was in the use of semi-parametric Bayesian approaches for HRF modelling (Ciuciu et al.,

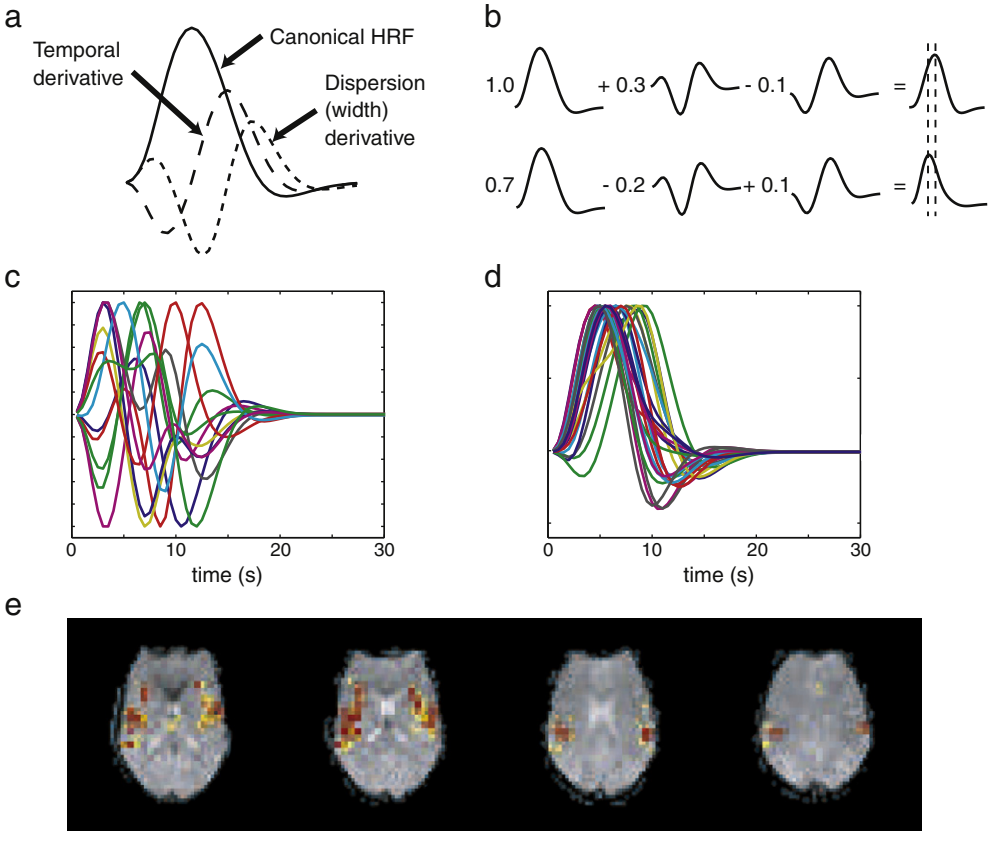


Fig. 6. The use of biophysical prior information on the shape of the haemodynamic response in fMRI data can provide increased sensitivity in detecting activity. (a) Example of a haemodynamic response function (HRF) basis set containing three basis functions. (b) Different linear combinations of the basis functions can be used to obtain different HRF shapes. (c) It is possible for many of the possible linear combinations of the basis set to produce nonsensical looking HRFs. (d) Bayesian biophysical priors can be used such that we constrain the inference to only those sets of parameters that give biophysically plausible HRF shapes. (e) Difference in thresholded activation between the HRF model with Bayesian priors and the HRF model without Bayesian priors for a single-event pain stimulus dataset: [red] voxels are active for both models, [yellow] voxels are active for just the constrained HRF model, and [blue] voxels are active for just the unconstrained HRF model. Note that there are no blue voxels visible.

2003; Goutte et al., 2000; Marrelec et al., 2003). "Semi-parametic" refers to the idea that the HRF does not have a fixed parametrised form, instead the HRF is allowed to have any form with a parameter describing the size of the HRF at each time point.

Without regularisation these models have too many parameters for stable inference. Bayesian regularisation approaches overcome this problem by placing priors on the HRF parameters that encode the prior belief that the HRF is temporally smooth. Such methods make no strong assumptions about the shape of the response function, and hence are suitable as exploratory approaches or perhaps when abnormal HRFs are possible.

Spatial models

A question challenging fMRI methods developers at the turn of the century was how to best handle spatial correlations in the data, alongside the idea that we expect activity to occur in clusters of voxels. Existing frequentist methods used a combination of pre-smoothing and spatial statistics based on Random Field Theory (Worsley et al., 1996a). However, these frequentist approaches relied on the subjective selection of a number of parameters, such as the amount of spatial smoothing to impose, and the choice of cluster forming thresholds. Bayes offered the possibility of a solution to these problems via the use of adaptive spatial priors on the regression parameters in the GLM.

