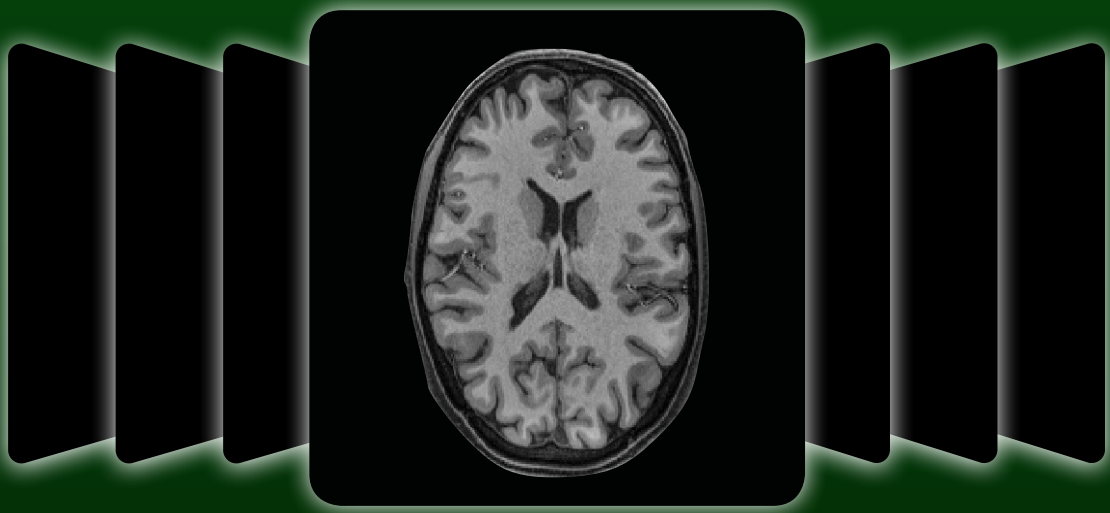
OXFORD NEUROIMAGING PRIMERS

Short Introduction to

Bayesian Inference for Neuroimaging



Michael Chappell
Mark Woolrich
Mark Jenkinson

Series editors:
Mark Jenkinson and Michael Chappell

PRIMER APPENDIX

Short Introduction to Bayesian Inference for Neuroimaging

Oxford Neuroimaging Primers

Series Editors: Mark Jenkinson and Michael Chappell

Introduction to Neuroimaging Analysis

Mark Jenkinson
Michael Chappell

Introduction to Perfusion Quantification using Arterial Spin Labelling

Michael Chappell
Bradley MacIntosh
Thomas Okell

Introduction to Resting State fMRI Functional Connectivity

Janine Bijsterbosch
Stephen Smith
Christian Beckmann

Primer Appendices

Short Introduction to Bayesian Inference for Neuroimaging

Short Introduction to Brain Anatomy for Neuroimaging

Short Introduction to MRI Physics for Neuroimaging

Short Introduction to MRI Safety for Neuroimaging

Copyright

1st Edition (v1.0)

Preface to the Oxford Neuroimaging Primers series

The Oxford Neuroimaging Primers are aimed to be readily accessible texts for new researchers or advanced undergraduates in neuroimaging who want to get a broad understanding of the ways in which neuroimaging data can be analyzed and interpreted. All primers in this series have been written so that they can be read as stand-alone books, although they have also been edited so that they "work together" and readers can read multiple primers in the series to build up a bigger picture of neuroimaging and be equipped to use multiple neuroimaging methods.

Understanding the principles of the analysis of neuroimaging data is crucial for all researchers in this field, not only because data analysis is a necessary part of any neuroimaging study, but also because it is required in order to understand how to plan, execute, and interpret experiments. Although MR operators, radiologists, and technicians are often available to help with data collection, running the scanner, and choosing good sequences and settings, when it comes to analysis, researchers are often on their own. Therefore, the Oxford Neuroimaging Primers seek to provide the necessary understanding of how to do analysis while at the same time trying to show how this knowledge relates to being able to perform good acquisitions, design good experiments, and correctly interpret the results.

The series has been produced by individuals (both authors and editors) who have developed neuroimaging analysis techniques, used these methods on real data, packaged them as software tools for others to use, taught courses on these methods, and supported people around the world who use the software they have produced. We hope that this means everyone involved has not only the experience to instruct, but also the empathy to support the reader. It has been our aim for these primers to not only lay out the core principles that apply in any given area of neuroimaging, but also to help the reader avoid common pitfalls and mistakes (many of which the authors themselves probably made first). We also hope that the series is also a good introduction to those with a more technical background, even if they have to forgo some of the mathematical details found in other more technical works. We make no pretense that these primers are the final word in any given area, and we are aware that the field of neuroimaging continues to develop and improve, but the fundamentals are likely to remain the same for many years to come. Certainly some of the advice you will find in these primers will never fail you—such as always look at your data.

Our intention with the series has always been to support it with practical examples, so that the reader can learn from working with data directly and will be equipped to use the know- ledge they have gained in their own studies and on their own data. These examples, including datasets and instructions, can be found on the associated website (www.neuroimagingprimers.org), and directions to specific examples are placed throughout each primer. As the authors are also the developers of various software tools within the FMRIB Software Library (FSL), the examples in the primers mainly use tools from FSL. However, we intend these primers to be as general as possible and present material that is relevant for all readers, regardless of the soft- ware they use in practice. Such readers can still use the example data available through the primer website with any of the major neuroimaging analysis toolboxes. We encourage all readers to interact with these examples, since we strongly believe that a lot of the key learning is done when you actually use these tools in practice.

Mark Jenkinson & Michael Chappell, Oxford, January 2017

Preface

This text is one of a number of appendices to the Oxford Neuroimaging Primers, designed to provide extra details and information that someone reading one of the primers might find helpful, but where it is not crucial to the understanding of the main material. This appendix specifically addresses the principles that underpin Bayesian Inference, as it is used in neuroimaging. In it we seek to go into more detail than we might in one of the primers, for those who want to understand more about how Bayesian Inference can be used for data analysis. In turn, this appendix also provides a high level introduction to individuals who are interested in developing their own Bayesian Inference methods, or find they need to select between different methods in a specific application.

