Bayesian statistics and modelling

Title: Bayesian Statistics and Modelling

Authors: Rens van de Schoot[1]*, Sarah Depaoli[2], Andrew Gelman[3], Ruth King[4], Bianca Kramer[5], Kaspar Märtens[6], Mahlet G. Tadesse[7], Marina Vannucci[8], Duco Veen[1], Joukje Willemsen[1], Christopher Yau[9, 10]

Affiliations

[1] Department of Methods and Statistics, Utrecht University, Utrecht, The Netherlands

[2] Department of Quantitative Psychology, University of California Merced, Merced, CA, USA

[3] Department of Statistics, Columbia University, New York, USA

[4] School of Mathematics, University of Edinburgh, Edinburgh, UK

[5] Utrecht University Library, Utrecht University, Utrecht, The Netherlands

[6] Department of Statistics, University of Oxford, Oxford, UK

[7] Department of Mathematics and Statistics, Georgetown University, Washington DC, USA

[8] Department of Statistics, Rice University, Houston, TX, USA

[9] Division of Informatics, Imaging & Data Sciences, University of Manchester, Manchester, UK

[10] The Alan Turing Institute, British Library, 96 Euston Road, London

Corresponding author: Rens van de Schoot: Department of Methods and Statistics, Utrecht University, P.O. Box 80.140, 3508TC, Utrecht, The Netherlands; Tel.: +31 302534468; E-mail address: a.g.j.vandeschoot@uu.nl.

Acknowledgements [AU: do any of the other authors want to add funding information?]
The first author (RvdS) was supported by a grant from the Netherlands organization for scientific research: NWO-VIDI-452-14-006. RK was supported by a Leverhulme research fellowship grant reference RF-2019-299.

Author contributions
Introduction (R.v.d.S., A.G.); Experimentation (R.v.d.S., S.D., J.W.); Results (R.v.d.S., R.K., M.G.T., M.V., D.V., K.M., C.Y.); Applications (S.D., R.K., K.M., M.G.T., M.V., C.Y.); Reproducibility and data deposition (B.K., D.V., S.D., R.v.d.S.); Limitations and optimizations (A.G.); Outlook (K.M., C.Y.); Overview of the Primer (R.v.d.S.).

Competing interests
The authors declare no competing interests.

ORCID :

RvdS: https://orcid.org/0000-0001-7736-2091

SD: https://orcid.org/0000-0002-1277-0462

AG: https://orcid.org/0000-0002-6975-2601

RK: https://orcid.org/0000-0002-5174-8727

BK: https://orcid.org/0000-0002-5965-6560

KM: https://orcid.org/0000-0002-7631-727X

MGT: https://orcid.org/0000-0003-2671-1663

MV: https://orcid.org/0000-0002-7360-5321

DV: https://orcid.org/0000-0002-8352-7574

JW: https://orcid.org/0000-0002-7260-0828

CY: https://orcid.org/0000-0001-7615-8523

Abstract

Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' Theorem. This Primer describes the stages involved in Bayesian analysis, from specifying the prior and data models, to deriving inference, model checking and refinement. Bayesian analysis has been successfully employed across a variety of research fields, including social sciences, ecology, genetics, medicine, and more. We discuss these applications and propose strategies for reproducibility and reporting standards. Finally, we outline the impact of Bayesian analysis in artificial intelligence, a major goal in the next decade.

[H1] Introduction

It all started with an essay written by Reverend Thomas Bayes, published by Richard Price[1], on inverse probability: how to determine the probability of a future event solely based on past events? It was Pierre Simon Laplace[2] who actually published the theorem we now know as Bayes' theorem (Box 1). The typical Bayesian workflow consists of three main steps (Figure 1). (1) The first ingredient has to do with knowledge available about the parameter in a statistical model without the data itself and is captured in the so-called prior distribution **[G]**. (2) The second ingredient is the information about the same parameters in the data; it is the observed evidence expressed in terms of the likelihood function **[G]** of the data given the parameters. Both prior distribution and likelihood function are combined via Bayes' Theorem and are summarized by (3) the so-called posterior distribution **[G]**, which is a compromise of the prior knowledge and the observed evidence. This joint distribution is also called a generative model. The posterior distribution reflects one's updated knowledge, balancing prior knowledge with observed data.

Although the idea of inverse probability and Bayes' theorem have been longstanding within mathematics, these tools have only become prominent in applied statistics in the past fifty years [3-10]. There are many reasons for using Bayesian methods: Sometimes researchers may be "forced into" the use of Bayes' theorem some models, for example mixture or multilevel models, require Bayesian methods to improve convergence issues[11], exact quantification of uncertainty, aid in model identification[12], produce more accurate parameter estimates[13], data augmentation or data fusion. We will describe much more advantages and disadvantages throughout the manuscript.

The goal of this primer is to provide an overview of the current and future use of Bayesian statistics across different fields of science and to provide an overview of literature that can be used for further study. Moreover, we use many examples how to actually implement a Bayesian model on real data, with all data and code is available for teaching purposes. We aim at a broader group of quantitative researchers working in science-related areas with at least some knowledge of regression modelling. In order to keep the current paper as general as possible with respect to implementing Bayesian methods, there are several concepts listed in Figure 1 that we will be focusing on, like priors and posteriors, and several that we will not specifically address, see the left part in the Figure. We also only briefly touch upon topics like model averaging, network analyses, utility functions/ loss functions without giving a full introduction and do not discuss topics like nonparametric methods. For the non-Bayesian parts we do not discuss we refer the interested reader to classical textbooks.[14,15] This Primer discusses the general framework, algorithms, and a Bayesian research cycle with a special focus on prior specifications (Experimentation). We discuss model fitting, a thorough example of variable selection and we provide an example calculation with posterior predictive checking (Results). Then, we describe how Bayesian statistics is being used in different fields of science (Applications), followed by guidelines for data sharing, reproducibility, and reporting standards (Reproducibility and Data Deposition). We conclude with a discussion of avoiding bias with incorrect models (Limitations and optimizations), and provide a look into the future with Bayesian Artificial Intelligence (Outlook).

[H1]Experimentation

There are several main issues included in this section. First, prior distributions are detailed, highlighting different levels of informativeness (informative, weakly informative, and diffuse priors). The selection of priors is often viewed as one of the more important choices that a researcher makes when implementing Bayesian methods since the priors can have a substantial impact on final model results. This is followed by a description of the prior predictive checking process, which can be used to assess whether the prior settings being implemented are viable. This section concludes with a description of how to determine the likelihood, which is combined with the prior to form the posterior. Given the important roles that the prior and the likelihood have in determining the posterior, it is imperative that prior and model selection be conducted with care.

H3: An Empirical Example - Predicting PhD delays

To illustrate many aspects of Bayesian statistics we provide an example based on real-life data. Note that we simplified the statistical model and the results are only meant for instructional purposes. Instructions for running the code is available for different software including additional data exploration steps[16]. Consider an empirical example of a study predicting PhD delays[17] in which the researchers asked 333 PhD recipients in The Netherlands how long it had taken them to finish their PhD thesis. Based on this information they computed the amount of delay as defined as the difference between planned and actual project time in months ($M = 9.97$, $min/max = -31/91$, $SD = 14.43$). Suppose we are interested in predicting PhD delay ($\boldsymbol{y}$) using a simple regression model, $\boldsymbol{y} = \beta_{age} + \beta_{age^2}. + \boldsymbol{\varepsilon}$, with age (in years) as a predictor, denoted by β_{age}, and we expect this relation to be quadratic, denoted by β_{age^2}. Also, the model contains an intercept, $\beta_{intercept}$ and we assume the residuals, $\boldsymbol{\varepsilon}$, are normally distributed with an unknown variance, σ_{ε}^2.We will refer to this example throughout the following sections to illustrate key concepts.

[H2]Formalizing Prior Distributions

Prior distributions– play a defining role in Bayesian statistics. Prior distributions, or priors, can come in many different distributional forms such as a normal, uniform, Poisson distribution, among others, see also the section Variable Selection for some examples of, so-called, Shrinkage priors. They can also represent different levels of informativeness; the information reflected in a prior distribution can be anywhere along the continuum of complete uncertainty to relative certainty. Although it is important to remember that priors can fall along this continuum, there are three main classifications of priors that are often used in the literature to capture the degree of (un)certainty surrounding the population parameter value: (1) informative, (2) weakly informative, and (3) diffuse. These classifications can be made based on the researcher's personal judgment. For example, a normal prior with a variance of 1000 may be considered diffuse in one setting and informative in another—it depends on the values of the parameter, as well as parameterization or scaling for the parameter.

Figure 2 illustrates the relationship between the likelihood, prior, and posterior for different prior settings for β_{age}. In this figure, the first column represents the prior distribution, which is normally distributed for the sake of this example. Notice that there are five different rows of priors, representing different prior settings (some varying in the level of informativeness. The second column represents the likelihood. The prior and the likelihood form together to create the posterior according to Bayes' rule. The third column

illustrates the prior, likelihood, and the resulting posterior, which is derived for illustrative purposes in the current section. In the next section Results we demonstrate how to obtain the posterior.

The individual parameters that control the amount of (un)certainty in the priors are called hyperparameters **[G]**. Take the normal distribution as an example. This distribution is defined by a mean and a variance which are the hyperparameters for the normal prior, and we can write this distribution as: $N(\mu_0, \sigma_0^2)$, where the hyperparameters represent the mean (μ_0) and variance (σ_0^2) for the prior, respectively. If the variance is relatively large, then it represents more uncertainty surrounding the mean, vice versa. For example, Figure 2 illustrates five prior settings in the first column with different values for μ_0 and σ_0^2. The diffuse and weakly informative priors (first three rows) show more spread, that is, a larger variance, compared to the informative priors (last two rows). The mean hyperparameter can be seen as the peak in the distribution.

An *informative prior* **[G]** is one that reflects a high degree of certainty surrounding the population parameter. Specifically, the hyperparameters for these priors are specified to express particular information reflecting a greater degree of certainty about the model parameters being estimated. In the case of a normal probability distribution, this would indicate that the prior would have a very small, or narrowed, variance. A researcher may want to use an informative prior when existing information suggests restrictions on the viable range of a particular parameter, or a relationship between parameters, like a positive but imperfect (population) correlation between susceptibility to various medical problem[18,19]. The information embedded in the informative prior can come from a variety of places, which is referred to as prior elicitation. Strategies for prior elicitation can be to ask an expert or a panel of experts to provide an estimate for the hyperparameters based on knowledge of the field[20-23], use the results of a previous publication or meta-analysis[24,25], or a combination thereof[26]. Consider the prior $\beta_{age} \sim N(2.5, 5)$, which was derived from a ShinyApp containing a visualization of how the different priors interact[27].

