Lecture Notes
Computational Cognitive Neuroscience
Introduction

Marcel van Gerven

Learning Goals

In this lecture, we lay the mathematical foundations on which various other lectures build. After studying these notes you should:

1. Understand what is meant by a random variable
2. Recall marginalisation and the product rule
3. Be able to apply Bayes' rule and explain its components
4. Understand ML and MAP estimation
5. Understand the gist of exact inference
6. Understand what is meant by conjugacy
7. Know what is meant by the predictive density and write down its associated formula
8. Know what is meant by Bayesian model comparison and write down the equation for the Bayes factor
9. Be able to outline why Bayesian statistics is important to cognitive (neuro)science

1 Introduction

Bayesian theory provides a theoretically sound mathematical framework for reasoning under uncertainty. With the development of more efficient inference algorithms and exponential growth in computer power, we have witnessed a Bayesian revolution, transforming areas such as astronomy, artificial intelligence, and bioinformatics. The appeal of the Bayesian approach is that it allows inferences to be drawn about unobserved (latent) variables of interest given data and background knowledge at hand.

Especially in cognitive neuroscience, the Bayesian approach is becoming increasingly important. On one hand since, in cognitive neuroscience, the object

of study is a collection of unobserved random processes about which we have ample background knowledge, yet which can only be probed indirectly and sparsely using different data acquisition methods. On the other hand, Bayesian theory has become a key explanatory mechanism for understanding brain function in general. After all, reasoning under uncertainty is exactly the problem that is solved by the brain.

2 Probability Theory

Model specification is often done in the language of *probability theory*. In this section we describe the foundations of probability theory.

Definition 2.1 (Probability space)**.** *A* probability space *consists of three parts:*

1. *A* sample space Ω *which is the set of possible states of the world (outcomes)*
2. *A set of* events $\mathcal{F}$ *where each event is a set containing zero or more outcomes*
3. *A* probability measure $\mathbb{P}$*, which is a function from events to probabilities satisfying the Kolmogorov axioms:*
   (a) $\mathbb{P}(\varnothing) = 0$
   (b) $\mathbb{P}(\Omega) = 1$
   (c) $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *where* $A_1, A_2, \ldots$ *is a collection of disjoint members of* $\mathcal{F}$*.*

Instead of using events, it is convenient to describe the world in terms of random variables.

Definition 2.2 (Random variable)**.** *A* random variable *is a function*

$$X \colon \Omega \to \mathbb{R}$$

from a sample space into the real numbers with the property that $\{\omega \in \Omega \colon X(\omega) \leq x\} \in \mathcal{F}$ *for each* $x \in \mathbb{R}$*.*

Definition 2.3 (Cumulative distribution function)**.** *Each random variable has an associated* (cumulative) distribution function

$$F \colon \mathbb{R} \to [0, 1]$$

given by $F(x) = \mathbb{P}(X \leq x)$*. Here,* $X \leq x$ *is shorthand for the event* $\{\omega \in \Omega \colon X(\omega) \leq x\}$*.*

We distinguish continuous and discrete random variables.

Definition 2.4 (Discrete random variable)**.** *A random variable X is called* discrete *if it takes values in some countable subset $\{x_1, x_2, \ldots\}$, only, of $\mathbb{R}$. X has* (probability) *mass function*

$$P\colon \mathbb{R} \to [0,1]$$

given by $P(x) = \mathbb{P}(X = x)$. It is related to the distribution function via $F(x) = \sum_{i\colon x_i \leq x} P(x_i)$.

Definition 2.5 (Continuous random variable)**.** *A random variable X is called* continuous *if its distribution function can be expressed as*

$$F(x) = \int_{-\infty}^{x} p(u)d(u) \;\; x \in \mathbb{R}$$

for some integrable function $p\colon \mathbb{R} \to [0,\infty]$ called the (probability) density function *of X. This completes the technical definition of discrete and continuous random variables.*

We can also consider random vectors consisting of multiple random variables, which leads to the definition of a joint distribution function.