We hope that this appendix, in keeping with the series as a whole, will be an accessible introduction to the topic of Bayesian Inference for those without a background in the physical sciences. Hence, we have concentrated on concepts rather than delving into any detailed mathematics. However, we also hope it is a good introduction to physical scientists meeting Bayesian Inference for the first time, perhaps before going on to more technical texts, such as those we include in the Further Reading at the end.

This appendix contains several different types of boxes in the text that are designed to help you navigate the material or find out more information for yourself. To get the most out of this appendix, you might find the description of each type of box below helpful.

Boxes

These boxes contain more technical or advanced descriptions of some topics covered in this appendix. None of the material in the rest of the appendix assumes that you have read these boxes, and they are not essential for understanding any of the other material. If you are new to the field and are reading this appendix for the first time, you may prefer to skip the material in these boxes and come back to them later.

> ***Box 4.1:* Frequency and**
>
> So far we have considered

Further Reading

At the end, we include a list of suggestions for further reading, including both articles and books. A brief summary of the contents of each suggestion is included, so that you can choose the most relevant references for you. None of the material in this appendix assumes that you have read anything from the further reading. Rather, this list suggests a starting point for diving deeper, but is by no means an authoritative survey of all the relevant material you might want to consult.

FURTHER READING

- Huettel, S. A., Song, A. ...

Whilst the principles of Bayesian Inference are well established and thus the material in this appendix will, we hope, be relevant for many years to come. Advances in the field continue and new techniques appear all the time, particularly with the growth in the closely related field of machine learning and artificial intelligence methods. Hence, all we hope for as authors is that this will be a useful introduction to what is a large and fascinating field of research that extends well beyond purely neuroimaging applications.

Michael Chappell, Mark Woolrich and Mark Jenkinson

Contents

1 Introduction 1

2 Generative Models 1

3 Inference 3

4 Priors 6

4.1 Biophysical 7

4.2 Regularisation priors 7

5 Hierarchies 8

6 Model Selection 9

6.1 Bayes Factors 9

6.2 Shrinkage Priors 9

7 Practicalities 10

7.1 Numerical Methods 11

7.2 Approximations 12

8 Conclusion 13

1 Introduction

Bayesian inference has become increasingly popular for the analysis of neuroimaging data. It is found in a great many advanced methods for a variety of neuroimaging techniques including BOLD fMRI, diffusion, and ASL perfusion. This means it is also quite common to use neuroimaging analysis software that contains a Bayesian inference method. Whilst it is largely unnecessary for the purposes of carrying out a neuroimaging study to have a deep understanding of Bayesian inference, particularly the practicalities of how the relevant computations are actually carried out, it can be helpful to understand the principles. This Short Introduction gives an overview of Bayesian inference as it is typically found in neuroimaging applications. Bayesian inference is is a very general method with many of applications and there are plenty of good books to consult if you would like to know more about how to use it to solve other data analysis problems (see Further Reading).

Bayesian inference operates in the world of probability and statistics. However, the way that it is commonly used in neuroimaging appears subtly different from the way you most probably met statistics in your education and certainly the way you are likely to have met Bayes theorem before, if you ever have. What has made Bayesian inference so popular is that is provides a mathematical framework which makes it possible to take a principled approach to the complex analysis problems that occur in neuroimaging. What it offers is a consistent way to handle uncertainty, whether this be the uncertainty of our pre-existing knowledge of the brain or our imaging equipment, or uncertainty introduced by noise, and then to quantify the resulting uncertainty in the estimates we make from the data.

2 Generative Models

Typically in neuroimaging we are looking to extract specific information about the brain from imaging measurements. Since these measurements are noisy, it is not possible to directly map the measured values to precise physical, biological or physiological quantities. For example, we cannot directly set up a rule that states “If the fMRI data looks exactly like X, then the brain is definitely active in area Y”. However, it is comparatively easy to turn the problem around and specify “If the brain is active in area Y, then the FMRI data should look like X”. Assuming that we can construct a suitable description of how the data has been generated then we can predict what the imaging data should look like for different underlying states (e.g., blood volume, flow and oxygenation in a brain region). This is referred to as generative modelling, since the thing that predicts the data is normally a mathematical model. It is the use of generative models with Bayes theorem that lies at the heart of Bayesian inference for most neuroimaging applications. Figure 1 illustrates an example of a generative model, in this case for fMRI data. Here the model itself includes both a description of the time-course of the stimulus that a participant in the experiment experienced and also the haemodynamic response function that links the stimulation to the MRI signal, which is based on changes in the haemodynamics. Notice that strictly the generative model also includes the addition of noise that occurs in the measurement process. Often, in many analysis methods, the effects of noise are treated implicitly; for example, least squares optimization implicitly assumes zero mean white noise.

The generative model is a natural way to incorporate our understanding of both the brain and the neuroimaging methods we are using to image it. The model allows us to make predictions of what the data will look like, but in practice, as we have already noted, what we want to do is take the data and use that to extract information about the brain. What we want, therefore, is to estimate information about the brain given some data. For example, in Figure 1, the model has parameters m