Finally, another method that can be used for prior elicitation involves implementing data-based priors, which are derived based on a variety of methods including maximum likelihood[28-31] or sample statistics[32-34]. Although data-based priors are relatively common, we do not recommend use of so-called "double-dipping" procedures, where estimation occurs based on the sample data and then results are used to to derive priors implemented (with the same sample data) for final model estimation. We refer the reader elsewhere[32] for more details on this topic. Instead, a hierarchical modelling strategy can be implemented, where priors can depend on hyperparameter values that are data-driven, for example sample statistics pulled from the data, thus avoiding the direct problems linked to "double-dipping." In some cases, an informative prior can produce a posterior that is not reflective of the population model parameter. There are circumstances when informative priors are needed, but it is also important to assess the impact these priors have on the posterior through a sensitivity analysis as discussed below.

A *weakly informative prior* **[G]** is typically not too diffuse, and it is not too restrictive either. In the case of a normal prior, a weakly informative prior would have a variance hyperparameter that exhibits wider

variance compared to an informative prior. Such priors will have a small impact on the posterior, depending on the scale of the variables, and the posterior results are still data driven.

Some researchers find this to be a nice middle ground regarding the informativeness of the prior. A researcher may want to use a weakly informative prior when some information is assumed about a parameter, but there is still a desired degree of uncertainty. For example, a weakly informative normal prior for the regression coefficient could allow 95% of the prior density mass to fall within values between -10 and 10 or between 0 and 10, see the two different examples in Figure 2, respectively. Essentially, weakly-informative priors do not supply any strict information, but yet are still strong enough to avoid inappropriate inferences that can be produced from a diffuse prior[35,36]. For this purpose a plausible parameter space should be specified capturing a range of plausible parameter values that is considered to be a reasonable range, thereby excluding improbable values and attaining only a limited density mass to implausible values. For example, if a regression coefficient is known to be near zero, then a weakly informative prior can be specified to reduce the plausible range between, for example, ±5. This prior would reduce the probability of observing out-of-bound values (e.g., a regression coefficient of 100) without being too informative.

Finally, *diffuse priors* **[G]** reflect a great deal of uncertainty about the model parameter. This form of priors represents a decision to not include knowledge about the value of the parameter being estimated. Such a prior would be represented by a distribution with a relatively flat density (Figure 2). A researcher may want to use a diffuse prior when there is a complete lack of certainty surrounding the parameter. In this case, the data will largely determine the posterior. Sometimes researchers will use the term "non-informative prior" as a synonym to "diffuse"[37]. However, we refrain from using this term because we argue that even a completely flat prior, for example, a so-called Jeffreys prior[38], is still providing information about the degree of uncertainty[39]. Therefore, no prior is really non-informative. Diffuse priors can be useful for expressing a complete lack of certainty surrounding parameters, but they can also have unintended consequences on the posterior[40]. For example, diffuse priors can have an adverse impact on parameter estimates via the posterior when sample sizes are small, especially under complex modelling situations involving meta-analytic models[41], logistic regression models[39], or mixture models[13]. In addition, improper priors are sometimes used with the intention of using them as diffuse priors. Although improper priors are common, and they can be implemented with relative ease within a variety of Bayesian programs, it is important to note that improper priors can lead to improper posteriors. We mention this caveat here because obtaining an improper posterior can impact the degree to which results can be substantively interpreted. Overall, we note that a diffuse prior can be used as a placeholder, in the same way that we might start with a simple statistical model with the intent to improve it as necessary. It may be that future analyses (e.g., with subsequent data) are conducted with more informative priors.

Overall, there is no right or wrong prior setting. Many times, diffuse priors can produce results that are aligned with the likelihood, whereas sometimes inaccurate (e.g., biased) results can be obtained with relatively flat priors[13]. Likewise, and as described above in the context of informative priors, an informative prior that is not centered in the same place as the likelihood can pull the posterior away from the likelihood. Because there can be an unintended impact of the priors - despite the level of informativeness - it is always important to conduct a prior sensitivity analysis in order to fully understand the influence that the prior settings have on posterior estimates. Especially when sample size is small, Bayesian

estimation with mildly informative priors is often used[9,42,43], but the prior specification might have a huge effect on the posterior results.

In addition, it is important to note that when priors do not conform with the likelihood, it is not necessarily evidence that there is an issue with the prior. It may be that the likelihood is at fault due to a misspecified model or biased data. In turn, the difference between the prior and the likelihood may be reflective of variation that is not captured by the prior or likelihood alone. These issues can be identified through a sensitivity analysis of the likelihood - for example, by modifying the model - in order to assess how the priors and the likelihood align.

Although it is important to distinguish between these different types of priors, there is an overarching issue that needs addressing. We would like to conclude this section with a final thought about the impact of priors. It is common for critics of Bayesian methods to point toward the subjectivity of priors as a potential downfall of the approach. We argue two distinct points here. First, many elements of the estimation process are subjective, including the model itself or the error assumptions. To place the notion of subjectivity solely on the priors is a misleading distraction from the fact that many other elements in the process are inherently subjective by nature. Second, priors are not necessarily a point of subjectivity. They can be used as tools to allow for data-informed shrinkage, enact regularization, or influence algorithms toward a likely high-density region and improve estimation efficiency. In turn, priors are typically defined through previous beliefs, information, or knowledge. Although beliefs can be characterized as subjective points of view from the researcher, information is typically defined as being outside of the researcher and something that can be rigorously quantified, and knowledge can be defined as objective and consensus-based. Therefore, we urge the reader to consider priors in this broader sense, and not simply as a means of incorporating subjectivity into the estimation process.

Lastly, the current section on informative, weakly informative, and diffuse priors was written in a general sense in that these terms can be used to help define univariate and multivariate priors. The majority of discussion presented in the current paper surrounds univariate priors placed on individual model parameters. However, these concepts can be extended to the multivariate sense, where priors are placed on, for example, an entire covariance matrix rather than a single element from a matrix. For more information on multivariate priors, see[44,45].

[H2]Prior Predictive Checking

Because the inference based on a Bayesian analysis is subject to the "correctness" of the prior, it is of importance to carefully check whether the specified model can be considered to be generating the actual data[46,47]. Note that priors are based on background knowledge and cannot be inherently wrong if the prior elicitation procedure is valid. There is an extensive history of expert elicitation across many different disciplines. MATCH[48] is a generic elicitation tool, but many elicitation problems require custom elicitation procedures and tools, see for instance[49-53] as examples of elicitation procedures designed for specific models. For an abundance of elicitation examples and methods, see the data base of over 67,000 elicited judgements[54], or the following collections[20,55,56]. However, even in case of a valid prior elicitation procedure, it is extremely important to understand the exact specification of the priors. This holds

especially for smaller sample sizes *in relation to* the complexity of the model, for numerous examples[9] In the case of smaller sample sizes, priors will exhibit a strong influence on the posteriors. The step of prior prediction is an exercise to improve the understanding of the priors specified and not a method for changing the original prior, unless the prior explicitly generates data that are incorrect.

Box[57] suggested deriving a prior predictive distribution **[G]** from the specified prior**.** The prior predictive distribution is a distribution of all possible samples that could occur if the model is true. In theory, a "correct" prior provides a prior predictive distribution similar to the true data generating distribution[46]. The prior predictive checking approach compares the observed data to the prior predictive distribution, and checks their compatibility[47]. The compatibility can be summarized by a p-value, describing how far out in the tails of the reference prior predictive distribution the observed data lie[58]. When the prior predictive-value **[G]** is "small", say 0.05, it would indicate that the observed data is unlikely to be generated by the model, and thus call it into question[47]. Evans and Moshonov[59] suggested restricting the approach pof Box to minimal sufficient statistics, i.e. statistics that are as efficient as possible in relaying information about the value of a certain parameter from a sample[60].

Young and Pettit[61] argue that measures being based on a tail area, such as the approaches of Box and Evans and Moshonov, do not produce the required behaviour; favouring the more precise prior if two priors are both specified at the correct value. They propose to use a Bayes factor **[G]** [62] to compare two priors, see also Box 3. All aforementioned methods leave the determination of the existence of prior-data conflict up to debate depending on an arbitrary cut-off value. The data agreement criterion[63] tries to resolve this issue by introducing a clear classification of prior-data conflict, removing the subjective element of the decision[64]. This is done at the expense of selecting an arbitrary divergence based criterion. An alternative has been developed[65] which computes whether the distance is surprising in relation to the expert's prior predictive distribution, see for a comparison of both criterion Lek et al[66]

H3: An Empirical Example - Predicting PhD delays - continued

Prior predictive checks can help prevent mistakes from being made. For instance, various software packages can notate the same distribution differently. The normal distribution can be specified by the hyperparameters mean and variance, mean and standard deviation or mean and precision. The precision is the inverse of the variance. For the last prior shown in Figure 2, we have mis-specified the prior variance, that is instead of using a variance of 5 we mis-specified the variance and used the inverse of the variance (i.e., a precision) instead (1/5=0.2), $\beta_{age} \sim N(2.5, 0.2)$. If a user is not aware of such differences, a prior which was intended to be weakly informative can easily turn into an informative prior distribution. The prior predictive checks in Figure 3 help to avoid misspecifications like this. Panel A displays a scenario in which precision was mistakenly used instead of variance for β_{age} , and displays an unexpected pattern for the prior predictive distribution. Panel B shows reasonable results for the prior predictive distribution for the correct implementation of the hyperparameters. Additionally, in panel C, the kernel density estimate (i.e., the estimate of the probability density function) **[G]**[67] of the observed data is displayed (y - in dark blue) which fall neatly in the distribution of the simulated data (y_{rep} - in light blue). The kernel densities for the prior predictive data are based on combinations of possible values of the different priors. Because of the combinations of uncertainty in the priors, the prior predictive kernel density estimates can be quite different from the observed data. The main focus for Panel C is to check

that the prior predictive kernel distributions are not order-of-magnitudes different from the observed data.

The scripts to reproduce the results are available at the Open Science Framework: https://osf.io/ja859/ DOI 10.17605/OSF.IO/JA859. Note that in this example the prior predictive distribution and the data are compared on the test statistics mean and standard deviation(sd). It is common to desire descent prior predictive performance on these simple statistics at least. The test statistic can however be chosen to reflect important characteristics of the data, e.g. skewness. It is common to desire descent prior predictive performance on these simple statistics at least. The test statistic can however be chosen to reflect important characteristics of the data, e.g. skewness.