Definition 2.6 (Joint distribution function)**.** *The joint distribution function of a* random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *on $(\Omega, \mathcal{F}, \mathbb{P})$ is the function*

$$F\colon \mathbb{R}^n \to [0,1]$$

given by $F(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$, where $\mathbf{X} \leq \mathbf{x}$ is shorthand for $X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n$.

This definition can be used to generalize to multivariate probability mass and density functions. In the following, for illustration, we mostly work with probability density functions $p(\mathbf{x})$ defined on continuous random vectors unless indicated otherwise. At this point it becomes convenient to introduce the notion of a conditional probability density function$p(\mathbf{x} \mid \mathbf{y})$ which expresses the probability density function for $\mathbf{x}$ conditional on the observation that $\mathbf{Y} = \mathbf{y}$.

3 Rules of Probability Theory

Probability theory depends on simple rules such as the sum rule and the product rule.

Definition 3.1 (Sum rule)**.** *The* sum rule *states*

$$p(\mathbf{x}) = \int_{-\infty}^{\infty} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}\,.$$

Here, we marginalise *over $\mathbf{y}$ to convert the joint density $p(\mathbf{x}, \mathbf{y})$ into a marginal density $p(\mathbf{x})$. In case of discrete random variables, density functions become mass functions and the integral is replaced by a sum.*

Definition 3.2 (Product rule)**.** *The* product rule *(also called* chain rule*) states that*

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})p(\mathbf{x} \mid \mathbf{y})\,.$$

Here, the joint density is expressed in terms of a marginal density $p(\mathbf{y})$ and a conditional density $p(\mathbf{x} \mid \mathbf{y})$.

4 Bayesian Statistics

Bayesian statistics is concerned with updating our beliefs in the face of uncertainty given observed data. As will be shown, we can use these updated beliefs to address the following three fundamental inferential problems [Kru10]:

1. Parameter estimation
2. Prediction
3. Model comparison

In Bayesian statistics, estimation of parameter values translates into the estimation of a *posterior density* $p(\theta \mid \mathcal{D}, M)$ for parameters of interest θ given observed data $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ and a model M. This posterior summarizes everything we can ever know about θ. Given a *likelihood function* $L(\theta) \equiv p(\mathcal{D} \mid \theta, M)$ and a *prior* $p(\theta \mid M)$, this posterior distribution is given by Bayes' rule:

$$p(\theta \mid \mathcal{D}, M) = \frac{p(\mathcal{D} \mid \theta, M)p(\theta \mid M)}{p(\mathcal{D} \mid M)} \tag{1}$$

where

$$p(\mathcal{D} \mid M) = \int p(\mathcal{D} \mid \theta, M)p(\theta \mid M)d\theta \tag{2}$$

is the normalizing constant. This normalizing constant is also known as the *marginal likelihood* or *evidence*. It plays an important role in Bayesian model comparison. The model M represents the functional form of the prior and likelihood as well as the possible hyper-parameters on which these functional forms depend. We refer to the combination of prior and likelihood as a *generative model*.[1] For convenience, the model M is often suppressed from the notation. For illustration, look at the following concrete example.

Example 4.1. *Consider the situation where you are presented with a test consisting of a series of n true/false questions of equal difficulty. After getting back the test results, you learn that you answered k out of n questions correctly. What is the probability θ of answering a new question correctly?*

We are asked to solve a parameter estimation problem for θ. In order to answer this question, the first step is to formalize the problem by writing down a generative model $p(k \mid n, \theta)p(\theta)$. Note that k plays the role of the data, θ is the

[1] Note that some researchers equate the likelihood with the generative model.

variable of interest, and n is considered a hyper-parameter (we know the number of questions beforehand. The second step is to choose appropriate probability distributions to complete the model.

For the likelihood function, we choose the following formalization in terms of a Binomial distribution:

$$p(k \mid n, \theta) = \text{Binomial}(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \,. \tag{3}$$

The Binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability θ.