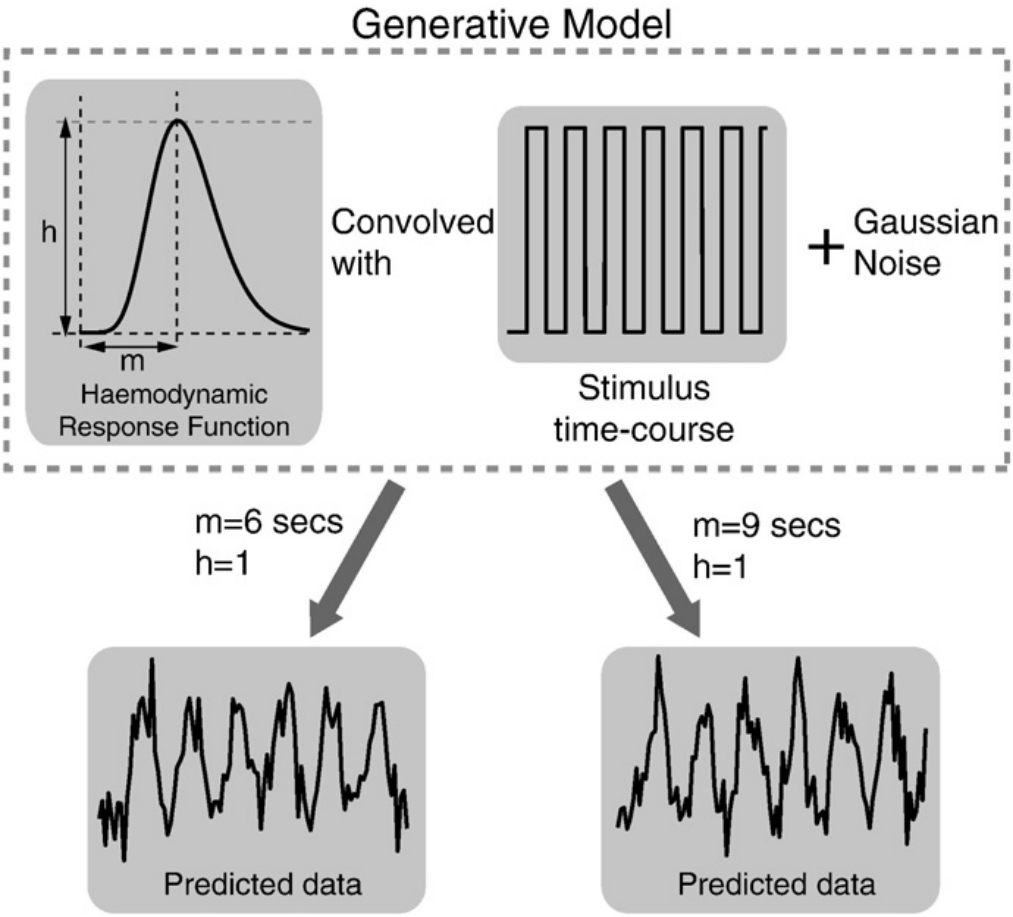


Figure 1: An example of a generative model for fMRI series analysis. Here it is assumed that the observed data can be described by the combination of Gaussian (white) noise and a deterministic component that represents the MRI signal, that itself is composed of a time-course related to the stimulation that the subject experienced, convolved with a haemodynamic response function parameterised by two parameters h and m. In this model everything is known apart from the parameters h and m and the amplitude of the noise - these would be the parameters to be estimated from the data. Note that using the generative model it is possible to simulate data according to any choice of the unknown parameters as illustrated at the bottom of the figure. This figure is reproduced with permission from Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M. A., Makni, S., Behrens, T., et al. (2009). Bayesian analysis of neuroimaging data in FSL. NeuroImage, 45(1, Supplement 1), S173-S186. http://doi.org/10.1016/j.neuroimage.2008.10.055

and h that tell us about the properties of the response to the stimulus and we want, for a given set of data, to work out what values those parameters had.

The classical approach to estimation is to do model-fitting, for example, by choosing the parameter values that minimize the squared error between the measured data and prediction from the generative model. A familiar example would be the problem of linear regression: finding the best straight line through a set of (noisy) data points. In that example, we are modeling the data by a straight line and we are seeking two parameters: the slope and intercept. This method allows us to extract a single 'best guess' estimate for each parameter in the model, but ignores the uncertainty we have in those parameters, for example due to noise. Returning to the example of linear regression; there may be multiple different straight lines that are all plausible for the data points, reflecting the fact that the data were noisy. Rather than being forced to choose a single solution, we might prefer to capture the range of solutions, as well as how plausible each solution is. Additionally, we might have some existing knowledge about the parameters in our model that we would like to

take into account, for example some values might be more plausible in healthy physiology (e.g., the intercept must be positive). The classical approach doesn't offer a systematic way to incorporate this information.

3 Inference

As we saw in section 2, we are aiming to estimate the parameters of a model given some measured data. We want to do this whilst also capturing uncertainty, including the range of possible solutions, as well as some measure of how well each solution explains the data. A natural way to do this is to use the concept of probability. We can use probability, a value strictly between 0 and 1 (inclusive), to capture our belief[1] in the values of the parameters in the model. Since the parameters in the generative models we use in neuroimaging can take a range of values it is most natural to resort to probability distributions, rather than single probability values. If you need a refresher on probability distributions you might like to look at Box 1.

Probability and probability distributions naturally enter into our calculations when we consider the noise, or errors, that appear in our measurements. These errors will most often be random (or at least we will assume they can be treated as such) and so we can model them as having arisen from a random process where the probability of producing an error of a given size can be specified.

The process of inference is a matter of finding the probability distribution for the model parameter(s) we are interested in, given the data we measured, which we write as

	$P(\theta \mid Y)$	(1)

where

θ are the parameters in the model,

Y are the measured data, and

the '|' symbol means 'given'.

Note that for a model with one parameter this is a conventional probability distribution as you might sketch on a piece of paper with the parameter on the x-axis and probability density on the y-axis (see Box 1). For a model with multiple parameters we can naturally extend the definition to multiple dimensions.

What we have written in equation 1 is called, in Bayesian inference, the posterior distribution and it arises from the inference process. Note that strictly it depends both on the model parameters and the model itself; if we changed the model we would expect a different answer. Thus we should have written $P(\theta \mid Y, M)$ to remind ourselves that the model matters. However, once we have settled on a generative model that we want to use for inference, we tend to leave 'M' out of the equations and consider it to be implied. But, as we will see in section 6, if we have more than one possible model it will be important to leave it in so that we can do model comparisons.