The first approaches to spatial regularisation used spatial MRF priors, and inferred using numerical integration approaches such as MCMC (Gössl et al., 2001a; Woolrich et al., 2001, 2004c). However, the computational expense of inferring on spatial models is even higher than with voxelwise temporal models, and so more computationally efficient approaches were needed. As such, Variational Bayesian approaches were soon developed (Penny et al., 2005; Woolrich et al., 2004b). This work on MRFs has since been generalised within the more flexible framework of spatial Gaussian Process priors, allowing for the modelling of spatial non-stationarities (Harrison et al., 2007) and the combining of spatial and non-spatial prior information (Groves et al., 2008).

The spatial priors contain control parameters that regulate their strength. Crucially, these "hyperparameters" are inferred via the Bayesian inference framework at the same time as the rest of the model. This demonstrates a key advantage of fully Bayesian approaches. Frequentist GLM inference approaches had become riddled with heuristically tuned or user specified variables. Fully Bayesian approaches offered the promise of analysis methods which were much less dependent on arbitrarily chosen variables.

Other approaches

Over the years a number of different spatial models have been proposed, including a Bayesian wavelets approach (Flandin and Penny, 2007). One notable alternative is to model the spatial distribution of activation maps using mixture models. As shown in Fig. **7**, these use a mixture of distributions representing the "active" and "non-active" voxels (Everitt and Bullmore, 1999; Hartvig and Jensen, 2000; Woolrich and Behrens, 2006; Woolrich et al., 2005).

The interest in using mixture modelling centred on the potential for greater inference flexibility. False positive rate (FPR) could be controlled by thresholding using the estimated distribution of the "non-active" voxels. However, now we could also look to approximately control the true positive rate (TPR) by thresholding using the probability of a voxel being "active". This may be of real importance when using fMRI for pre-surgery planning (Bartsch et al., 2006).

However, the wider-spread use of mixture modelling has been hampered by the presence of structured noise artefacts (e.g. spontaneous networks of activity, stimulus correlated motion) that violate the distributional assumptions that need to be made. It maybe that more sophisticated structured noise modelling approaches are needed to render the distributional assumptions valid (Makni et al., 2008).

Mixture modelling has also been tackled using the relatively recent development in the Bayesian machine learning community of *non-parametric Bayes*. This is a bit of a misnomer as it refers to the inference of models where there are a massive number of parameters. For example, Dirichlet process priors (Fergusson, 1973) are used in non-parametric mixture models where there are effectively an infinite number of distributions modelled (also termed infinite mixture models). These allow inference to be made on the number of classes in the mixture given the data. Such methods have been used in the context of FMRI to infer on active regions using a spatial mixture model (Kim and Smyth, 2006), and in the context of diffusion MRI to classify brain regions according to their connections as seen through tractography (Jbabdi et al., 2008).

Hierarchical multi-subject models

Following the first applications of Bayes to fMRI time series data, attention soon turned to the group level. Arguably, it was also around this time that the fMRI methods field in general began to realise that ensuring statistical validity of *first-level* single session statistics may not be the most important thing to focus on. After all, the vast majority of fMRI neuroscience studies were being used to address questions about activation effects in populations of subjects. Having sensitive and valid *group* statistics was the bottom line.

Fig. **8** illustrates an example scenario where the question of interest involves estimating the difference in activation between two groups of subjects. This question is addressed by having different GLMs at the session, subject and group level in a hierarchical fashion. Hierarchical modelling is something that fits very naturally into the Bayesian framework via a cascade of conditional probabilities. As such Friston et al. (2002) proposed the first Bayesian group inference approach using a hierarchical model.

Summary statistic approaches

However, in fMRI the human and computational costs involved in data analysis are relatively high, and so it was desirable to be able to make group-level inferences using the *results* of separate first-level analyses. This is commonly referred to as the *"summary statistics"* approach, as only the summary statistics from a lower level in the hierarchy are needed in the analysis of the next level. The dominant frequentist group analysis approach at the time used GLM regression parameter estimates as summary statistics (Holmes and Friston, 1998). However, this was only optimal under certain conditions. For example, it required balanced designs, i.e., all lower-level design matrices need to be identical.

At that stage it was not clear how to provide a summary statistics approach without these restrictions, as there was no clear way to obtain a parametric frequentist solution. In contrast, Bayes was able to provide us an answer. In Woolrich et al. (2005) we use Bayes to show how to use the summary statistics approach without restrictions, by passing up information about not only the effect sizes from the lower levels, but also their variances. This formed the basis of FSL tool, FLAME, for doing group inference.

FLAME has a number of distinct advantages over the pre-existing frequentist-derived approach (Holmes and Friston, 1998). Firstly, FLAME permits the analysis of fMRI data where the lower-level design matrices have different structure from each other (e.g., contain behavioural scores as regressors). Secondly, as illustrated in Fig. **8**b, FLAME can provide more accurate variance estimation and can also increase the ability to detect real activation. We have recently extended this approach to the between-study level, to carry out coordinate based meta-level analysis (CBMA) augmented with Gaussian Process priors for doing spatial interpolation (Salimi-Khorshidi et al., 2011).