[H2] Determining the Likelihood Function

The likelihood, which is used in both Bayesian and frequentist inference [68], is the conditional probability distribution $p(y|\theta)$ of the data y given parameters θ. In Bayesian inference, the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ comes into the posterior as a function of $\boldsymbol{\theta}$ for observed data $\boldsymbol{y}$. The likelihood function summarises the information of the following elements: a statistical model that stochastically generates all the data, a range of possible values for $\boldsymbol{\theta}$, and the observed data. In a Bayesian model, the likelihood function is part of the generative model, the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\theta}$. Because the concept of likelihood is not specific to Bayesian methods, we do not provide a more elaborate introduction of the statistical concept here. Instead, the interested reader is directed to the paper by Etz[69] for an introduction of how likelihood underlies common frequentist and Bayesian statistical methods and to the work of Pawitan[70] for a complete mathematical explanation on this topic.

Much of the discussion surrounding Bayesian inference focuses on the choice of priors, and there is a vast literature on potential defaults[71,72] The inclusion of prior knowledge in the form of a prior is the most noticeable difference between frequentist and Bayesian methods and a source of controversy. However, as argued by Gelman, Simpson and Betancourt[71], a prior can in general only be interpreted in the context of the likelihood with which it will be paired. The importance of the likelihood often gets left out of the discussion, even though the specified model for the data - instantiated by the likelihood function – is the foundation for the analysis[73].

In some cases, specifying a likelihood function can be very straightforward, see Box 2 for an example. However, in practice the underlying data-generating model is not always known. Researchers often naively choose a certain distribution out of habit or because they cannot change it (easily) in the software. The choice of the statistical data-generating model is subjective (based on background knowledge) and should therefore be well understood and described in detail. Robustness checks should be performed to verify the influence of the choice of the likelihood function on the posterior results[72]. Although most research in the theory of Bayesian robustness has concerned the sensitivity of the posterior to imprecision solely in the prior, a few contributions have focussed on the problem of robustness with respect to the likelihood, see for instance[74-76] and references therein.

[H1] Results

After specifying the prior and the likelihood, in this section we assume the data has been collected and we describe the posterior parts of Figure 1. That is, we explain how a model can be fitted to data with the goal of obtaining a posterior distribution, how to select variables, and why posterior predictive checking would be needed. In practice, model building is an iterative process. Any Bayesian model (which includes both the prior distribution and the probability model for data given parameters, which serves also as the likelihood function) can be viewed as a placeholder which can later be improved, in response to the availability of new data, lack of fit to existing data, or simply a process of refinement of the model. Box[57], Rubin[77], and chapter 6 of Gelman et al.[73] discuss the fluidity of Bayesian model building, inference, diagnostics, and model improvement.

[H2] Model Fitting

Once the general model structure has been formulated to describe the data, and the associated likelihood function derived, the next step is to fit the model to the observed data to estimate the model parameters. Although the statistical models necessarily simplify reality, they aim to capture the main processes driving the data. Models may differ substantially in their complexity, taking into account the different mechanisms acting on the system and sources of stochasticity and variability. Some examples of the types of data and associated models are provided in *Applications*. Fitting the models to the observed data permits the estimation of the model parameters, or functions of these, leading to an improved understanding of the system, and associated underlying factors via relevant interpretable quantities given the data.

There are two main paradigms for model fitting and parameter estimation: Bayesian and frequentist. These approaches differ fundamentally. Within the Bayesian framework probabilities are assigned to the model parameters, describing the associated uncertainties; whereas the frequentist framework focuses on the expected long-term outcomes of an experiment. The corresponding implications is that frequentist methods focus on producing a single point estimate for each model parameter, such as the maximum likelihood estimate, (with an associated uncertainty interval: the confidence interval); whereas in Bayesian statistics, the focus is on estimating the entire posterior distribution of the model parameters. This posterior distribution of often summarised, for simplicity, via associated point estimates (such as the posterior mean or median) and an interval estimate in the form of a credible interval (i.e. an interval that contains a given % of the posterior distribution). Direct inference on the posterior distribution is typically not possible as the mathematical equation describing the posterior distribution is typically both high-dimensional (the number of dimensions is equal to the number of parameters) and of a very complex form. In particular, the expression for the posterior distribution is typically only known up to a constant of proportionality, with the denominator expressible as a function of only the data, where this function is not available in closed form but expressible as an analytically intractable integral. We note that this intractability of the posterior distribution was the primary practical reason that Bayesian statistics was discarded by many scientists for the alternative frequentist statistics. However, the seminal paper by Gelfand and Smith[78] transformed the data analytic world, describing how Markov chain Monte Carlo (MCMC) **[G]**, a technique for sampling from a probability distribution, can be used to fit models to data within the Bayesian paradigm. [79]

MCMC is able to indirectly obtain inference on the posterior distribution via simulation[79]. In particular, MCMC permits a set of sampled parameter values of arbitrary size to be obtained from the posterior distribution of interest, despite the posterior distribution being high dimensional and only known up to a constant of proportionality. These sample values are used to obtain empirical estimates of the posterior distribution of interest, which can be estimated up to the desired accuracy by increasing the number of sampled parameter values, if necessary. We note that due to the high dimensionality of the posterior distribution it is often useful to focus on the marginal posterior distribution of each parameter, defined by marginalising (or integrating) out over the other parameters (i.e. dimensions). Marginal distributions are useful for focusing on individual parameters but by definition do not provide any information on the relationship between the parameters.

Whilst MCMC is the most common algorithm used in Bayesian analyses, there are other model-fitting algorithms, see Table 1 for a non-exhaustive overview of MCMC techniques of sampling and approximation techniques. We refer the interested reader for running the PhD-example with different estimators to[80,81]. In this article for posterior inference, we focus on MCMC which combines two concepts: (i) obtain a set of parameter values from the posterior distribution (using the Markov chain **[G]** , or the first "MC"); and (ii) given sampled parameter values obtain a distributional estimate of the posterior and associated posterior statistics of interest (using Monte Carlo **[G]** , or the second "MC"). We discuss each of these "MC" components in turn, in reverse order.

Consider concept (ii) "Monte Carlo". Suppose we have a set of parameter values from some distribution. Monte Carlo integration permits estimation of this distribution using associated empirical estimates[82]. For example, to estimate distributional summary statistics, such as the mean, variance or symmetric 95% credible interval of a parameter we use the corresponding sample mean, variance and 2.5% and 97.5% quantile parameter values. Similarly, probability statements can be estimated (such as the probability that a parameter is positive/negative; or lies in the range [a,b]) as the proportion of the sampled values that satisfy the given statement; while the posterior marginal density of any given parameter can be obtained via kernel density estimation, which uses a non-parametric approach for estimating the associated density from which sampled values have been drawn[83].

However, in general, it is not possible to directly and independently sample parameter values from the posterior distribution. This leads to concept (i) the "Markov chain". The idea is to obtain a sample from the posterior distribution by constructing a Markov chain with some specified first-order transition kernel which defines the distribution of the parameters at iteration t+1, given their state at time t, such that the resulting stationary/equilibrium distribution of the Markov chain is equal to this posterior distribution of interest. Thus, if we run the Markov chain long enough so that it has reached its stationary distribution, subsequent realisations of the chain can be regarded as a (dependent) sample from the posterior distribution and used to obtain the corresponding Monte Carlo estimates, see for an example Figure 4A. We emphasise that the sampled parameter values obtained from the Markov chain are auto-correlated, in that the parameter values are dependent on their previous values in the chain, and generated via the first order Markov chain. The Markov chain is defined by the specification of the initial parameter values and transition kernel **[G]**. There are standard approaches for defining the transition kernel so that the corresponding stationary distribution is the correct posterior distribution: such as the Gibbs sampler[84]; Metropolis-Hastings algorithm[85,86]; and Hamiltonian Monte Carlo[87].

Obtaining posterior inference, by fitting models to observed data can be complicated due to model complexities or data collection processes. For example, for random effect models or in the presence of latent variables, the likelihood may not be available in closed form, but only expressible as an analytically intractable integral (over the random effect terms or latent variables). Alternatively, the likelihood may be available in closed form, for example, for a finite mixture model (or discrete latent variable model), but where the likelihood is multimodal leading to slow mixing within a standard MCMC approach. In such circumstances data augmentation is often used[88], where we define additional variables, or auxiliary variables **[G]**, such that the joint distribution of the data and auxiliary variables (often referred to as the "complete data" likelihood) is now available in closed form and quick to evaluate. For example, for a random effects model, the auxiliary variables correspond to the individual random effect terms (that would previously have been integrated out); for a finite mixture model the auxiliary variables correspond to the mixture component that each observation belongs to. A new joint posterior distribution is then constructed over both the model parameters and auxiliary variables, which is defined to be proportional to the complete data likelihood and associated parameter priors. A standard MCMC algorithm can then be applied that obtains a set of sampled parameter values over both the model parameters and auxiliary variables. Considering the values of only the model parameters of interest within the Markov chain, essentially discarding the auxiliary variables, provides a sample from the original (marginal) posterior distribution of the model parameters given the observed data. Finally we note that the auxiliary variables may themselves be of interest themselves in some cases, and inference on these can be easily obtained via the sampled values.

The transition kernel determines the MCMC algorithm, describing how the parameter values (and any other additional auxiliary variables) are updated at each iteration of the Markov chain. In order for the stationary distribution of the Markov chain to be the posterior distribution of interest, the transition kernel is specified such that it satisfies some relatively straightforward rules. The transition kernel is typically defined via some proposal distribution – this name arises as the process of updating the parameter values involves proposing a set of new parameter values from some distribution which, in the general case, are subsequently either accepted or rejected with some probability, where this acceptance probability is a function of the proposal distribution. If the proposed values are accepted the Markov chain moves to this new state; if the values are rejected the Markov chains remains in the same state at the next iteration. Thus, the transition kernel is non-unique with many general choices for the proposal distribution. For example these include the posterior conditional distribution (i.e. the Gibbs sampler; where the acceptance probability in the updating step is equal to unity), Metropolis-Hastings random walk sampler (randomly perturbing the parameter values from their current values), slice sampler and no-U-turn sampler, amongst many others. We do not focus further on the internal mechanics of the MCMC algorithm here as there is a wealth of literature on this topic and also associated computational tools and programs for performing a Bayesian analysis via an MCMC approach (see later in this section).