For the prior, we choose the following formalization in terms of a Beta distribution:

$$p(\theta \mid \alpha, \beta) = \text{Beta}(\theta \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \,. \tag{4}$$

The Beta distribution is a family of continous distributions defined on the interval [0,1] and parameterized by positive shape parameters α and β. The resulting generative model can equivalently be written as

$$\begin{aligned} k \mid n, \theta &\sim \text{Binomial}(n, \theta) \\ \theta \mid \alpha, \beta &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

where $\sim$ reads *is distributed as*. Given a generative model and observed data, we can compute a posterior probability distribution using *probabilistic inference*.

Updating of a prior $p(\theta)$ to a posterior $p(\theta \mid \mathbf{y})$ is known as probabilistic inference. In general, computing a posterior distribution $p(\theta \mid \mathbf{y})$ is hard. Firstly, due to the evidence term $p(\mathbf{y})$ in Equation (1) which usually means computing a difficult integral. Secondly, because drawing samples from the posterior is challenging, especially in high-dimensional spaces, since there is no obvious way to generate samples without visiting all possible states [Mac03].

5 Point Estimation

An often used simplification is to compute a point estimate of θ rather than a full posterior distribution. Consider again Bayes' rule shown in Equation (1). Note that the evidence does not involve the parameters θ, and is given by a single number that ensures that the area under the posterior distribution equals one. Therefore, Equation (1) is also written as

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta) p(\theta) \tag{5}$$

which states that the posterior is proportional to ($\propto$) the likelihood times the prior. A reasonable point estimate is the maximizer of this quantity:

$$\hat{\theta} = \arg\max_{\theta} \{p(\mathcal{D} \mid \theta) p(\theta)\}$$

which is known as the *maximum a posteriori* (MAP) estimate. The MAP estimate corresponds to the mode of the posterior distribution. It is a convenient choice since it reduces a difficult integration problem to an easier optimization problem. In case we have a flat prior (all values of θ are equally likely), the MAP estimate reduces to the *maximum likelihood* (ML) estimate:

$$\hat{\theta} = \arg\max_{\theta} \{p(\mathcal{D} \mid \theta)\} \ .$$

In order to find the MAP estimate one approach is to use the partial derivatives of $\log\{p(\mathcal{D} \mid \theta)p(\theta)\}$ to find a local optimum (maximizer of $p(\theta \mid \mathcal{D})$). This optimum can sometimes be found by setting (partial) derivatives to zero and solving the resulting equations. Otherwise, gradient ascent algorithms can be used.

Example 5.1. *Consider finding the MAP estimate of θ for the Beta-Binomial model. In this case, we have*

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \left\{ \binom{n}{k} \theta^k (1-\theta)^{n-k} \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \right\} \\ &= \arg\max_{\theta} \left\{ \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \right\} \end{aligned}$$

Taking derivatives, we obtain

$$\begin{aligned} \frac{d}{d\theta}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} &= (k+\alpha-1)\theta^{k+\alpha-2}(1-\theta)^{n-k+\beta-1} - \\ &\quad (n-k+\beta-1)\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-2} \end{aligned}$$

Setting to zero and simplifying, we obtain

$$\hat{\theta} = (k+\alpha-1)/(n+\alpha+\beta-2)\,.$$

Though the MAP and ML estimates are often used for computational convenience, they have their drawbacks. The most obvious drawback is that we have no measure of uncertainty regarding our point estimate $\hat{\theta}$ (see [Mur12, Mac03] for additional drawbacks).

6 Exact Inference

The Beta-Binomial model is one of the models for which an analytical solution is available. The availability of such an analytical solution depends on the notion of *conjugacy*. That is, the prior and likelihood are conjugate when the posterior takes the same functional form as the prior distribution. This also means that we obtain a closed-form expression for the posterior. For the Beta-Binomial model we obtain

$$p(\theta \mid k, n, \alpha, \beta) = \frac{1}{B\left(k+\alpha, n-k+\beta\right)} \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \qquad (6)$$

which is equal to a Beta distribution with shape parameters $k+\alpha$ and $n-k+\beta$.