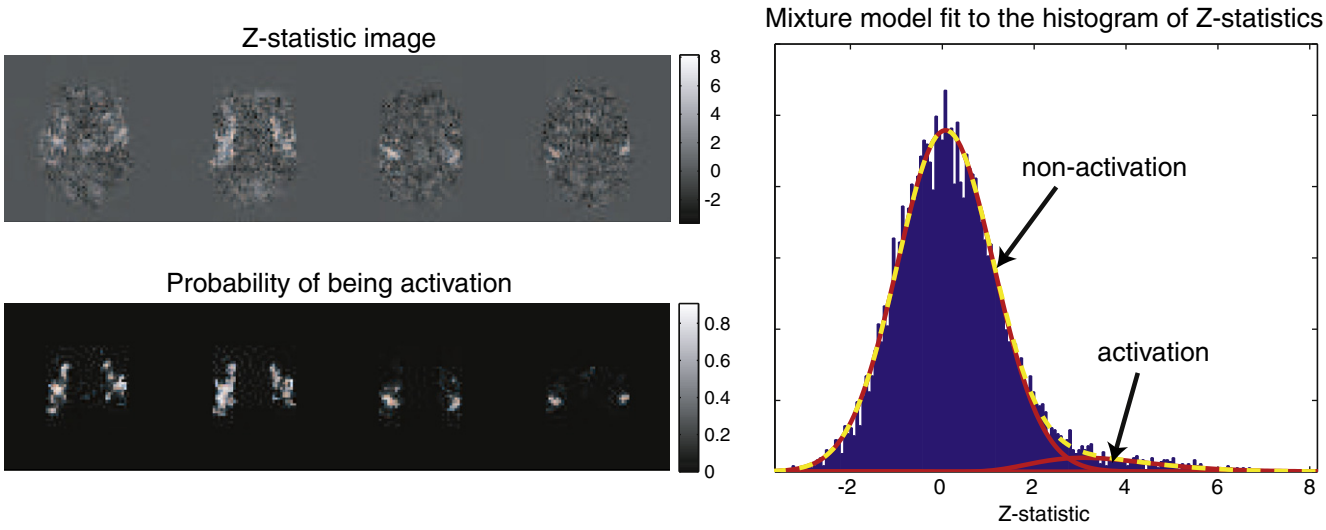


Fig. 7. Mixture modelling of a *Z*-statistic image. [Top-left] Four slices of a *Z*-statistic image obtained from fitting a GLM to the fMRI data for a single subject for a pain stimulus condition. [Right] Mixture model fit to the histogram of *Z*-statistics. Non-activating voxels are modelled as coming from a close-to-zero mean Gaussian distribution and activating voxels as coming from a Gamma distribution. [Bottom-left] Image showing the probability that a voxel is activating — this information can be used in thresholding to approximately control the true positive rate (TPR) as an alternative to null hypothesis testing.