Beyond the necessity of specifying a transition kernel, such that the corresponding stationary distribution is the posterior distribution of interest, the choice of transition kernel defines the performance of the MCMC algorithm in terms of how long the Markov chain needs to be run to obtain reliable inference on the posterior distribution of interest. Trace plots **[G]** of the parameters display the value of the parameters over iteration number. One-dimensional trace plots are most commonly plotted that describe the parameter value at each iteration of the Markov chain (on the y-axis) against iteration number (on the x-axis) and are often a useful exploratory tool (Figure 4A). They provide a visualisation of the chain in terms

of how each parameter is exploring the parameter space, often referred to as mixing, which, if poor, require changes to the specified transition kernel; and also for identifying when the Markov chain has reached its stationary distribution. Recall that the Markov chain only converges to the posterior distribution, so that realisations of the chain prior to convergence to its stationary distribution are discarded – this was originally called the burn-in but we prefer the term warm-up.[89] The most common technique applied to assess convergence is the $\hat{R}$ statistic **[G]** [90,91] where multiple independent runs of the MCMC algorithms are run and the within-chain variability and between-chain variability compared (Figure 4B). Ideally, each of the multiple chains should be started from different (over-dispersed) starting values (and using different random seeds) to provide greater initial variability across the Markov chains, to make it more likely that non-convergence of the chain to the stationary distribution will be identified, for example, if different sub-modes of the posterior distribution are being explored. $\hat{R}$ is defined to be the ratio of the within- and between-chain variability. Values close to 1 for all parameters and quantities of interest suggest the chain has sufficiently converged to the stationary distribution, so that future realisations can be regarded as a sample from the posterior distribution of interest (Figure 4B). Once the stationary distribution is reached, a further question relates to how many iterations are needed to obtain reliable Monte Carlo estimates (i.e. for sufficiently small Monte Carlo error). To assess this, batching the sampled values is often used which involves sub-dividing the sampled values into non-overlapping "batches" of consecutive iterations and considering the variability of the estimated statistic using the sampled values in each batch [92].

Additionally, to determine if the entire posterior parameter space has been explored the effective sample size (ESS) of the sampled parameter values may be obtained. The ESS roughly expresses how many independent sampled parameter values contain the same information as the autocorrelated MCMC samples- recall that the sampled MCMC values are not independent as they are generated via a first-order Markov chain. Note that 'sample size' in the ESS does not refer to sample size of the data but can be seen as the effective length of the MCMC chain instead of the actual length of the chain. Low sampling efficiency is related to high autocorrelation (so that the variability of the parameter values is small over successive iterations) and non-smooth histograms of posteriors, which in turn could point towards potential problems in the model estimation or weak identifiability of the parameters[51]. Therefore, when problems occur in obtaining reliable Monte Carlo estimates, a good starting point is to sort all variables based on ESS and investigating the ones with the lowest ESS first. ESS is also useful for diagnosing the sampling efficiency for a large number of variables[93].

For further discussion of MCMC-related issues, see for example[73,94,95]. There are now many standard computer packages for implementing Bayesian analyses, and a summary of the main packages are given in Table 2 (see also *Reproducibility and data deposition*), which have subsequently led to the explosion of Bayesian inference across many scientific fields (for examples, see *Applications*). Many of the available packages perform the MCMC algorithm as a black-box (though often with options to change default settings), permitting the analyst to focus on the prior and model specification, and avoid any technical coding. Note there are many additional packages that make it easier to work with the sometimes heavily code-based software, for example the packages BRMS[96] and Blavaan[97] in R for making it easy to use Stan.[98]

H3: Empirical Example - Continued

The priors for the PhD delay example were updated with the data and posteriors were computed in Stan[98]. All scripts to reproduce the results are available at the Open Science Framework: DOI 10.17605/OSF.IO/JA859. The trace plot of four independent runs of the MCMC algorithms for $\beta_{intercept}$ is shown in Figure 4A and displays stability post-burn in. Also, the associated $\hat{R}$ statistic stabilizes after approximately 2,000 iterations, see Figure 4B. The prior and posterior distributions are displayed in Panels 4C-E. The posterior parameter estimates can be summarized using, for example, the median of the posterior distributions. Based on these point summaries, it appears the delay peaks at around the age of 50, with an explained variance of only of 6%. If we compare our prior and posterior predictive distributions, we are less uncertain and more consistent in what we expect after observing the data. So, accurate predictions of delay for individual cases may not be possible, but we can predict general trends at group level.

[H2] Variational inference

As we have outlined, Bayesian analysis consists of a number of stages including detailed model development, including specifying the prior and data models, the derivation of exact inference approaches based on MCMC, and model checking and refinement (Figure 1). Each is ideally treated independently, separating model construction from its computational implementation. The focus on exact inference techniques has spurned considerable activity in developing Monte Carlo methods which are considered as a gold standard for Bayesian inference. Monte Carlo methods for Bayesian inference adopt a simulation-based strategy for approximating the high-dimensional integrals required to compute posterior quantities. An entirely alternative approach is to produce functional approximations of the posterior using approaches including Variational Inference [G] (VI)[99] or Expectation Propagation[100]. In the following, we describe the variational approach, also known as variational methods or variational Bayes, due to its popularity and prevalence of use in machine learning.

Variational inference begins by constructing an approximating distribution to approximate the desired, but intractable, posterior distribution. Typically, the approximating distribution is chosen from a family of standard probability distributions, e.g. multivariate Normal, and further assumes that some of the dependencies between the variables in our model are broken. In the case, where the approximating distribution assumes all variables are independent, this gives us the well-known "mean-field approximation". The approximating distribution will be specified up to a set of "variational parameters" that we optimise to find the best posterior approximation by minimising the Kullback-Leibler divergence to the true posterior. As a consequence, variational inference reposes Bayesian inference problems as optimisation rather than as sampling problems and can be solved using numerical optimisation, i.e. gradient descent. When combined with subsampling-based optimisation techniques such as stochastic gradient descent, variational inference makes approximate Bayesian inference possible for complex large-scale problems.

Variational methods therefore transform the inference problem into an optimisation task to identify the parameters of the approximation that minimise its discrepancy with respect to the true posterior. In Bayesian machine learning (see also the Outlook section), coordinate descent approaches for optimisation, have generally given way to stochastic optimisation approaches which provide further scalability benefits in the presence of large data sets[101-103]. Stochastic gradient descent uses only subsets

of the data (mini-batches) to compute noisy estimates of the gradients whilst still retaining convergence guarantees. However, there is no free lunch, unless the true posterior belongs to the pre-specified family of approximating distributions, it is often difficult to determine how good the variational approximation represents the true posterior.

[H2] Variable Selection

Variable selection is the process of identifying the subset of predictors to include in a model. It is a major component of model building along with determining the functional form of the model. Variable selection is especially important in situations where a large number of potential predictors is available. The inclusion of unnecessary variables in a model has several disadvantages, such as increasing the risk of multicollinearity, lacking enough samples to estimate all model parameters, overfitting the current data thus leading to poor predictive performance on new data, and making the model interpretation more difficult. For example, in genomic studies where high-throughput technologies are used to profile thousands of genetic markers, only a few of those predictors are expected to be associated with the phenotype or outcome under investigation. Methods for variable selection can be categorized into those based on hypothesis testing and those that perform penalized parameter estimation. In the Bayesian framework, hypothesis testing approaches use Bayes factors and posterior model probabilities, while penalized parameter estimation approaches specify shrinkage priors **[G]** that induce sparsity **[G]**, as discussed below. Bayes factors are often used when dealing with a small number of potential predictors as they involve fitting all candidate models and choosing between them, whereas penalization methods fit a single model and thus can scale up to larger dimensions.

We provide a brief review of these approaches in the context of a classical linear regression model, where the response variable from n independent observations, $\boldsymbol{y}$, are related to p potential predictors defined in an $n \times p$ covariate matrix $\boldsymbol{X}$ via the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. The regression coefficients $\boldsymbol{\beta}$ capture the effect of each covariate on the response and $\boldsymbol{\varepsilon}$ are the residuals assumed to follow a Normal distribution with mean 0 and variance σ^2. Bayes factors[62] (Box 3) can be used to compare and choose between candidate models, where each candidate model would correspond to a hypothesis. Unlike frequentist hypothesis testing methods, Bayes factors do not require the models to be nested. In the context of variable selection, each candidate model corresponds to a distinct subset of the p potential explanatory variables[104,105]. These 2^p possible models can be indexed by a binary vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)'$, where $\gamma_j = 1$ if covariate X_j is included in the model, that is $\beta_j \neq 0$, and $\gamma_j = 0$ otherwise. Let $M_{\boldsymbol{\gamma}}$ be the model that includes the X_j's with $\gamma_j = 1$. Prior distributions for each model $p(M_{\boldsymbol{\gamma}})$ and for the parameters under each model $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 | M_{\boldsymbol{\gamma}})$ are specified, and Bayes factors $BF_{\boldsymbol{\gamma b}}$ are evaluated to compare each model $M_{\boldsymbol{\gamma}}$ to one of the models taken as a baseline, $M_{\boldsymbol{b}}$. The posterior probability, $p(M_{\boldsymbol{\gamma}}|\boldsymbol{y})$, for each model can be expressed in terms of the Bayes factors as

$$p(M_{\boldsymbol{\gamma}}|\boldsymbol{y}) = \frac{BF_{\boldsymbol{\gamma b}}\, p(M_{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{\gamma}'} BF_{\boldsymbol{\gamma}'\boldsymbol{b}}\, p(M_{\boldsymbol{\gamma}'})}$$

where the denominator sums over all considered models $M_{\boldsymbol{\gamma}'}$. The models with largest posterior probabilities would correspond to those with the highest amount of evidence in their favor among the

ones under consideration. When p is relatively small (say $p < 20$), all 2^p variable subsets and their posterior probabilities can be evaluated. The model with highest posterior probability (the maximum *a posteriori* model) may be selected as the one most supported by the data. Alternatively, the covariates with high marginal posterior inclusion probabilities, $p(\gamma_j = 1|\boldsymbol{y}) = \sum_{X_j \in M_\gamma} p(M_{\boldsymbol{\gamma}}|\boldsymbol{y})$, may be selected. For moderate to large p, this strategy is not practically feasible as an exhaustive evaluation of all 2^p possible models becomes computationally expensive. Instead, shrinkage priors that induce sparsity, either by setting the regression coefficients of non-relevant covariates to zero or by shrinking them towards zero, are specified and MCMC techniques are used to sample from the posterior distribution..