The posterior distribution contains a lot of information, since all plausible parameter values (or combinations) are represented. However, for the purposes of interpretation and getting an 'answer' it doesn't provide a neat summary of what the data tells us. Since the posterior is a probability distribution we can use all the standard tools of statistics on it. For example, we could take the mean of the distribution to give a representative 'best' estimate and calculate the variance or standard deviation to get a measure of the variability in the estimate.

---

[1] We might call it 'confidence', but we don't commonly use that term in this way because it could get confused with the statistical concept of a confidence interval.

Box 1: Probability Distributions

For discrete events, such a tossing a coin or rolling a die, the associated probability distribution records the probability associated with each outcome, e.g. 0.5 for Heads and 0.5 for Tails. When using Bayesian inference in neuroimaging we are more often concerned with continuous rather than discrete variables. For example, the parameter of interest might be the signal change associated with the BOLD effect. In this case, it no longer makes sense to talk about the probability of a particular value, say measuring a signal of 1.0, because there are an infinite number of 'very close' values that we might measure depending upon how finely we divide up the measurement range. For example, when you say you have measured 1.0 it is not typical to distinguish this from 1.001 or 0.999, as 1.0 implies that the value is rounded to one decimal place and thus represents all values in the range 0.995 to 1.049. Hence, it only makes sense to associate a single probability value to a measurement value of 1.0 if we also define a range of values that we would treat the same as 1.0.

Following this logic we can define a continuous probability distribution function that records the probability density for all possible values, to get the probability we then specify a range of values and compute the area under the function (via integration). By definition, the combined (integrated) probability across all possible values is 1. Hence the area under the probability density function (PDF) will be unity, i.e. if we integrate the function from minus infinity to plus infinity we get a value of one. The role that integration takes in handling continuous probability distributions has very important consequences for the process of performing Bayesian inference.

There are a number of properties we might extract from a probability distribution (or formally the PDF) that provide summary information about the parameter the distribution represents. The most familiar will be a measure of the average value, something we might want to take as our 'best guess' for the parameter, or the 'expected' value. Various other measures exist as well, including the median and mode. We might also be interested in some measure of variability in the parameter value and thus might compute the variance (or standard deviation), or confidence intervals.

Figure 2 shows the Standard Normal (or Gaussian) distribution, which has a mean of 0 and a standard deviation of 1. Gaussian distributions get used a lot in probability modelling partly because they are an appropriate representation of a measurement that is subject to white noise, but also because they are reasonably easy to handle mathematically.

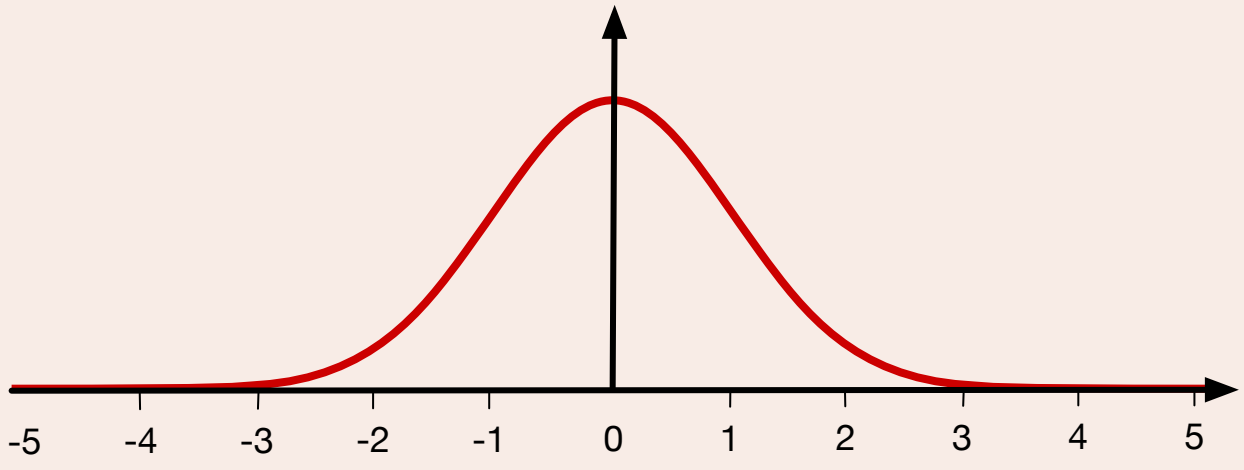


Figure 2 The Standard Normal distribution probability density function. This has a mean of zero and variance (and standard deviation) of one.

If, as is likely, the model has multiple parameters and thus the posterior distribution exists in multiple dimensions we can also extract the distribution for a single parameter (or combination thereof). For example, we might want to concentrate only on the parameters that have a specific neurological or physiological meaning, such as perfusion in Arterial Spin Labelling, and not on other 'nuisance' parameters related to individual physiology or the neuroimaging device, such as the amplitude of the noise. This process is called marginalisation and allows us to extract the distribution of only the parameter(s) we are interested in, taking into account the uncertainty in all of the others, see Box 2.

The result of Bayesian inference is the generation of the posterior distribution, and it is Bayes theory that we need to get us there. This very compactly summarises the whole of Bayesian inference as

$$P(\theta \mid Y) = \frac{P(Y \mid \theta)P(\theta)}{P(Y)} \quad (2)$$

where we meet three terms

$P(\theta)$, the prior distribution. This captures, in the form of a probability distribution, any prior information we have about the parameters 'before the experiment began'.

$P(Y \mid \theta)$, the likelihood. This corresponds to the generative model. Importantly this includes a description of the random process that causes the data to be noisy, hence it is written as a probability distribution[2].

$P(Y)$, the evidence. This is the 'probability of the data'. Remembering that all of the terms in equation 2 are dependent upon the generative model chosen, strictly it is the probability of the data given the model, or $P(Y \mid M)$. We can interpret the evidence as a measure of how well the model describes the data accounting for all possible combinations of parameters in the model, something we will return to in section 6.