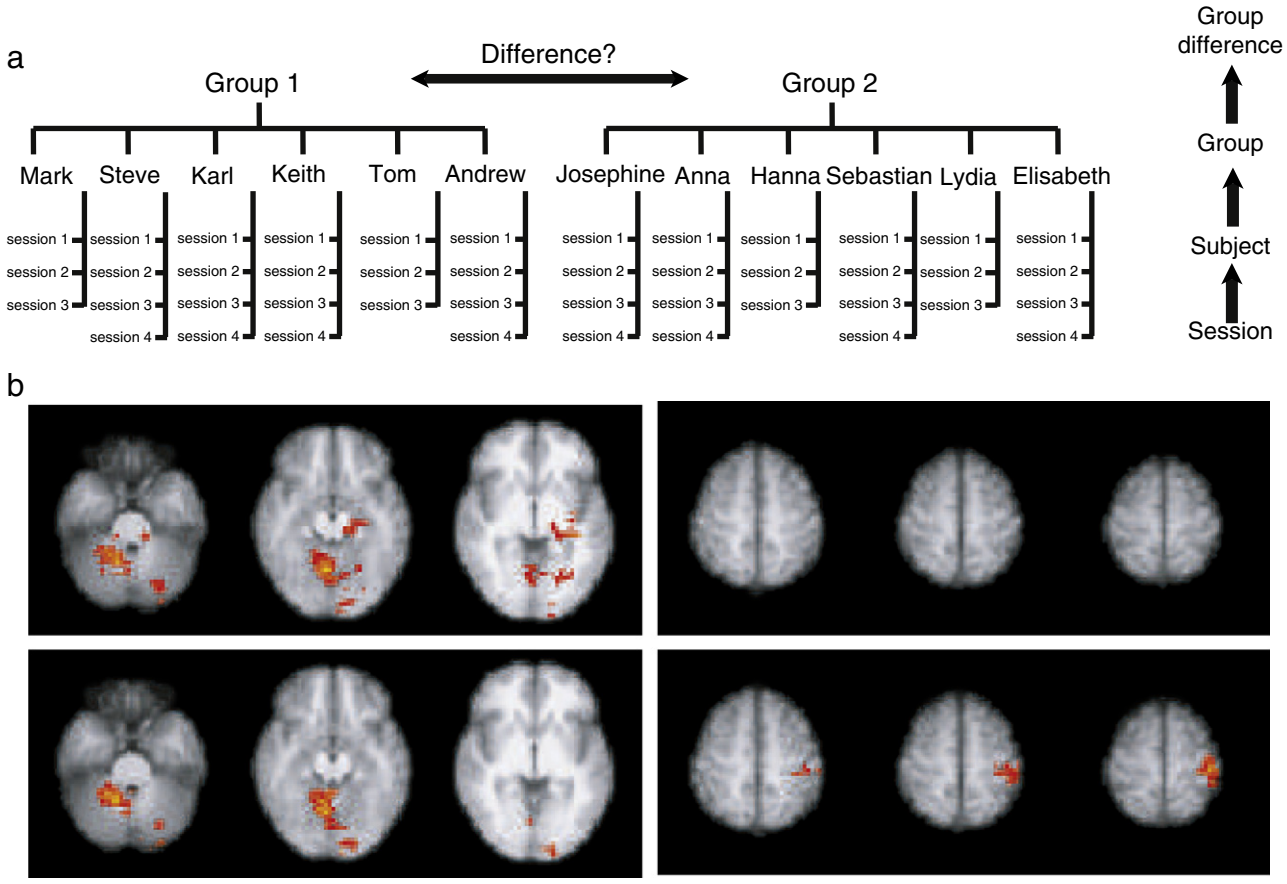


Fig. 8. (a) Hierarchical GLM for the analysis of group fMRI data. With a summary statistics approach the GLMs are estimated one level at a time and summary statistics are passed up to the next level of the hierarchy. (b) Thresholded group activation for two different group analysis approaches and for two different motor tasks. [top] A frequentist-derived approach that only uses first-level regression parameter estimates as summary statistics, and [bottom] the Bayesian-derived FLAME approach that augments the first-level regression parameter estimate summary statistics with their variances. [left] Index finger tapping vs rest. There is a general decrease in z-statistics in the FLAME approach, demonstrating the increased accuracy of the cross-subject random-effects variance estimation, due to the use of the first-level variances. [right] Sequential finger tapping vs index finger tapping. There is a general increase in z-statistics in the FLAME approach, demonstrating increased efficiency in parameter estimation by down-weighting subjects that have high first-level variance.

Model selection

Bayesian inference has the ability to be able to assess the evidence in favour of a particular model. It is tempting to say that this made it possible to compare fMRI models of different complexity for the first time. But it was possible to effectively do model selection prior to this by using f-tests, albeit only with linear models. Furthermore, this all does a disservice to cross validation methods, which can also assess

models by balancing the goodness of fit with model complexity (Carew et al., 2003; Strother et al., 2002). Nonetheless, the pursuit of model selection via the Bayesian method has provided the fMRI community with some invaluable insights.

In order to evaluate a model using Bayes, the quantity that needs to be evaluated is the denominator of the posterior PDF given in Eq. (2), i.e, the probability of the data given the model. This quantity, often termed the *model evidence*, accounts for both *accuracy* (data fit) and *complexity* of the model given the observed data. Model selection consists of calculating this quantity for a given number of models, and selecting the model with highest evidence.

As with general Bayesian inference, estimating the Bayesian model evidence for most practical cases requires approximations. Early examples of this were to select the model order of autoregressive models of fMRI noise using the Variational free energy (Penny et al., 2003), or the Deviance Information Criterion (DIC) (Woolrich et al., 2001, 2004c) as approximations to the model evidence. Other applications of fMRI model selection followed including approaches for choosing between different haemodynamic models (Stephan et al., 2007b) and different spatial priors (Penny et al., 2007).

The power of this approach became particularly apparent when it was used to select different Dynamic Causal Models (DCM) Penny et al. (2004). The combined use of Bayesian model comparison and DCM provided a novel framework for evaluating competing hypotheses about the architecture of networks of brain connectivity for the first time. Recent developments also allow for model comparisons across families of models to overcome problems with "model dilution" when there are a large number of models to compare, or if there is heterogeneity over subjects in group comparisons (Penny et al., 2010).

There are, as with any method, a few health warnings and challenges attached to the use of Bayesian model evidence, but which often get overlooked. First, it is important for users of the method to understand what can be achieved. A classic error is to assume that the model evidence will always assign the highest evidence to the "true model". Putting aside the question of whether it is meaningful to talk about a "true model" for real data, the model evidence only tells us about the best model given the data. For example, the data in question may only support a model of limited complexity, and more complex models may be preferred if more, or higher SNR, data were available.

Another often overlooked issue is the dependency of model evidence on the noise model. If the noise model is wrong then erroneous model comparisons will result. For example, if noise correlations are not properly accounted for then the correct balance of model complexity and goodness of fit can be lost (Groves et al., 2008). This may be an area in which cross validation methods are safer than using the Bayesian model evidence.