Various shrinkage priors have been proposed over the years. A widely used shrinkage prior **[G]** is the spike-and-slab prior **[G]**, which uses the latent binary indicator vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p) \in \{0,1\}^p$ to induce a mixture of two distributions on β_j , one peaked around zero (spike) to identify the zero elements and the other a flat distribution (slab) to capture the non-zero coefficients[106,107]. The discrete spike-and-slab formulation[106] uses a mixture of a point mass at zero and a flat prior (see Figure 5A), while the continuous spike-and-slab prior[107] uses a mixture of two normal distributions (see Figure 5B). Another widely used formulation puts the spike-and-slab prior on the variance of the regression coefficients[108]. After specifying prior distributions for the other model parameters, MCMC algorithms are used to explore the large model space and yield a chain of visited models. Variable selection is then achieved through the marginal posterior inclusion probabilities, $P(\gamma_j = 1|\boldsymbol{y})$. Integrating out the parameters $\boldsymbol{\beta}$ and σ^2 can accelerate the MCMC implementation, while speeding up its convergence and mixing. Various computational methods have also been proposed to rapidly identify promising high posterior probability models, by combining variable selection methods with modern Monte Carlo sampling techniques[109,110](see also Table 1).

Another class of regularization priors that have received a lot of attention in recent years are continuous shrinkage priors[111-113]. These are unimodal distributions on β_j that promote shrinkage of small regression coefficients towards zero, similarly to frequentist penalized regression methods that accomplish regularization by maximizing the log-likelihood function subject to a penalty[114]. The least absolute shrinkage and selection operator (lasso)[114], for instance, uses the penalty function $\lambda \sum_{j=1}^{p} |\beta_j|$ with λ controlling the level of sparsity. The lasso estimate of β_j can be interpreted as a Bayesian posterior mode estimate using independent Laplace priors for the regression coefficients. Motivated by this connection, the Bayesian lasso[111] specifies conditional Laplace priors on $\beta_j|\sigma^2$. It should be noted that Bayesian penalization methods do not shrink regression coefficients to be exactly zero, as the lasso penalization does. Instead, the variable selection is carried out using credible intervals for β_j or by defining a selection criterion on the posterior samples. Many continuous shrinkage priors can be parametrized as a scale mixture of normal distributions, which facilitates the MCMC implementation. For example, the Laplace prior in the Bayesian lasso can be obtained as a scale mixture of normals with an exponential mixing density. The exponential mixing distribution has a single hyperparameter, which limits its flexibility in differentially shrinking small and large effects (see Figure 5C). This limitation can be overcome by using a class of shrinkage priors that introduce two shrinkage parameters, which respectively control the global sparsity and the amount of shrinkage for each regression coefficient. The resulting marginalized priors for β_j are characterized by a tight peak around zero that shrinks small coefficients to zero, and heavy tails that prevent excessive shrinkage of large coefficients. These priors are known as global-local shrinkage priors[113]. The horseshoe prior **[G]**, for example, achieves this by specifying a normal distribution for the

regression coefficient, β_j, conditional on its scale parameters, which in turn, follow half-Cauchy distributions[112](see Figure 5D). A comprehensive review and thorough comparison of the characteristics and performance of different shrinkage priors can be found in [115]. Bayesian variable selection methods have been extended to a wide variety of models. Extensions to multivariate regression models include spike-and-slab priors that select variables as relevant to either all or none of the responses[116], as well as multivariate constructions that allow each covariate to be relevant for subsets and/or individual response variables[117]. Other extensions include generalized linear models, random effects and time-varying coefficient models[118,119], mixture models for unsupervised clustering[120], and estimation of single and multiple Gaussian graphical models[121,122]. The forthcoming Handbook of Bayesian Variable Selection[123] presents a comprehensive review and highlights recent developments.

[H3] Examples of Recent Applications of Bayesian Variable Selection in Biomedical studies

The variable selection priors for linear models described in the Results section have found important applications in biomedical studies. We briefly discuss some examples of recent applications of Bayesian variable selection methods.

The advent of high-throughput technologies has made it possible to measure thousands of genetic markers on individual samples. Linear models are routinely used to relate large sets of biomarkers to disease-related outcomes, and variable selection methods are employed to identify the significant predictors. In Bayesian approaches, additional knowledge about correlation structure among the variables can be easily incorporated into the analysis. For example, in models with gene expression data, spike-and-slab variable selection priors incorporating knowledge on gene-to-gene interaction networks have been employed to aid the identification of predictive genes[124], as well as the identification of both relevant pathways and subsets of genes[125]. Other successful applications of Bayesian variable selection priors have been in genome-wide association studies (GWAS), where hundreds of thousands of single nucleotide polymorphisms (SNPs) are measured in thousands or tens of thousands of individuals, with the goal of identifying genetic variants that are associated with a single phenotype or a group of correlated traits.[126,127]

Air pollution is a major environmental risk factor for morbidity and mortality. Small particles produced by traffic and industrial pollution can enter the respiratory tract and have adverse health effects. Particulate matter exposure and their health effects exhibit both spatial and temporal variability. For a treatment of Bayesian hierarchical models for spatial data we refer readers to[128]. Spatially varying coefficients models with spike-and-slab priors inducing spatial correlation have been proposed to identify pollutants associated to adverse health outcomes over a whole region, as well as in different subregions[129]. Over the past couple of decades, a number of -omic studies have been conducted to investigate the effects of environmental exposures on genomic markers and gain a better understanding of the mechanisms underlying lung injury from exposure to air pollutants. Multivariate response models with structured spike-and-slab priors that leverage the dependence across markers have been proposed to identify and estimate the joint effect of pollutants on DNA methylation outcomes[117].

In neuroscience, neuroimaging studies often employ functional magnetic resonance imaging (fMRI), a non-invasive technique that provides an indirect measure of neuronal activity by detecting blood flow

changes. These studies produce massive collections of time series data, arising from spatially distinct locations of the brain, on one or multiple subjects. In a typical task-based experiment, the whole brain is scanned at multiple times while the subject performs a series of tasks. The objective of the analysis is to detect those brain regions that get activated by the external stimulus. Bayesian approaches to general linear models that employ spatial priors have played an important role in the analysis of such data, as they allow a flexible modelling of the correlation structure of the data[130]. Spike-and-slab variable selection priors that incorporate structural information on the brain have been investigated within a wide class of spatio-temporal hierarchical models for the detection of the activation patterns[131,132]. Other applications of Bayesian variable selection priors in fMRI analysis have been in brain connectivity studies. Here, fMRI data are measured, on subjects typically at rest, with the aim of inferring how brain regions interact with each other and how information is transmitted between them. Among other approaches, multivariate vector autoregressive linear models have been investigated as a way to infer effective (i.e., directed) connectivity. Continuous shrinkage priors as well as structured spike-and-slab prior constructions have been employed for the selection of the active connections[133,134]. Bayesian variable selection methods have been successfully applied to a number of other biomedical areas, involving longitudinal data, functional data, survival outcomes and case-control studies, to mention a few.

[H2] Posterior Predictive Checking

Once a posterior distribution for a particular model is obtained, it can be used to simulate new data conditional on this distribution. Those simulations can be used for, at least, three purposes: First, to check if the simulated data from the model resemble the observed data. To this end, one could compare kernel density estimate of the observed data to density estimates for the simulated data[67]. Second, a more formal posterior predictive checking approach can be taken to evaluate if the model can be considered a good fit with the data generating mechanism[67,77,135-137]. Any parameter-dependent statistic or discrepancy can be used for the posterior predictive check[136]. This is similar to how prior predictive checks can be used but much more stringent in the comparison[67]. Because posterior distributions are usually more concentrated on the parameter space compared to prior distributions, the tails of the predictive distributions are more concentrated and tail-area probabilities for any observed statistic or discrepancy are hence more sensitive. The sensitivity of the posterior predictive checks is useful because if realistic models are used, the expectation is that these are well calibrated in the long-term average[77], for more details see *Limitations and opimizations.* Third, posterior predictive distributions can be used to extrapolate beyond the observed data to predict what data we would expect for new situations based upon our model, e.g. in time series. The first two uses of posterior predictive checking should be used with care. There is a risk of over adjusting and refining models to much to the details of a specific data set. An example of this third kind of use of posterior predictive distributions can be found in the time series of Figure 6. The analysis highlights how daily webpage views can be decomposed into non-periodic changes, holiday effects, weekly seasonality, and yearly seasonality effects. Based on the posterior distributions for the particular model, posterior predictive distributions were simulated for the observed and future data, naturally becoming more uncertain when they are further ahead due to accumulated uncertainty. It is also important to be aware that in temporal models some challenges in terms of posterior inference that are inherent to the spatial and/or temporal dependencies[44,138-140].

H3: Empirical Example – Time Series Wikipedia page views

To illustrate the use of posterior predictive distributions suppose that it is of interest to know how many pageviews a webpage has, and what time related factors might be relevant. Consider the Wikipedia page views for the premier league, the highest English professional soccer league, obtained using the 'wikipediatrend'[141] R package. The scripts are available at the Open Science Framework: https://osf.io/7yrud/ - DOI 10.17605/OSF.IO/7YRUD. The decomposable time series model[142] implemented in the 'prophet'[143] R package, allows the estimation of trends with non-periodic changes (Figure 6A), holiday effects (B), weekly seasonality (C), and yearly seasonality effects (D). Notable effects in this time series are the peaks of interest surrounding the start of the seasons in August, the end of the seasons in May, and the dip on 29-04-2011 – the wedding day of Prince William and Catherine Middleton. Additionally, a decrease in webpage views occur on each Christmas day, and notable increases occur on Boxing day and at the start of the year when traditionally matches are played during the Christmas break. The model is estimated using observed data in the period between January 1st 2010 and January 1st 2018. Based on the posterior distributions for the particular model, posterior predictive distributions can be simulated for the observed and future data. In panels E and F posterior predictive distributions at each time point can be seen. In general, the simulated data from the model resembles the observed data for the observed time frame. The posterior predictive distributions for future time points are more uncertain when they are further ahead due to accumulated uncertainty. Notice that increases and decreases in page views are accurately predicted for future page views, with the exception of increased interest in July 2018 which might relate to the final stage of the World cup Soccer at that time.

[H1] Applications

Bayesian inference has been used across all fields of science. We describe a few examples here but there are many other areas of application such as philosophy, pharmacology, economics, physics, political science and beyond.

[H2] Social and Behavioural sciences

A recent systematic review examining the use of Bayesian statistics found that the social and behavioural sciences (e.g., psychology, sociology, and political sciences) have experienced an increase in empirical Bayesian work[4]. The number of Bayesian publications has been steadily rising since about 2004, with more notable increases in the last decade. In part, this focus on Bayesian methods has been due to the development of more accessible software, as well as a focus on publishing tutorials aimed at applied social and behavioural scientist researchers. The increase in prevalence of Bayesian methods is also due to the continued use of Bayes' rule as a theory for developmental processes.