***Box 2:** Marginalisation*

The aim of Bayesian inference is to arrive at a probability distribution over all of the parameters in our data generation model, so that we not only have a 'best' estimate of the value, but also a measure of the variability or uncertainty in the values. Often our model will have many parameters, but we are only interested in a subset, or even only one parameter. For example, the amplitude of the noise might be a parameter because it will vary from dataset to dataset, but we often don't care what the specific value is in any given dataset. So typically all we want is the probability distribution associated with the parameters of interest accounting for all possible values of the other parameters. This is achieved by the process of marginalisation and exploits a simple property of a probability distribution: that if you integrate the area under a distribution across the full range of parameter values (from $-\infty$ to $+\infty$) the result will be one. For a distribution that relies upon multiple variables, if we integrate across all possible values for all of the variables, apart from the one we are interested in, we will be left with the distribution for the remaining variable alone: the marginal distribution. This is the same as saying what is the distribution for just this parameter taking into account all the possible values of all of the other parameters. As we will see in section 7 this integral may not be trivial and is where most of the 'effort' in Bayesian inference is expended in practice.

[2] If we are being picky we should say that the likelihood is a probability distribution in Y (the data), but not a probability distribution with respect to θ (the parameters).

The process of applying Bayes theorem is the specification of the generative model and thus the likelihood, and selection of appropriate priors (see Section 4). With that it is possible to obtain the posterior distribution subject to calculation of the evidence term. As the evidence doesn't depend on the parameters in the model, it is simply a value that needs to be calculated to get the posterior distribution scaled correctly. However, the calculation needed is generally very tricky and a range of solutions have been developed that try to get around this computation, something that we will return to in section 7. Often we might try avoid the calculation of the evidence term and only find the posterior distribution 'up to scale', i.e. get the 'shape' of the distribution and not get the scaling correct. However, whilst this avoids the integration associated with getting the scaling correct, we still typically end up needing to do integration to get certain useful statistics such as the mean. Although, this doesn't prevent us from getting some measures, such as the mode (which is unaffected by the scale), something we will return to in section 7.

4 Priors

A major feature of Bayesian inference is the requirement to specify priors on all the parameters in the model. This is often also the biggest source of controversy when Bayesian methods are used, with the criticism being leveled that the imposition of priors biases the results. The response to this is that you cannot infer information from data without making assumptions; e.g., the act of choosing a generative model is one such assumption. Priors in the Bayesian inference framework provide a mathematical way to express assumptions about the parameters and the ability to specify prior assumptions can also offer a substantial advantage in many applications. Specifying priors does, however, offer the opportunity to introduce bias: it is ultimately up to the user to choose them wisely and be aware of the implications of their choices.

You can view the process of Bayesian inference, going from the prior distribution to the posterior distribution, as a method for updating our knowledge of the parameters in the model using some data. Thus the posterior distribution from one analysis might form the prior for a subsequent analysis where we have acquired some more new data[3]. Under this view of Bayesian inference the posterior distribution reflects some combination of the prior knowledge and the new data. If the new data happens to not be very informative, maybe it is very noisy, then the posterior will largely reflect the prior distribution still. If the data contains a lot of information about the parameter in question, maybe the measurement technique is particularly sensitive to a given parameter, then the information from the data (the Likelihood) will dominate. In this case, the likelihood can vastly outweight the prior if the information from the data is far greater than that being provided by the prior. This means that if you set a very informative prior, e.g., a narrow distribution around a chosen value, and your data does not give you much new information, do not be surprised if the posterior 'reverts' to looking the same as the prior. The inference is doing what it is meant to and simply telling you that you haven't gained any new information.

In practice there is also a wide variety of choice of prior distributions that can be made, to express both knowledge and ignorance about particular parameters. There are two categories of particular relevance in neuroimaging.

[3] Don't be tempted to think that you take the posterior from the analysis of some data and use it as a prior in a new analysis with the same data to get a 'better' or more certain answer. It is a remarkably common misconception that this is valid, but it is bad practice to 'reuse' your data in this way.

4.1 Biophysical

The prior distribution captures information we know about the biology, physiology or physics of the system and parameter(s) in question. For example, we might know from other measurements, or the literature more widely, the typical value and range of values associated with a parameter. This type of prior can often be seen as a way of placing a 'soft' constraint on a parameter value in the model. So rather than placing hard-limits outside which the parameter is not allowed to vary, we can instead specify a range that is more probable, whilst still allowing for extreme values if the data demands it. This is something that can be useful in pathological cases, where many of our 'normal' assumptions might be invalid.

4.2 Regularisation priors

The biophysical prior provides some form of regularisation, or constraint, on a parameter based on biophysical knowledge. We can extend the idea further and say that we want to constrain the value of a given parameter in some way based on other information. A classic example for imaging would be that of a 'spatial prior', where we want the prior to reflect the fact that the variation across an image, of the parameter in question, is likely to be smooth. In this case the prior is defined in terms of the parameter values in neighboring voxels, potentially with some scaling for distance, e.g., a Euclidean distance metric. Regularisation is quite commonly used in classical model fitting methods, often included in the cost function that is being minimised with an extra regularisation term[4]. Generally in classic model fitting this term has an unknown weighting factor that also has to be set, or somehow determined from the data. The advantage of the spatial priors is that the influence, or contribution, of the prior is automatically determined as part of the inference process, providing some adaptation of the spatial regularisation with the data. However, like other spatial regularisation methods there is still some form of spatial extent parameter involved that needs to be set. This might be a user choice, although could be included within the inference as an extra (hyper-) parameter to be determined from the data (with its own prior), something that is reasonably natural in Bayesian inference using a hierarchical model, that we will meet in section 5.

Although there is a lot of flexibility when it comes to choice of priors, and thus it is in principle possible to choose a prior that genuinely reflects knowledge (or absence of knowledge) about the parameters, in practice there are a more limited set of 'convenient' priors that people tend to use in practice. Often, convenience means that we choose a form of the prior that makes the implementation of the inference algorithm easier. It is for this reason that Gaussian distributions are often used, even if they can also be justified in many situations because we might expect the underlying distribution of the parameter to be normal. Like all methods, sometime we have to accept compromises to get a workable algorithm, and then be aware what consequences these choices might have on the results.