Finally, there is the challenge of how to ensure objectivity in the Bayesian model evidence, with regards to the particular problem that models are not generally invariant to reparametrisation. This is where the model is rewritten using a transformation of the parameters (e.g. working with $\tilde{x} = log(x)$ rather than x), such that the likelihood remains the same, but the marginalisation (integration) is carried out on the new parameter. The problem is that different parameterisations can produce different model evidences, and there is no principled reason to prefer one parameterisation over another. The use of Jeffrey's priors might be a solution, as these render the inference invariant to reparametrisation (Woolrich et al., 2004a). However, Jeffrey's priors are improper and so result in an unnormalised (infinite) evidence. This is related to the "Marginalisation Paradox" whereby marginalising in different orders gives different results, and is also potentially a problem when inferring marginal distributions on parameters. To my knowledge this is an unsolved problem that has received insufficient attention to date.

Online model selection

Bayesian model comparison, alongside non-Bayesian cross validation approaches, has been a very useful method for shedding light on the relative merits of different models. However, when the number of potential models gets very large, both cross validation and standard Bayesian model evidence estimation approaches, become computationally impractical. This can occur when we are considering nested sub-models in which different unique combinations of model parameters are zero. In this case there is a combinatorial explosion that occurs in the number of models to compare. For example, a model with N such parameters would result in 2^N model comparisons.

This is where online approaches to model selection come into their own. A good example of online model selection is the use of shrinkage priors, also known as Automatic Relevance Determination (ARD) priors (MacKay, 1995). In its most straightforward form ARD consists of placing a Gaussian prior, with zero mean and unknown variance, on any model parameter in question. The prior variance, σ^2, is then inferred alongside the rest of the model, such that if $\sigma^2 \to 0$ then the parameter is knocked out of the model. This was used in Woolrich et al. (2004c) to select the presence of an undershoot in an HRF model, and the model order of temporal autoregressive models. Smith et al. (2003) used a related approach to do variable selection of the regression parameters in the GLM. More recently this has been used in a Bayesian implementation of ICA for determining the number if independent components (Groves et al., 2011), and for carrying out online feature selection in Bayesian decoding of fMRI data (Friston et al., 2008a).

Brain connectivity

In the last few years there has been a considerable trend away from the spatial mapping of task related activity towards inferring brain connectivity. This is motivated by the idea that connectivity gets us closer to the actual mechanisms of brain function, and has also been somewhat fueled by the emergence of the resting state phenomenon. The stories behind the use of fMRI to understand brain connectivity and resting state networks are the topics of other articles in this issue. However, it is worth noting here that many of the benefits and techniques of Bayes honed on task-fMRI activation mapping, are being translated into understanding brain connectivity. Many examples of this have already been mentioned in this article, including the approach of Dynamic Causal Modelling (DCM).

DCM for fMRI is a method that performs Bayesian inference on the full generative model of a network of connected brain areas. It combines differential equations describing the neuronal interactions between brain regions with those describing the biophysical haemodynamics (Friston et al., 2003; Stephan et al., 2007a). DCM has since been extended to handle stochastic differential equations, which model spontaneous endogenous neuronal fluctuations. This makes it possible to infer on resting state fMRI using DCM for the first time (Daunizeau et al., 2009; Li et al., 2011).

The complexity of DCM is a very good example of the ability of Bayes to provide us with a mathematical framework for inferring on any generative model. It is hard to conceive of how one would attempt to do statistical inference (rather than just point estimation) on models such as DCM without a Bayesian framework. There are perhaps those who would argue that this is one of the dangers of Bayes, in that you can faithfully turn the Bayesian handle, oblivious to the potential pitfalls of the inappropriateness of the model, accuracy of the inference approximations, and problems with local minima. For this reason Bayesian methods should always be verified with sensible validation approaches, such as the use of more time consuming but more accurate inference approaches and perhaps cross validation (Daunizeau et al., 2011b).

DCM is an important step in the field of brain connectivity, in that it employs biologically plausible generative models (often known as "effective connectivity"). As pointed out in Friston (2011), using non-generative model measures of connectivity based on statistical dependence measures such as correlation (known as "functional connectivity") can leave the neuroscientist with spurious results. This is due to the

danger that functional connectivity measures can change between conditions or groups simply due to changes in, for example, the signal-to-noise (SNR) of the data. Effective connectivity approaches, such as DCM, effectively model the SNR and so are robust to these problems.

Job done? Well not quite. Using DCM for practical *network discovery* is still an unsolved problem due to the large number of models that must be searched over. A potential solution is the recent idea of *post hoc* Bayesian model comparison, in which the Bayesian model evidence for nested submodels can be approximated from the inference on the full model with all parameters being non-zero (Friston and Penny, 2011; Friston et al., 2010b). However, it remains to be seen if this can reliably work on network models with large numbers of nodes (>10).