Specifically, there have been two parallel uses of Bayesian methods within the social and behavioural sciences: theory development and estimation. The field has experienced an increase in use with respect to each of these two perspectives.

Bayes' rule has been used as an underlying theory for understanding reasoning, decision-making, cognition, and theories of mind. This implementation has been especially prevalent within developmental psychology and related fields. For example, Bayes' rule was used as a conceptual framework for cognitive development in young children, capturing how children develop an understanding of the world around them[144]. Bayesian methodology has also been discussed in terms of enhancing cognitive algorithms used for learning. Specifically, Gigerenzer and Hoffrage[145] discussed the use of frequencies, opposed to probabilities, as a method to improve upon Bayesian reasoning. In another seminal paper, Slovic and Lichtenstein[146] discussed how Bayesian methods can be used for judgement and decision-making processes. Within this area of the social and behavioural sciences, Bayes' rule has been used as an important conceptual tool for developing theories and understanding developmental processes.

The second way that Bayes' rule is used within the social and behavioural sciences, and the focus of much of the current paper, is as a tool for estimation.

The social and behavioural sciences are a terrific setting for implementing Bayesian inference. The literature is rich with information that can be used to derive prior knowledge. In turn, informative priors are useful in complex modelling situations, which are common in the social sciences, as well as in cases of small sample sizes. Likewise, certain models (e.g., some multidimensional item response theory models) used to explore education outcomes and standardized tests are intractable using frequentist methods and require the use of Bayesian methods.

There have been many tutorials aimed at explaining Bayesian methods to empirical researchers in a variety of subsections of the social and behavioural sciences. To highlight the scope of tutorials, a systematic review of Bayesian methods in the field of psychology uncovered 740 eligible regression-based papers using this approach. Of these, 100 papers (13.5%) were tutorials for implementing Bayesian methods, and an additional 225 papers (30.4%) were either technical papers or commentaries on Bayesian statistics. Some examples of tutorials within this field are as follows. Hoijtink et al.[147] discussed the use of Bayes factors for informative hypotheses within cognitive diagnostic assessment. They illustrated how Bayesian evaluation of informative diagnostics hypotheses can be used as an alternative approach to the traditional diagnostic methods. There is added flexibility with the Bayesian approach since informative diagnostic hypotheses can be evaluated using the Bayes factor using only data from the individual person being diagnosed. Lee[148] published an overview of how Bayes' theorem can be used within the field of cognitive psychology. They discuss how Bayesian methods can be used to develop more complete theories of cognitive psychology, account for observed behaviour in terms of different cognitive processes, explain behaviour on a wide range of cognitive tasks, and provide a conceptual unification of different cognitive models. Depaoli et al.[149] showed how Bayesian methods can benefit health-based research being conducted within psychology. Specifically, they highlighted how informative priors via expert knowledge and previous research can be used to better understand the physiological impact of a health-based stressor. In this research scenario, frequentist methods would not have produced viable results because the sample size was relatively small for the model being estimated (data were expensive to collect and analyse and the population was difficult to access for sampling). Finally, Kruschke[150] presented the simplest example using a *t*-test geared toward experimental psychologists, showing how Bayesian methods can benefit the interpretation of any model parameter. This paper highlights the Bayesian way of interpreting results, focusing on the interpretation of the entire posterior rather than a point estimate.

Methodologists have been attempting to guide applied researchers toward using Bayesian methods within the social and behavioural sciences. Although the implementation has been slower to catch on (e.g., the systematic review found only 167 regression-based papers (22.6%) were empirical applications using human samples), some subfields are regularly publishing work implementing Bayesian methods.

The field has gained many interesting insights to psychological and social behaviour through Bayesian methods, and the substantive areas where this work has been conducted are quite diverse. For example, Bayesian statistics helped to: uncover the role that craving suppression has in smoking cessation[151], make population forecasts based on expert opinions[152], examine the role that stress related to infant care has in divorce[153], examine the impact of the President of the United States' ideology on U.S. Supreme Court rulings[154], and predict behaviours that limit the intake of "free sugars" in one's diet[155].

These examples all represent different ways in which Bayesian methodology is captured in the literature. It is common to find papers that highlight Bayes' rule as a mechanism to explain theories of development and critical thinking[144], are expository [149,150]), focus on how Bayesian reasoning can inform theory through use of Bayesian inference[148], and papers using Bayesian modelling to extract findings that would have been difficult using frequentist methods[151]. Overall, there is broad use of Bayes' rule within the social and behavioural sciences.

We argue that the increased use of Bayesian methods in the social and behavioural sciences is a great benefit to improving substantive knowledge. However, we also feel that the field needs to continue to develop strict implementation and reporting standards so that results are replicable and transparent, as discussed in the next section. We believe that there are important benefits to implementing Bayesian methods within the social sciences, and we are optimistic that a strong focus on reporting standards can make the methods optimally useful for gaining substantive knowledge.

[H2] Ecology

Applying Bayesian analyses to ecological applications has become increasingly widespread due to both philosophical arguments and practical model-fitting advantages. This is combined with readily available software, see Table 2, and numerous publications describing Bayesian ecological applications using a range of software packages (see for example[156-162] amongst many others). The underlying Bayesian philosophy is attractive in many ways within ecology[163] as it permits: the incorporation of external, independent, prior information within a rigorous framework (such information may be from previous studies on the same/similar species or from using inherent knowledge of the biological processes)[164,165]; the ability to make direct probabilistic statements on parameters of interest (such as survival probabilities, reproductive rates, population sizes and future predictions)[158]; the calculation of relative probabilities of competing models (for example, the presence/absence of density dependence or environmental factors in driving the dynamics of the ecosystem) which in turn permit model-averaged estimates incorporating both parameter and model uncertainty. The ability to provide probabilistic statements is particularly useful in relation to wildlife management and conservation. For example, King et al[166] provide probability

statements in relation to the level of population decline over a given time period, which in turn provides probabilities associated with species' conservation status.

A Bayesian approach is also often applied in practice for pragmatic reasons.Many ecological models are complex (for example, they may be spatio-temporal in nature, high-dimensional and/or involving multiple interacting biological processes) leading to computationally expensive likelihoods that are slow to evaluate; while imperfect or limited data collection processes often lead to missing data and associated intractable likelihoods. In such circumstances standard Bayesian model-fitting tools, such as data augmentation, may permit the models to be fitted; whereas in the alternative frequentist framework, additional model simplifications or approximations may be required. The application of Bayesian statistics in ecology is vast and encompasses a range of spatio-temporal scales from an individual organism level to ecosystem level, from understanding the population dynamics of the given system[167], modelling spatial point pattern data[168], to population genetics, to estimating abundance[169] or assessing conservation management[170].

Ecological data collection processes are generally from observational studies, where a sample is observed from the population of interest using some given data survey protocol. In general, the survey should be carefully designed, taking into account the ecological question(s) of interest and so that it minimises the complexity of the model required to fit to the data to be able to answer the given question with a high degree of accuracy. Nevertheless, due to data collection problems (which may, for example, be as a result of equipment failure or due to poor weather conditions), or inherent data collection problems (for example it is not possible to record any individual level information, such as breeding status, if an individual is unobserved), associated model-fitting challenges may arise. Such challenges may include (but are far from limited to) irregularly spaced observations in time (possibly due to equipment failure or motion sensor detections), measurement error (for example, in relation to population counts or disease/breeding status of individuals made from visual observations), missing information (such as individual covariate information or global environmental factors) and multi-temporal and/or spatial scales where different aspects of data are recorded at different temporal scales (for example, hourly GPS location data of individuals; daily environmental data collected at fixed locations; monthly aerial/satellite photographs and annual censuses). The data complexities that arise, combined with associated modelling choices, may lead to a range of model-fitting challenges which can often be more easily addressed within the Bayesian paradigm.

For a given ecological study, separating out the individual processes acting on the ecosystem is an attractive mechanism for simplifying the model specification process.[167] For example, state-space models provide a general and flexible modelling framework that describe two distinct types of processes: (i) the system process and (ii) the observation process. The system process describes the true underlying state of the system and how this changes over time. These states may be univariate (such as population size) or multivariate (such as location data); and the system process may describe multiple processes acting on the system (such as birth/reproduction/dispersal/death). However, we are typically not able to observe the true states without some associated error: the observation process describes how the observed data relate to the true (unknown) states. These general state-space models span many applications, including for example, animal movement[171]; population count data[172]; capture-recapture-type data[166]; fisheries stock assessment[173]; and biodiversity[174] (for a review and further applications, see for example[167,175,176]).

Bayesian model-fitting tools, such as MCMC with data augmentation[177], sequential Monte Carlo or particle (P)MCMC,[178-180] permit general state-space models to be fitted to the observed data without the need to specify further restrictions on the model specification (such as distributional assumptions) or make additional likelihood approximations.

The process of collecting data continues to evolve with advances in technology, for example, use of GPS geo-location tags and associated additional accelerometers; remote sensing; use of drones for localised aerial photographs; unmanned underwater vehicles; motion-sensor camera traps; citizen science etc. The use of these technological devices has led to new forms of data, and in greater quantity, and associated model-fitting challenges, providing a fertile ground for Bayesian analyses.

[H2] Genetics

Genetics and genomics have been a popular application of Bayesian methods. In genome-wide association studies (GWAS), Bayesian approaches have provided a powerful alternative to frequentist approaches for assessing the evidence of population associations between genetic variants and a phenotype of interest[181]. These include approaches for incorporating genetic diversity (e.g. admixture[182]), fine-mapping to identify causal genetic variants[183], imputation of genetic markers not directly measured using reference populations[184] and meta-analysis for combining information across studies. These applications further benefit from the use of marginalisation in order to account for modelling uncertainties when drawing inferences. More recently, large cohort studies such as the UK Biobank (UKBB)[185] have collated heterogeneous datasets (e.g. imaging, lifestyle, routinely collected health data) alongside genetic information that have expanded the methodological requirements for identifying genetic associations with complex (sub)phenotypes. For example, a Bayesian analysis framework TreeWAS[186] has extended genetic association methods to allow for the incorporation of tree-structured disease diagnosis classifications by modelling the correlation structure of genetic effects across observed clinical phenotypes. This approach incorporates prior knowledge of phenotype relationships that can be derived from a diagnosis classification tree (e.g. ICD-10).