[4] It is quite common to come across model-fitting or optimization problems that include regularisation terms that call themselves Bayesian. There are good arguments to be made that this is reasonable, as you can view the regularisation term as a sort of prior. However, often these methods have not been derived starting from Bayes theorem and thus do not necessarily exhibit all the properties we attribute to Bayesian inference in this primer appendix. This leads to a debate that we will not pursue further here!

5 Hierarchies

Thus far we have seen how a prior distribution can express prior knowledge, or information, about the parameters in the model so that it is included as part of the inference. These prior distributions will themselves often have parameters, e.g., the mean and standard deviation of a normally distributed prior, which we must choose to reflect our prior knowledge. There may also be situations where these parameters depend on other information and/or we want to infer them from the data. An example of this, that we have already met, is the idea of a spatial prior, where the prior information we have about the parameter value in one voxel depends on the values in the neighbouring voxel. In fact, the idea of values in one voxel depending on values elsewhere in brain, or global parameters associated with the whole brain, or other structures/processes in the brain, is relatively common. In these cases there is a hierarchy of relationships between parameters in the Bayesian inference. Most typically there will be voxelwise parameters (which are usually the ones of interest) that will have priors, whose parameters in turn depend upon a common parameter or parameters. A good example being diffusion tractography, illustrated in Figure 3, where we have a generative model in each voxel that models the diffusion of water in that specific bit of tissue, but this can be constrained by properties of the white matter tract to which the tissue in the voxel belongs. Out ultimate aim in this case is, from the whole brain's worth of data, to identify the white matter tracts, i.e., finding the connections in the brain using diffusion data from all of our individual voxels. We would quite like to do one 'big' inference where we not only examine the data in each voxel, but arrive at the larger-scale properties of the tracts themselves. Using our hierarchical model to relate the data we observe in each voxel to the 'global' parameters associated with the tracts

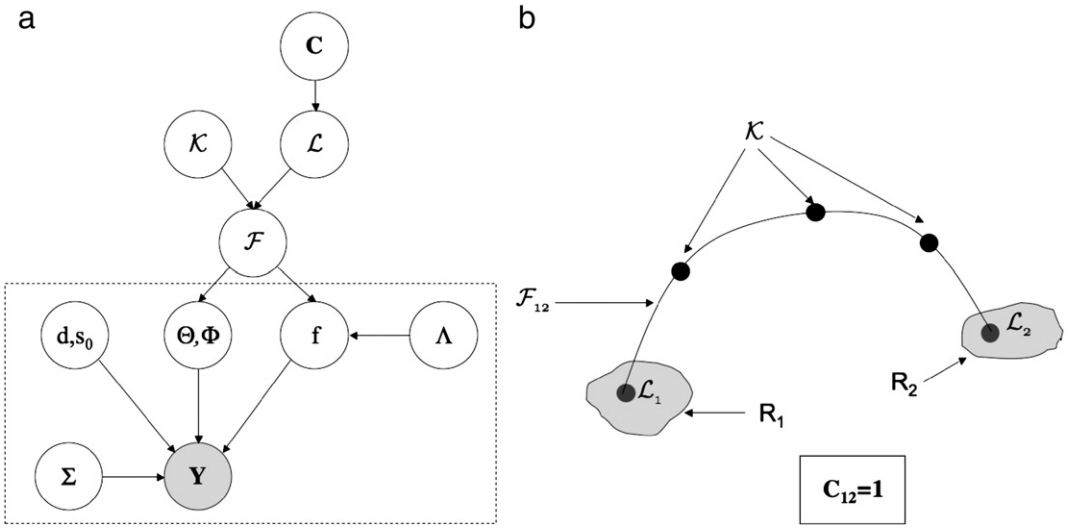


Figure 3: A hierarchical model in use in Bayesian inference for diffusion tractography. (a) The model is composed of a local part that applies in each voxel of the data, within the dashed box, and global constraints. The data, Y, is modelled locally by a combination of parameters that capture water diffusion in the voxel, d and s0, volume fraction, f, and diffusion directional parameters, Θ and Φ, as well as a noise contribution, Σ. The parameters that are to be estimated in each voxel are subject to constraints that arise from global parameters, F, K, L, related to the tract of tissue to which the voxel under consideration belongs. (b) An illustration of a tract connecting regions R1 and R2 which is described by the associated global parameters F, K and L. This figure reproduced with permission from Jbabdi, S., Woolrich, M. W., Andersson, J. L. R., & Behrens, T. E. J. (2007). A Bayesian framework for global tractography. NeuroImage, 37(1), 116-129. http://doi.org/10.1016/j.neuroimage.2007.04.039.

allows us to infer all of the desired properties in a single step. This is in contrast to a classical analysis where you would analyse each voxel in turn (e.g., to find a value for the dominant diffusion direction) and then attempt to trace, going from one voxel to the next, a path associated with a given white matter tract.

6 Model Selection

In section 3 we wrote down Bayes theorem for the estimation of parameters where we already chosen the specific model that we wanted to use to describe the data. Thus the choice of model was an assumption we made. Being more explicit, we could have written out Bayes theorem showing our dependence upon the model

$$P(\theta \mid Y, M) = \frac{P(Y \mid \theta, M)P(\theta, M)}{P(Y, M)} \quad (3)$$

In practice, we might have more than one model that we believe could explain the data we have observed, as quite often there will be simpler and more complex models. Normally we cannot be certain if the more complex model is a good choice, perhaps because we are not certain that the data quality is good enough to tell us all the information we need for the more complex model. Thus we would like some way to compare different models to each other and use this to select the most appropriate one[5].

6.1 Bayes Factors

Up to now we have ignored the denominator in the expression of Bayes theorem: the evidence. But, if you consider equation 3 you will notice that it is a function of the model and the data, but not of the parameters in the model. Thus, the evidence can potentially tell us something about the fit of the model to the data, taking into account all possible values for the parameters in the model. This metric should be a useful way to determine whether our model describes the data well or not and might allow us to compare one model to another. As we noted in section 3, calculation of the evidence term is difficult; thus, using it to evaluate the model fit can present issues. Instead, it is possible to find ways to calculate the relative evidence of one model to another - often called the Bayes factor. This factor can then be used to say whether one model is more favourable than another. However, it is still often difficult to say with confidence that model A is definitely better than model B in general, as it might just be better on this particular dataset and we have no very rigorous way to interpret the absolute value of a Bayes factor to say there is a statistically significant difference.