Another particular challenge that Bayes can potentially address is how to go about designing the best experiments, e.g. by changing the form of the stimuli, such that the experiment has the best chance of distinguish between the competing DCMs (Daunizeau et al., 2011a). Indeed such innovations could have important applications in optimising data acquisition in other settings as well.

Discussion

Bayesian approaches have made a substantial impact on fMRI analysis over the last 18 years. A testament to this is the spin-off benefits that have occurred in areas of neuroscience related to fMRI. I have occasionally been privileged to be involved in a couple of important areas where this has occurred.

A good example is the use of Bayesian methods for doing probabilistic diffusion tractography from diffusion weighted MRI data. Using techniques honed on fMRI models, Bayesian MCMC methods allowed us to infer on local (voxelwise) generative models of diffusion using Bayes to obtain PDFs on local white matter fibre directions (Behrens et al., 2003, 2007a). Probabilistic tractography can then be used to track through these local fibre directions, to infer on anatomical white matter connections between two distant points (Behrens et al., 2003). This work, inspired by the development of Bayesian methods in fMRI, constitutes the FSL diffusion tractography toolbox (FDT).

Another example, is in the use of Bayesian methods in computational neuroscience to model the brain as a Bayesian learner (Dayan and Abbott, 2001; Doya, 2007). We have used this approach to provide evidence of the mechanisms of learning in decision making tasks using fMRI data (Behrens et al., 2007b, 2008). These are good example of the use of neuroimaging data to test *specific* mechanisms of brain function, rather than simply looking for correlates of task or behavioural variables. See Kiebel et al. (2009), Daunizeau et al. (2010), Friston et al. (2010a), Vilares and Kording (2011) for other examples of modelling the brain as a Bayesian learner.

Bayesian methods have also been developed in parallel in related functional neuroimaging modalities such as MEG and EEG. Some Bayesian M/EEG methods pre-date related fMRI approaches, for example the work by Baillet and Garnero (1997) and Schmidt et al. (1999). However, there also more recent examples of developments in fMRI Bayesian methods feeding back into M/EEG methods development. This includes MEG source reconstruction methods (Friston et al., 2008b; Woolrich et al., 2011), and approaches that build on fMRI DCM to develop DCM for M/EEG using more sophisticated neuronal models (Kiebel et al., 2006; Penny et al., 2009).

In my time in the field, I have often heard Bayesian critics say that there is nothing particularly magical about Bayesian inference; that many of the attributes of Bayes that Bayesians hold dear have equivalent approaches in non-Bayesian statistics. Indeed, forays into the statistical literature indicate that convergence is often being sought, or simply pointed out (Guyon et al., 2010). Nonetheless, the intuitive and mathematical framework provided by the one simple equation that is Bayes rule, has made it possible for fMRI methods developers to navigate with more clarity through the methodological possibilities. As computational power increases and the field works with more complex models, the clarity that the Bayesian methods brings to the table suggests that it should stay at the forefront of neuroimaging methods development for some time to come.

Acknowledgments

Funding for Mark Woolrich is from the Wellcome Trust. Thanks to Adrian Groves, Mark Jenkinson and Stephen Smith for their help with the article.