7 Making predictions

Another key question in Bayesian statistics is how to make optimal predictions. In Bayesian statistics, prediction for a new datapoint $\mathbf{y}$ given previously observed data $\mathcal{D}$ is formalized by the *predictive density*

$$p(\mathbf{y} \mid \mathcal{D}) = \int p(\mathbf{y} \mid \theta) p(\theta \mid \mathcal{D}) d\theta \tag{7}$$

where we take into account uncertainty about the model parameters θ by integrating them out. Often, the predictive density is approximated by replacing the integral by the MAP estimate:

$$p(\mathbf{y} \mid \mathcal{D}) = p(\mathbf{y} \mid \hat{\theta})$$

where $\hat{\theta} = \arg\max_\theta p(\theta \mid \mathcal{D})$.

We again consider an example to illustrate the predictive density.

Example 7.1. *Consider the situation where you have estimated the posterior distribution θ of answering a new question correctly based on known test results. Suppose you are presented with a new test consisting of a series of m true/false questions of equal difficulty. What is your prediction concerning the number of correct outcomes y on this test?*

Making use of Equation (6), the predictive density for this problem is given by

$$p(y \mid m, k, n, \alpha, \beta) \quad = \quad \binom{m}{y} \frac{B(y + k + \alpha, m - y + n - k + \beta)}{B(k + \alpha, n - k + \beta)} .$$

Again, for this particular model the predictive density can be analytically computed, which does not hold in general.

8 Model Comparison

In the preceding section, we have considered prediction under a given model. That is, we used a generative model M with a fixed set of hyper-parameters α and β. However, this begs the question whether M was the correct model to begin with.

In a Bayesian setting, one would actually want to take uncertainty about the model into account and compute the following alternative to the predictive density:

$$p(\mathbf{y} \mid \mathcal{D}) = \sum_{M \in \mathcal{M}} p(\mathbf{y} \mid \mathcal{D}, M) p(M \mid \mathcal{D}) \tag{8}$$

where $\mathcal{M}$ is the set of all possible models. This procedure is known as *Bayesian model averaging* [HMRV99]. Note that this is an expensive procedure since we need to compute the predictive density $p(\mathbf{y} \mid \mathcal{D}, M)$ for all possible models.

Rather than summing over models for the purpose of prediction, we might be interested in identifying an optimal model, given by

$$\hat{M} = \arg\max_{M \in \mathcal{M}} p(M \mid \mathcal{D}) \,. \tag{9}$$

It is hard to consider all possible models and in practice one often deals with selecting one of two competing hypotheses M_1 and M_2 on how the data is generated. In that case, we can make a relative statement about which model is more likely by computing the *Bayes factor*:

$$K = \frac{p(M_1 \mid \mathcal{D})}{p(M_2 \mid \mathcal{D})} = \frac{p(\mathcal{D})^{-1} p(\mathcal{D} \mid M_1) p(M_1)}{p(\mathcal{D})^{-1} p(\mathcal{D} \mid M_2) p(M_2)} = \frac{p(\mathcal{D} \mid M_1)}{p(\mathcal{D} \mid M_2)} \tag{10}$$

where we made use of the assumption that all models are equally likely a priori. This result shows that the question which model to select can be addressed using the model evidence $p(\mathbf{y} \mid M)$ as outlined in Equation (2). A Bayes factor of $K > 1$ means that model M_1 is more strongly supported by the data than model M_2. Kass and Raferty [KR95] provide the following scale for the interpretation of K:

K	**Evidence in favour of M_1**
1 to 3	Barely worth mentioning
3 to 20	Positive
20 to 150	Strong
>150	Very Strong

It is important to note that Bayesian model comparison automatically incorporates *Occam's razor*, which states that simple explanations of the data are preferred over complex explanations of the data.