6.2 Shrinkage Priors

An alternative way of doing model selection, where one model can be expressed as a (simpler) subset of another model, is to use shrinkage priors (also called Automatic Relevancy Determination,

[5] A potential limitation of Bayesian inference is that is depends upon the specification of the model. Thus, it cannot truly handle the situation where the model is flawed, in this situation it could be over-confident in the inferred parameters. Even where we compare various different models, we are still implicitly assuming that one of them can fully represent the data. In practice, at least for neuroimaging applications, this is not generally a cause for concern, but can be an argument in favour of frequentist approaches.

or ARD, in some Bayesian literature). Broadly, this method imposes a prior on a given parameter in the model whose mean is fixed but whose width, e.g., the standard deviation of a normally distributed prior, is to be determined from the data. This uses the idea of hierarchical models from section 5, as the prior has a parameter, the width, to be determined, that itself will have its own prior distribution. Now, if the width of the shrinkage prior is determined from the data to be 'wide', the associated parameter in the model is 'free' and will be determined largely from the data. If the width is determined to be small, the prior shrinks, and the parameter is forced toward the prior mean. If that prior mean is set such that it effectively removes a part of the model (e.g., when the parameter weights a term in the model such that a value of zero means that the contribution of that model term in generating the signal is also zero) this effectively removes that part of the model and thus reduces the complexity: selecting a simpler model. Since this is built in the idea of a hierarchical model, it is perfectly possible to realize such a scheme with Bayesian inference (not withstanding any computational practicalities). The shrinkage prior means that if the data doesn't support the inclusion of the extra component in the model, then it will tend to be removed, rather than simply fitting to noise in the data. Additionally, if the data is particularly noisy then the shrinkage prior will tend to remove the contribution of the extra part of the model if it only slightly modifies the signal (where such changes are swamped by noise), favoring the simpler model.

7 Practicalities

Up to this point in we have concentrated on what you can do with Bayesian inference, but we have avoided talking about the computations needed to apply it in practice. In principle, if we can write down the likelihood from our generative model along with suitable priors then we have all the information we need to find the posterior distribution. Unfortunately, as we noted in section 6, the final step often isn't trivial and often the issue is computing the evidence term, or performing some other integration over the distribution.

Since the evidence term doesn't depend on the parameters and yet those are what we are interested in capturing in the posterior, we might consider instead using a reduced version of Bayes' theorem

$$P(\theta \mid Y) \propto P(Y \mid \theta)P(\theta) \tag{4}$$

Here we accept that we cannot evaluate the evidence, but it is a constant (with respect to the parameters) and thus we can get the posterior up to a scaling factor. All we then need to do is find a way to scale the posterior distribution correctly. Since the posterior is a probability distribution, if we integrate over all of the parameters (from $-\infty$ to $+\infty$) we know we should get a value of one. Thus if we do the integration with the scaled posterior and get a different value, that will tell us the appropriate scale parameter. Here we hit a problem: doing the required integrals for anything but a few simple combinations of models and distributions is not analytically possible[6]. There are only a few relatively simple problems for which an analytic solution for the posterior distribution can be found. For everything else we have to resort to numerical or approximate methods.

It is tempting to ask whether we couldn't avoid doing this integration at all and simply extract the information we want from the scaled posterior. For example, we could potentially search the scaled posterior to find the maximum point and thus the mode of the distribution, which we might treat as our 'best guess' of the parameter values. This is a widely used approach - it is called Maximum A Posteriori (MAP) - but, it only provides a single 'point estimate' for the parameters, and throws away

[6] If you have read the box on marginalisation you will spot that this process is the same as marginalising over all of the parameters.

all the information about uncertainty that we could obtain from the full posterior probability distribution. Additionally, the mode of the distribution is not always a good solution, for example in a bi-modal distribution one of the modes would be completely neglected. Thus, MAP might be seen as a somewhat incomplete form of Bayesian inference[7]. The mode is not the only summary measure we might extract from a probability distribution. We are likely to be interested in the mean, or the median, and the standard deviation or some confidence interval or intervals. Unfortunately, all of these require an integration of the posterior distribution, and thus we cannot avoid doing integration.

7.1 Numerical Methods

We won't, in this short introduction, provide a detailed description of practical methods for Bayesian inference, there are far better (and far longer) books on this to be found in Further Reading. We will simply provide a summary of some of the main type of solutions that you might meet.

Grid evaluation

The simplest numerical solution is simply to calculate the scaled posterior over an evenly spaced grid of parameter values and sum up the values. This comes with two pretty obvious limitations: firstly, how do you evaluate the function over a grid of all possible values (from minus infinity to plus infinity); secondly, as the number of parameters in the model increases, the number of evaluations grows exponentially. In practice, grid evaluation is hopeless for anything other than reasonably trivial problems with very few model parameters.

Numerical integration

Strictly speaking, grid evaluation is a form of numerical integration. There are a range of more efficient methods for numerical integration, that make sensible choices about the density of the samples to be used, and thus reduce the problem to one that is more computationally feasible.