References

Baillet, S., Garnero, L., 1997. A bayesian approach to introducing anatomo-functional priors in the eeg/meg inverse problem. IEEE Trans. Biomed. Eng. 44, 374–385.
Bartsch, A., Homola, G., Biller, A., Solymosi, L., Bendszus, M., 2006. Diagnostic functional mri: illustrated clinical applications and decision-making. J. Magn. Reson. Imaging 23, 921–932.
Behrens, T., Woolrich, M., Jenkinson, M., Johansen-Berg, H., Nunes, R., Clare, S., Matthews, P., Brady, J., Smith, S., 2003. Characterisation and propagation of uncertainty in diffusion weighted MR imaging. Magn. Reson. Med. 50, 1077–1088.
Behrens, T.E.J., Berg, H.J., Jbabdi, S., Rushworth, M.F.S., Woolrich, M.W., 2007a. Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? NeuroImage 34, 144–155.
Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007b. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–1221.
Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S., 2008. Associative learning of social value. Nature 456, 245–249.
Carew, J.D., Wahba, G., Xie, X., Nordheim, E.V., Meyerand, M.E., 2003. Optimal spline smoothing of fmri time series by generalized cross-validation. NeuroImage 18, 950–961.
Chappell, M., 2009. Variational bayesian inference for a nonlinear forward model. IEEE Signal Process. 57.
Ciuciu, P., Poline, J.B., Marrelec, G., Idier, J., Pallier, C., Benali, H., 2003. Unsupervised robust nonparametric estimation of the hemodynamic response function for any fmri experiment. IEEE Trans. Med. Imaging 22, 1235–1251.
Daunizeau, J., Friston, K.J., Kiebel, S.J., 2009. Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models. Physica D 238, 2089–2118.
Daunizeau, J., den Ouden, H.E.M., Pessiglione, M., Kiebel, S.J., Stephan, K.E., Friston, K.J., 2010. Observing the observer (i): meta-bayesian models of learning and decision-making. PLoS One 5, e15554.
Daunizeau, J., Friston, K., Stephan, K., 2011a. Optimizing experimental design for identifying networks in the brain using fmri. 17th Annual Meeting of the Organisation for Human Brain Mapping 58 (2), 312–322.
Daunizeau, J., David, O., Stephan, K.E., 2011b. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. NeuroImage 58 (2), 312–322.
Dayan, P., Abbott, L.F., 2001. "Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems". The MIT Press; 1st edition (December 1, 2001).
Descombes, X., Kruggel, F., von Cramon, D.Y., 1998. fmri signal restoration using a spatio-temporal markov random field preserving transitions. NeuroImage 8, 340–349.
Doya, K., 2007. Bayesian brain: probabilistic approaches to neural coding. MIT Press, 326.
Everitt, B.S., Bullmore, E.T., 1999. Mixture model mapping of the brain activation in functional magnetic resonance images. Hum. Brain Mapp. 7, 1–14.
Fergusson, T., 1973. A bayesian analysis of some nonparametric problems. Ann. Stat. 1, 209–230.
Fienberg, S., 2006. When did bayesian inference become "bayesian". Bayesian Anal. 1, 1–40.
Flandin, G., Penny, W.D., 2007. Bayesian fmri data analysis with sparse spatial basis function priors. NeuroImage 34, 1108–1125.
Frank, L.R., Buxton, R.B., Wong, E.C., 1998. Probabilistic analysis of functional magnetic resonance imaging data. Magn. Reson. Med. 39, 132–148.
Friston, K., 2002. Bayesian estimation of dynamical systems: an application to fmri. NeuroImage 16, 513–530.
Friston, K.J., 2011. Functional and effective connectivity: a review. Brain Connect. 56.
Friston, K., Penny, W., 2011. Post hoc bayesian model selection. NeuroImage 56, 2089–2099.
Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and bayesian inference in neuroimaging: theory. NeuroImage 16, 465–483.
Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.
Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34, 220–234.
Friston, K., Chu, C., Mouraomiranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008a. Bayesian decoding of brain images. NeuroImage 39, 181–205.
Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G., Mattout, J., 2008b. Multiple sparse priors for the m/eeg inverse problem. NeuroImage 39, 1104–1120.
Friston, K.J., Daunizeau, J., Kilner, J., Kiebel, S.J., 2010a. Action and behavior: a free-energy formulation. Biol. Cybern. 102, 227–260.
Friston, K.J., Li, B., Daunizeau, J., Stephan, K.E., 2010b. Network discovery with dcm. NeuroImage.
Genovese, C., 2000. A Bayesian time-course model for functional magnetic resonance imaging data (with discussion). J. Am. Stat. Assoc. 95, 691–703.
Gössl, C., Auer, D.P., Fahrmeir, L., 2001a. Bayesian spatiotemporal inference in functional magnetic resonance imaging. Biometrics 57, 554–562.