9 Making decisions

Inference as such is not of much use when it is not accompanied by decisions. E.g., model comparison entails committing to a hypothesis, predictions entail the selection of an action, et cetera. This is the realm of decision theory. Define a *loss function*

$$L(\theta, a(\mathbf{x}))$$

which quantifies the loss when θ parameterizes the true model (state of the world) and $a(\mathbf{x})$ is the action performed based on observations x.

Bayesian decision theory [Rob07] states that the optimal decision rule given an observation $\mathbf{x}$ is the one which minimizes the posterior expected loss:

$$\mathbb{E}\left[L(\theta, a(\mathbf{x})) \mid \mathbf{x}\right] = \int L(\theta, a(\mathbf{x})) p(\theta \mid \mathbf{x}) d\theta$$

Extension to decision making over time leads to *optimal control theory*. Optimal control theory lays the groundwork for theories of animal learning and motor control.

10 Relevance for Cognitive (Neuro)science

I have given a very brief overview of some of the essential ingredients underlying Bayesian statistics. Bayesian statistics is of relevance to cognitive (neuro)science in the following ways:

1. It provides the ingredients for building computational models that can deal with uncertainty. For example, when modeling how visual input leads to neuronal activity, we need to take into account sensory noise.

2. It provides a principled approach to data analysis. For example, when we try to estimate how brain networks are anatomically connected using diffusion imaging (a magnetic resonance imaging technique), we need to take into account that the observed data is affected by various sources of noise.

3. It allows us to make optimal predictions. For example in the context of real-time functional magnetic resonance imaging or in brain-computer interface research.

4. It provides a normative account of what the brain is doing. That is, the brain essentially tries to infer the causes of its sensory input in order to make more optimal decisions. This is known as the Bayesian brain hypothesis and relies on the three fundamental inferential problems: parameter estimation, prediction and model comparison.

11 Further Reading

A good mathematical textbook on the foundations of probablity theory is [GS01]. Further details on Bayesian statistics can be found in various textbooks [Kru10, Bar13, Mur12, Bis07, Mac03, LW14]. Bayesian Data Analysis is particularly recommended [GCS+13]. For an extensive introduction to probability theory and Bayesian inference, read Chapters 2 and 3 of [Mac03]. Chapter 23 of [Mac03] gives an overview of useful probability distributions. For more details on model comparison, consult Chapter 28 of [Mac03]. A good overview of Bayesian decision theory is given by [Rob07]. The MacKay book is freely available for download and is highly recommended. If you want to learn more about machine learning in neuroscience, consult e.g. [Hel15, Mar15, FVK+14].

References

[Bar13] David Barber. *Bayesian Reasoning and Machine Learning*. 2013.

[Bis07] C Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[FVK+14] Jeremy Freeman, Nikita Vladimirov, Takashi Kawashima, Yu Mu, Nicholas J Sofroniew, Davis V Bennett, Joshua Rosen, Chao-Tsung Yang, Loren L Looger, and Misha B Ahrens. Mapping brain activity at scale with cluster computing. *Nature Methods*, 11:941–950, July 2014.

[GCS+13] A Gelman, J B Carlin, H S Stern, DB Dunson, A Vehtari, and D B Rubin. *Bayesian Data Analysis*. Chapman \& Hall/CRC, 2013.

[GS01] GR Grimmett and DR Stirzaker. *Probability and random processes*. Oxford University Press, 3rd edition, 2001.

[Hel15] Moritz Helmstaedter. NeuroView The mutual inspirations of machine learning and neuroscience. *Neuron*, 86(1):25–28, 2015.

[HMRV99] JA A Hoeting, D Madigan, AE E Raftery, and CT T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, 14(4):382–401, 1999.

[KR95] Robert E Kass and Adrian E Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[Kru10] J K Kruschke. *Doing Bayesian Data Analysis*. Academic Press, 2010.

[LW14] M D Lee and EJ Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2014.

[Mac03] D J C MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[Mar15] Eve Marder. Understanding brains: Details, intuition, and big data. *PLOS Biology*, 13(5):e1002147, 2015.

[Mur12] K P Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[Rob07] C P Robert. *The Bayesian Choice*. Springer, 2007.