Sampling methods

This is another category that strictly fits within numerical integration. However, unlike the approaches we have considered already, in this case rather than trying to define the points we want to evaluate a priori, we determine the right set as we proceed, attempting to use the scaled posterior to guide us as to regions of the parameter space where we should do most of the evaluations (take the most samples). There are a myriad number of ways of doing this, but one you may well meet is called Markov-Chain Monte Carlo (MCMC), in fact you are most likely to meet Metropolis-Hastings, which is a specific variant of MCMC. The 'Monte Carlo' in the name suggests that we are taking random samples, the Markov chain relates to the statistical properties of the process by which we attempt to explore the parameters (i.e., go from one random sample to another). You are most likely to meet MCMC because it is regarded as the 'gold-standard' numerical Bayesian inference technique, in that it is guaranteed to give you samples from the true posterior distribution. The problem being that this is only guaranteed in the limit of having a very large

[7] Some would say this is a rather generous description.

number of samples and it is largely impossible to determine if you have converged to that limit or not. MCMC proceeds in iterations, called jumps, that produce a string of samples from the posterior distribution which can then be used to numerically normalise it and/or calculate properties such as the mean. Since it relies on convergence, it is common to discard early jumps, the 'burn-in' period. Also to ensure that the samples are truly independent, you may only keep a subset of those produced after the burn-in period. In general, you will probably find that neuroimaging software tools using this method will have made decisions about these aspects of the inference for you, and you play with them at your peril. However, you will probably find that the 'number of jumps' specified will be large (say 5000) and thus the inference won't be very quick! Although these methods can be amenable to parallel computing and thus exploit the specific computing architectures of GPUs (Graphics Process Units).

7.2 Approximations

The numerical methods we have considered so far all make approximations that rely on a set number of samples being able to represent the true distribution, but this inevitably involves a lot of calculations. An alternative is to directly seek to approximate the posterior distribution (which is a function) that we cannot perform integration on, with a function that is easier to work with and that we can integrate more straightforwardly.

The Laplace approximation

This is in essence an extension of the MAP approach we discussed above. After finding the MAP point solution we are still missing information about the variability at that point. What we can then do is to numerically evaluate the variations near the MAP solution (essentially to measure the curvature) and use that to find the parameters of a normal distribution that matches locally. Our final approximate posterior is thus a normal (multivariate Gaussian) distribution centred at the MAP with a width matched to the underlying scaled posterior. From this it is easy to derive the standard deviation or other measures of uncertainty.

Variational methods

Another approach is to specify the functional form of the approximate posterior as a known probability distribution, or a combination of distributions, e.g. the product of a series of distributions. The distributions will be described by some parameters and the correct choice of these parameters will achieve the best approximation of the true posterior. The challenge then is finding a suitable cost function that reflects the errors in the approximation and minimising it. A reasonably popular approach is to minimize the Kullback–Leibler divergence between approximate and true posteriors - as this is a measure of how similar two probability distributions are. In practice, there are still challenges hidden in this approach - for example, the right choice of approximating distribution, given the likelihood function and the priors, to make the cost function evaluation tractable. However, there are various examples of where this approach has been successfully used in popular neuroimaging methods, and with a considerable computational saving over MCMC.

8 Conclusion

The aim of this Short Introduction has been to give an overview of Bayesian Inference, in an attempt to explain why it is so popular in application to analysis of neuroimaging data. As we have seen, at least for neuroimaging, Bayesian Inference provides a mathematical framework for performing estimation that is built on probability theory and distributions. This makes it very convenient for noisy imaging data where very often we have a range of prior knowledge, both in terms of models of the data and our experience of particular parameters, that we want to utilise within the analysis. Whilst Bayesian methods can provide a principled framework for developing algorithms, they can be misused (and have their own flaws). Hopefully, this Short Introduction provides a sufficient information to quip you to look for sensible uses of Bayesian Inference, appreciate some of the choices that have been made that give rise to the resulting algorithms and be able to spot where certain choices could have an effect on the estimates that you need to watch out for. If you find the idea of Bayesian Inference appealing and want to apply it yourself to data analysis, the Further Reading section provides a range of resources both for neuroimaging and more broadly.

FURTHER READING

- MacKay, D. (2003). Information Theory, Inference and Learning Algorithms.Cambridge University Press.
  - This book provides a very good introduction to the concepts of inference, as well as methods of applying it. You might want to follow the advice given in the book and read selected chapters if you are only interested in the sort of inference we have discussed in this Primer Appendix and not other topics of Information Theory. Conveniently you can access a downloadable copy from the website:
  - http://www.inference.org.uk/itila/book.html
- Sivia, D.S. (2006). Data Analysis: A Bayesian Tutorial, 2nd Ed. Oxford University Press.
  - This is a good general technical introduction to using Bayesian Inference for data analysis.
- Woolrich, M. W. & Chappell, M. A. (2015). Bayesian Model Inversion. Brain Mapping (pp. 509-516). Elsevier.
  - A more mathematical introduction to Bayesian Inference than this Short Introduction, but also set in the context of neuroimaging.
  - https://doi.org/10.1016/B978-0-12-397025-1.00325-0
- Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M. A., Makni, S., Behrens, T., et al. (2009). Bayesian analysis of neuroimaging data in FSL. NeuroImage, 45(1, Supplement 1), S173-S186.
  - An introduction/review of Bayesian inference applied to neuroimaging applications as found in the FMRIB Software Library.
  - http://doi.org/10.1016/j.neuroimage.2008.10.055
- Gelman, Carlin, Stern, Dunson, Vehtari & Rubin (2014). Bayesian Data Analysis, 3rd Ed. CRC Press.
  - A very comprehensive text on Bayesian inference, not for the faint hearted.

It is not necessary to understand all of the details of the algorithms used to analyse neuroimaging data. However, some understanding of the concepts that lie behind these algorithms can be helpful to be able to design good experiments, make appropriate analysis choices, and interpret results carefully. In this Short Introduction we will outline the basics of Bayesian Inference, a popular mathematical framework for data analysis and machine learning that is used in many areas of neuroimaging analysis.

This text is one of a number of appendices to the Oxford Neuroimaging Primers, designed to provide extra details and information that someone reading one of the primers might find helpful, but where it is not crucial to the understanding of the main material. This appendix specifically addresses the principles that underpin Bayesian Inference, as it is used in neuroimaging. In it we seek to go into more detail than we might in one of the primers, for those who want to understand more about how Bayesian Inference can be used for data analysis. In turn, this appendix also provides a high level introduction to individuals who are interested in developing their own Bayesian Inference methods, or find they need to select between different methods in a specific application.