Gössl, C., Fahrmeir, L., Auer, D.P., 2001b. Bayesian modeling of the hemodynamic response function in bold fmri. NeuroImage 14, 140–148.
Goutte, C., Nielsen, F.A., Hansen, L.K., 2000. Modeling the haemodynamic response in fmri using smooth fir filters. IEEE Trans. Med. Imaging 19, 1188–1201.
Groves, A., Chappell, M., Woolrich, M., 2008. Combined spatial and non-spatial prior for inference on mri time-series. NeuroImage.
Groves, A.R., Beckmann, C.F., Smith, S.M., Woolrich, M.W., 2011. Linked independent component analysis for multimodal data fusion. NeuroImage 54, 2198–2217.
Guyon, I., Saffari, A., Dror, G., Cawley, G., 2010. Model selection: beyond the bayesian/frequentist divide. J. Mach. Learn. Res. 11, 61–87.
Harrison, L.M., Penny, W., Ashburner, J., Trujillo-Barreto, N., Friston, K.J., 2007. Diffusion-based spatial priors for imaging. NeuroImage 38, 677–695.
Hartvig, N.V., Jensen, J.L., 2000. Spatial mixture modeling of fmri data. Hum. Brain Mapp. 11, 233–248.
Holmes, A.P., Ford, I., 1993. A bayesian approach to significance testing for statistic images from pet. Ann. Nucl. Med. 7.
Holmes, A., Friston, K., 1998. Generalisability, random effects & population inference. Fourth Int. Conf. on Functional Mapping of the Human Brain: NeuroImage, p. S754.
Jbabdi, S., Woolrich, M., Behrens, T., 2008. Multiple-subjects connectivity-based parcellation using infinite mixture models. Fourteenth Int. Conf. on Functional Mapping of the Human Brain.
Josephs, O., Turner, R., Friston, K., 1997. Event-related fMRI. Hum. Brain Mapp. 5, 1–7.
Kershaw, J., Ardekani, B.A., Kanno, I., 1999. Application of bayesian inference to fmri data analysis. IEEE Trans. Med. Imaging 18, 1138–1153.
Kiebel, S.J., David, O., Friston, K.J., 2006. Dynamic causal modelling of evoked responses in eeg/meg with lead field parameterization. NeuroImage 30, 1273–1284.
Kiebel, S.J., von Kriegstein, K., Daunizeau, J., Friston, K.J., 2009. Recognizing sequences of sequences. PLoS Comput. Biol. 5, e1000464.
Kim, S., Smyth, P., 2006. Hierarchical Dirichlet Processes with Random Effects. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), MIT Press, NIPS, pp. 697–704.
Li, B., Daunizeau, J., Stephan, K.E., Penny, W., Hu, D., Friston, K., 2011. Generalised filtering and stochastic dcm for fmri. NeuroImage.
MacKay, D., 1995. Developments in probabilistic modelling with neural networks — ensemble learning. Proceedings of the Third Annual Symposium on Neural Networks. Springer, Nijmwegen, Netherlands, pp. 191–198.
Makni, S., Beckmann, C., Smith, S., Woolrich, M., 2008. Combining ica and glm for fmri data analysis. Fourteenth Int. Conf. on Functional Mapping of the Human Brain.
Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., Poline, J.B., 2003. Robust bayesian estimation of the hemodynamic response function in event-related bold fmri using basic physiological information. Hum. Brain Mapp. 19, 1–17.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.
Penny, W., Kiebel, S., Friston, K., 2003. Variational bayesian inference for fmri time series. NeuroImage 19, 727–741.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. NeuroImage 22, 1157–1172.
Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fmri time series analysis with spatial priors. NeuroImage 24, 350–362.
Penny, W., Flandin, G., Trujillo-Barreto, N., 2007. Bayesian comparison of spatially regularised general linear models. Hum. Brain Mapp. 28, 275–293.
Penny, W.D., Litvak, V., Fuentemilla, L., Duzel, E., Friston, K., 2009. Dynamic causal models for phase coupling. J. Neurosci. Methods 183, 19–30.
Penny, W.D., Stephan, K.E., Daunizeau, J., Rosa, M.J., Friston, K.J., Schofield, T.M., Leff, A.P., 2010. Comparing families of dynamic causal models. PLoS Comput. Biol. 6, e1000709.
Salimi-Khorshidi, G., Nichols, T., Smith, S., Woolrich, M., 2011. Using gaussian-process regression for meta-analytic neuroimaging inference based on sparse observations. IEEE Trans. Med. Imaging.
Schmidt, D.M., George, J.S., Wood, C.C., 1999. Bayesian inference applied to the electromagnetic inverse problem. Hum. Brain Mapp. 7, 195–212.
Smith, M., Pütz, B., Auer, D., Fahrmeir, L., 2003. Assessing brain activity through spatial bayesian variable selection. NeuroImage 20, 802–815.
Stephan, K.E., Harrison, L.M., Kiebel, S.J., David, O., Penny, W.D., Friston, K.J., 2007a. Dynamic causal models of neural system dynamics: current state and future extensions. J. Biosci. 32, 129–144.
Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007b. Comparing hemodynamic models with dcm. NeuroImage 38, 387–401.
Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework. NeuroImage 15, 747–771.
Vilares, I., Kording, K., 2011. Bayesian models: the structure of the world, uncertainty, behavior, and the brain. Ann. N. Y. Acad. Sci. 1224, 22–39.
Woolrich, M., Behrens, T., 2006. Variational Bayes inference of spatial mixture models for segmentation. IEEE Transactions on Medical Imaging 25 (10), 1380–1391.
Woolrich, M., Brady, M., Smith, S., 2001. Hierarchical fully Bayesian spatio-temporal analysis of FMRI data. Seventh Int. Conf. on Functional Mapping of the Human Brain.
Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., 2004a. Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage 21 (4), 1732–1747.
Woolrich, M., Behrens, T., Smith, S., 2004b. Constrained linear basis sets for HRF modelling using variational Bayes. NeuroImage 21, 1748–1761.
Woolrich, M., Jenkinson, M., Brady, J., Smith, S., 2004c. Fully Bayesian spatio-temporal modelling of FMRI data. IEEE Trans. Med. Imaging 23, 213–231.
Woolrich, M., Behrens, T., Beckmann, C., Smith, S., 2005. Mixture models with adaptive spatial regularization for segmentation with an application to fmri data. IEEE Trans. Med. Imaging 24, 1–11.
Woolrich, M., Hunt, L., Groves, A., Barnes, G., 2011. MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. Neuroimage 57 (4), 1466–1479.
Worsley, K.J., Friston, K.J., 1995. Analysis of fmri time-series revisited–again. NeuroImage 2, 173–181.
Worsley, K., Marrett, S., Neelin, P., Vandal, A., 1996a. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. 4 (1), 58–73.
Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., 1996b. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. 4, 58–73.