Beyond genetics, the availability of multiple molecular data types (“multi-omics”) has also attracted Bayesian solutions to the problem of multimodal data integration. Bayesian latent variable models can be used as an unsupervised learning approach to identify latent structures that correspond to known or previously uncharacterised biological processes across different molecular scales. Multi-Omics Factor Analysis (MOFA)[187] uses a Bayesian linear factor model to disentangle sources of heterogeneity that are common across multiple modalities from those specific to individual data modalities.

In recent years, high-throughput molecular profiling technologies have advanced to allow the routine -omics analysis of individual cells[188]. This has led to a methodological revolution with an explosion of novel approaches to account for the challenges of modelling single cell measurement noise, cell-to-cell heterogeneity, high-dimensionality, large sample sizes (millions of cells) and perturbation effects from, for instance, genome editing[189]. Cellular heterogeneity lends itself naturally to Bayesian hierarchical modelling and formal uncertainty propagation and quantification due to the layers of variability induced by tissue-specific activity, heterogenous cellular phenotypes within a given tissue and stochastic molecular expression at the level of the single cell. In BASiCS[190] this approach is used to account for cell-specific normalisation constants, technical variability to decompose total gene expression variability into technical and biological components.

Deep neural networks have also been utilised to specify flexible, non-linear conditional dependencies within hierarchical models for single cell -omics. SAVER-X[191] couples a Bayesian hierarchical model with a pretrainable deep autoencoder to extract transferable gene−gene relationships across datasets from different laboratories, variable experimental conditions and divergent species to denoise novel target datasets. While in scVI[192], hierarchical modelling is used to aggregate information across similar cells and genes to infer the distributions that underlie observed expression values. Approximate and scalable inference in both applications is enabled through the use of mini-batch stochastic gradient descent [**G**] (the latter within a variational setting) - a standard technique with modern use of deep neural networks - that allow these models to be fitted to hundreds of thousands to millions of cells (see also the outlook section).

Bayesian approaches have also been popular for cancer genomics where large-scale cancer genomic datasets[193] have enabled a data-driven approach to identifying novel molecular changes that drive cancer initiation and progression. Bayesian network models[194] have been developed to identify the interactions between mutated genes and capture mutational patterns (signatures) that highlight key genetic interactions that potentially allow for genomic-based patient stratification for clinical trials and the personalised use of therapeutics.

Bayesian methods have been important in answering questions about evolutionary processes in cancer. Several Bayesian approaches for phylogenetic analysis of heterogeneous cancers enable the identification of the distinct subpopulations that can exist with tumours and the ancestral relationships between these through the analysis of single cell and bulk tissue sequencing data[195]. These models therefore consider the joint problem of learning a mixture model (number and identity of the subpopulations) and graph inference (phylogenetic tree).

[H1] Reproducibility and Data Deposition

Proper reporting on statistics, including sharing of data and scripts, is a crucial element in the verification and reproducibility of research[196]. A typical workflow for good research practices across the research workflow that can contribute to reproducibility is displayed in Figure 7. We demonstrate where the Bayesian research cycle (Figure 1) and the *When to Worry, and how to Avoid the Misuse of Bayesian Statistics* checklist[149] (Box 4) fit in the wider context of transparency in research. In this section we highlight some important aspects of reproducibility and data /script deposition.

Allowing others to assess the statistical methods used, including access to the underlying data if possible, can help in interpreting the results, assess the suitability of the parameters used, and detect and fix errors. Reporting practices are not yet consistent across many fields, nor across journals in individual fields. Within the systematic review on Bayesian statistics in psychology[4], huge discrepancies within reporting practices and standards were uncovered in the social sciences. For example, of the 167 regression-based Bayesian papers using human samples in Psychology, 31% did not mention the priors that were implemented, 43.1% did not report on chain convergence, and only 40% of those implementing informative priors conducted a sensitivity analysis. We view this as a major impediment to the implementation of Bayesian statistics within the social and behavioural sciences, as well as other fields of research.

Specifically, for Bayesian methods there are many dangers in naïvely using priors. That is, the exact influence of the priors is often not well understood, and priors might have a huge, sometimes unwanted, impact on the study results. Therefore, one might want to pre-register the specification of the priors (and likelihood) when possible, e.g. in a confirmatory study when the actual statistical model is known beforehand. Moreover, akin to many elements of frequentist statistics, some Bayesian features can be easily misused. For example, the impact of priors on final model estimates can be easily overlooked. A researcher may estimate a model with certain priors and be unaware that using different priors with the same model and data can result in substantively different results. In both cases, the results could look completely viable, for example, chains appeared to be converged, posteriors appear viable and informative. Without examining the impact of priors through a sensitivity analysis and prior predictive checking, the researcher would not be aware of how sensitive results are to changes in the priors. Consider the prior variance in the PhD delay example for β_{age} which was mis-specified as being a precision instead of a variance.

Also, reporting on Bayesian statistics is not consistent with reporting on frequentist statistics, since there are elements included in the Bayesian framework that are fundamentally different from frequentist settings. Therefore, the *WAMBS-* checklist[149] was developed to promote proper use and reporting of Bayesian methods. We offer an updated version (WAMBS, version 2) here (Box 4).

To enable reproducibility and allow others to rerun Bayesian statistics on the same data with, e.g. other priors, model or likelihood functions for sensitivity analyses[197], it is important that the underlying data and code used are properly documented and shared, following the FAIR principles[198,199]: Findable, Accessible, Interoperable and Reusable. Preferably, data and code are shared in a trusted repository[200] rather than as supplemental information in a journal, with their own persistent identifier (such as a doi) and tagged with metadata describing the dataset or codebase. This also allows the dataset and code to be recognized as separate research outputs and allows other to cite them accordingly[201]. Repositories can be general (such as Zenodo), language-specific such as CRAN for `R` packages, and `PyPI` for Python code, or domain-specific[201]. As data and code require different license options, metadata, and other attributes, data are generally best stored in dedicated data repositories, which can be general or discipline-specific[202]. Some journals, like Nature Research' Scientific Data, have their own list of recommended data repositories (https://www.nature.com/sdata/policies/repositories). To make depositing data and code easier for researchers, two repositories (Zenodo and Dryad) are exploring collaboration to allow deposition of code and data through one interface, with data stored in Dryad and code in Zenodo (https://blog.datadryad.org/2020/03/10/dryad-zenodo-our-path-ahead/). Many scientific journals adhere to TOP guidelines[203] for transparency and openness in research, which specify requirements for code and data sharing.

Verification and reproducibility do not only require access to the data, but also to the code used in Bayesian modelling, ideally replicating the original environment the code was run in, with all dependencies documented either in a dependency file accompanying the code or by creating a static container image than provides a virtual environment to run the code in[202]. Open source software should be used as much as possible, as open sources reduce the monetary and accessibility threshold to replicating scientific results. Moreover, it can be argued that closed source software keeps part of the academic process hidden, including from the researchers who use the software. However, open-source software is only truly accessible with proper documentation (e.g. listing dependencies and configuration

instructions in Readme files, commenting code to explain functionality, and including a comprehensive reference manual when releasing packages).

[H1] Limitations and Optimizations

Bayesian inference is optimal conditional on the assumed model. That is, Bayesian posterior probabilities are calibrated in long-term average, if parameters are drawn from the prior distribution and data are drawn from the data distribution. That is, events with stated probability occur with that frequency in the long term, when averaging over the generative model. In practice, our models are never correct; this is where the limitations come from. There are two ways we would like to overcome these limitations: by identifying and fixing problems with the model, and by demonstrating that certain inferences are robust to reasonable departures from the model. There are many examples of model checks, see the sections on prior and posterior predictive checking, and robustness checks, like sensitivity analyses and checklists like the WAMBS (see Box 4), in the Bayesian literature.

Even the simplest and most accepted Bayesian inferences can have serious limitations. For example, suppose an experiment is conducted yielding an unbiased estimate z of a parameter θ which represents the effect of some treatment. If this estimate z is normally distributed with standard error s, we can write $z \sim Normal(\theta, s)$, a normal distribution parameterized by its location and scale parameter. Suppose that θ has a flat uniform prior distribution, then the posterior distribution is $\theta \sim N(z, s)$. These are all familiar calculations. Now suppose we observe $z = s$; that is, the estimate of θ is 1 standard error from zero. In practice, this would be considered statistically indistinguishable from noise, in the sense that such an estimate could occur by chance, even if the true parameter value were zero. But the Bayesian calculation gives a posterior probability $\Pr(\theta > 0|z) = 0.84$. Would you really be willing to offer 5-to-1 odds on a bet that $\theta > 0$, given these data? If not, in what sense can we say this probability is calibrated?

The answer is that the probability is calibrated if you average over the prior. You can't average over a uniform distribution on an infinite range, so let's consider a very diffuse prior, for example $\theta \sim N(0{,}1000)$, where we are assuming that s is roughly on unit scale. Under this model, when z is observed to equal s, the parameter θ will be positive approximately 84% of the time. The reason why the 84% probability doesn't seem correct is that the uniform, or very diffuse, prior does not generally seem appropriate. In practice, studies are designed to estimate treatment effects with a reasonable level of precision. True effects may be one or two standard errors from zero, but they are rarely 5 or 10 or 100 standard errors away. In this example, Bayesian inference if taken literally would lead to over-certainty: an 84% posterior probability corresponds to the willingness to bet at 5-to-1 odds. There is a positive way to look at this story, though: the evident problem with the bet allowed us to recognize that prior information was available that we had not included in our model. Moreover, a weakly informative prior such as $\theta \sim Normal(0, s)$ does not change the posterior by much, as then the posterior becomes normal $Normal(0{,}5\ s, 1/\mathrm{sqrt}(2)s)$, so $\Pr(\theta > 0|z) = 0.76$, and the betting odds only change to roughly 4:1. Ultimately, only a strong prior will make a big difference. Bayesian probabilities are only calibrated when averaging over the true prior or population distribution of the parameters.

More generally, Bayesian models can be checked by comparing posterior predictive simulations to data[136] and by estimating out-of-sample predictive error [204]. There is a benefit to strong prior distributions

that regularize (constrain parameters to reasonable values) to allow the inclusion of more data while avoiding overfitting. More data can come from various sources, including additional data points, additional measurements on existing data, and prior information summarizing other data or theories. All methods, Bayesian and otherwise, require subjective interpretation in order to tell a plausible story, and all models come from researcher decisions. The point is that any choice of model has implications. For example, the flat prior is weak in the sense of providing no shrinkage of the estimate, but it is strong in the sense of leading to an inappropriate level of certainty about the sign of theta.

[H1] Outlook

The widespread adoption of Bayesian Statistics across disciplines is a testament to the power of the Bayesian paradigm for the construction of powerful and flexible statistical models within a rigorous and coherent probability framework. Modern Bayesian practitioners have access to a wealth of knowledge and techniques that allows the creation of bespoke models and computational approaches for particular problems. While probabilistic programming languages, such as Stan, can take away much of the implementation details for many applications allowing the focus to remain on the fundamentals of modelling and design.

Nevertheless, an ongoing challenge for Bayesian Statistics is the ever-growing demands posed by increasingly complex real-world applications. These are often associated with issues such as large datasets and uncertainties regarding model specification. All of this occurs within the context of rapid advances in computing hardware, the emergence of novel software development approaches and the growth of "data sciences" which has attracted a larger and more heterogeneous scientific audience than ever before.

In particular, in recent years, the revision and popularisation of the term "artificial intelligence" (AI) to encompass a broad range of ideas including Statistics and Computation has blurred the traditional boundaries between disciplines. This has been hugely successful in popularising probabilistic modelling and Bayesian concepts outside of its traditional roots in Statistics but has also seen transformations in the way Bayesian inference is being carried out and new questions about how Bayesian approaches can continue to be right at the innovative forefront of AI research.

Driven by the need to support large-scale applications involving datasets of increasing dimensionality and sample numbers, Bayesians have exploited the growth of new technologies centred around Deep Learning (DL). This includes deep learning programming frameworks (e.g. `TensorFlow`,[205] `PyTorch`[206]) that greatly simplify the use of and computations with deep neural networks (DNN) that permit the construction of more expressive, data-driven models that are immediately amenable to inference techniques using off-the-shelf optimisation algorithms and state-of-the-art hardware (multicores, GPUs, TPUs). In addition to providing a powerful tool to specify flexible and modular generative models, DNNs have also been employed to develop new approaches for approximate inference and stimulated a new paradigm for Bayesian practice that sees the integration (not separation) of statistical modelling and computation at its core.

An archetypal example is the "Variational Autoencoder" (VAE)[207]. VAEs have been successfully used in a variety of applications, including single cell genomics[191,192], and they provide a general modelling

framework that has led to a number of extensions including latent factor disentanglement[208-210]. The underlying statistical model is actually a simple Bayesian hierarchical latent variable model. This model maps high-dimensional observations to low-dimensional latent variables that are assumed to be normally distributed through functions defined by DNNs. Variational inference (VI) is used to approximate the posterior distribution over the latent variables. However, in standard VI we would introduce a local variational parameter for each latent variable, in which case the computational requirements would scale linearly with the number of data samples. VAEs use a further approximation process known as *amortization* to replace inference over the many individual variational parameters with a single global set of parameters that are used to parameterise a DNN (known as a *recognition network*) that outputs the local variational parameters for each data point.

Remarkably, when the model and inference are combined and interpreted together, the VAE has an elegant interpretation as an encoding-decoding algorithm: It consists of a probabilistic *encoder* - a DNN that maps every observation to a distribution in the latent space - and a probabilistic *decoder* - a complementary DNN that maps each point in the latent space to a distribution in the observation space. Thus, model specification and inference have become entangled within the VAE, demonstrating the increasingly blurry boundary between principled Bayesian modelling and algorithmic DL techniques. Other recent examples include the use of DNNs to construct probabilistic models that define distributions over possible functions[211-213], build complex probability distributions by applying a sequence of invertible transformations (normalizing flows)[214,215] and define models for exchangeable sequence data[216].

The expressive power of DNNs and their utility within model construction and inference algorithms come with compromises that are fertile ground for further Bayesian research. The trend toward entangling models and inference has popularised these techniques for large-scale data problems but fundamental Bayesian concepts remain to be fully incorporated within this paradigm. Marginalisation, model averaging, decision theoretic approaches rely on accurate posterior characterisation which remains elusive due to the challenge posed by high-dimensional neural network parameter spaces[217]. While Bayesian approaches to neural network learning have been around for decades[218-221], further investigation into prior specifications for modern Bayesian deep learning models which involve complex network structures is required to understand how priors translate to specific functional properties[222].

Recent debates within the field of artificial intelligence have questioned the requirement for Bayesian approaches and highlighted potential alternatives. For instance, Deep Ensembles[223] have been shown to be alternatives to Bayesian methods for dealing with model uncertainty. However, more recent work has shown that ``Deep Ensembles" can actually be reinterpreted as approximate Bayesian model averaging[224]. Similarly, "Dropout" is a regularization approach popularised for use in the training of deep neural networks to improve robustness by randomly dropping out nodes during the training of the network[225]. Dropout has been empirically shown to improve generalizability and reduce overfitting. Bayesian interpretations of dropout have emerged linking it to forms of Bayesian approximation of probabilistic deep Gaussian processes[226]. While the full extent of Bayesian principles have not yet been generalised to all recent developments in artificial intelligence, it is nonetheless a success that Bayesian thinking is deeply embedded and crucial to a number of innovations that have arisen. The next decade is sure to bring a new wave of exciting innovative developments for Bayesians Intelligence.

1112

Tables

Table 1. **A non-exhaustive overview of sampling and approximation techniques**

Name	Short description
MCMC	Markov chain Monte Carlo
Metropolis-Hastings (MH)	Updating algorithm uses general proposal distribution, with an associated accept/reject step for the proposed parameter value(s).[85,86]
Reversible jump (RJ)MCMC	Extension of MH algorithm to permit trans-dimensional moves within parameter space – most often applied in presence of model uncertainty.[33,227]
Hamiltonian Monte Carlo	Special case of MH algorithm based on Hamiltonian dynamics.[87]
No-U-Turn sampler (NUTS)	An extension to Hamiltonian Monte Carlo that optimizes the generation of candidate points.[228]
Gibbs sampler	Special case of MH algorithm where the proposal distribution is the corresponding posterior conditional distribution, with an associated acceptance probability of 1.[84]
Particle (P)MCMC	Combined sequential Monte Carlo algorithm and MCMC used when the likelihood is analytically intractable[178]
Evolutionary Monte Carlo	MCMC algorithm that incorporates features of genetic algorithms and simulated annealing.[229]
Other	
Sequential Monte Carlo	Algorithm based on multiple importance sampling steps for each observed data point - often used for on-line or real-time processing of data arrivals.[230]
Approximate Bayesian Computation	Approximate approach, typically used when the likelihood function is analytically intractable or very computationally expensive.[231]
Integrated nested Laplace approximations (INLA)	Approximate approach developed for the large class of latent Gaussian models, which includes, for example, generalized additive spline models, Gaussian Markov processes and random fields.[232]
Variational Bayes	Variational Inference describes a technique to approximate posterior distributions via simpler approximating distributions. Optimisation is used to adapt the variational parameters within these approximating distributions to make them as close to the true posterior distribution as possible using the KL-divergence as a measure of discrepancy[99].

Table 2. **A non-exhaustive summary of commonly used and open Bayesian software programs.**

Software Package	Summary	Type of sampling	System specifications
General-purpose Bayesian inference software			
BUGS[233-235] (Bayesian Inference Using Gibbs Sampler) / JAGS[236] (Just Another Gibbs Sampler)	The original general-purpose Bayesian inference engine, in different incarnations. Uses Gibbs and Metropolis sampling. Windows based software (WinBUGS[233]), with user-specified model and black-box MCMC algorithm. Developments include an open source version (OpenBUGS[235]) also available on Linux and Mac (using WINE); and parallel algorithm version (MultiBUGS[237]). R packages are available for calling BUGS from R (such as R2WinBUGS[238], R2OpenBUGS[238] and BRugs[239]). JAGS[236] (Just Another Gibbs Sampler) is an open source variation of BUGS which can run cross-platform and can run from R via rjags[236].	MCMC	OpenBUGS = Windows, Linux, Mac (using WINE) MultiBUGS = Windows JAGS = all platforms.
PyMC3[240]	Framework for Bayesian modeling and inference entirely within Python; includes Gibbs sampling and Hamiltonian Monte Carlo		
Stan[98]	General-purpose Bayesian inference engine using Hamiltonian Monte Carlo; can be run from R, Python, Julia, Matlab, and Stata. Open source software that implements efficient Hamiltonian Monte Carlo (HMC). Versions available for R, Python, MATLAB, Julia and Stata.	MCMC (Hamiltonian Monte Carlo)	All platforms
NIMBLE[241]	Generalization of the Bugs language in R; includes sequential Monte Carlo as well as MCMC. Open source R package using BUGS/JAGS-model	MCMC and sequential Monte carlo	All platfoms

	language to develop a model; and different algorithms for model fitting including MCMC and sequential Monte Carlo approaches including the ability to write novel algorithms.		
Programming languages that can be used for Bayesian inference			
`TensorFlow Probability`[242,243]	A Python library for probabilistic modelling built on Tensorflow[205] from Google.	MCMC	Python 3.5 – 3.8 Ubuntu 16.04 or later Windows 7 or later (with C++ redistributable) macOS 10.12.6 (Sierra) or later (no GPU support) Raspbian 9.0 or later
`Pyro`[244]	Probabilistic programming language built on `Python` and `PyTorch`[206].	MCMC	
`Julia`[245]	In addition to `Stan`, numerous other probabilistic programming libraries are available for the `Julia` programming language including `Turing.jl`[246] and `Mamba.jl`[247].	MCMC	Windows macOS Linux FreeBSD
Specialized software doing Bayesian inference for particular classes of models			
`JASP`[248] (Jeffreys's Amazing Statistics Program)	JASP is a user friendly higher-level interface, offering standard analysis procedures in both their classical and Bayesian form. It is open source and relies upon a collection of open-source R packages.		Windows MAC Linux
`R-INLA`[232]	Open source `R` package for implementing `INLA`.[249] Fast inference in `R` for a certain set of hierarchical models using nested Laplace approximations.	INLA	All platforms

GPstuff[250]	Fast approximate Bayesian inference for Gaussian processes using expectation propagation; runs in Matlab, Octave, and R.		Unix and Windows Matlab

Figures Headings

Figure 1. The Bayesian Research Cycle.

Typical steps needed for a research cycle using Bayesian statistics. The first part of the Bayesian Research Cycle, indicated with (A) is identical to any research cycle: starting with reading literature, defining a problem, specifying the research question and hypothesis[14,15]. The analytic strategy should be pre-registered to enhance transparency. The second part of the Bayesian Research Cycle, indicated with (B) is specifically for a Bayesian workflow. It includes formalizing prior distributions based on background knowledge and prior elicitation, determining the likelihood function by specifying a data generating model and including observed data, and obtaining the posterior distribution as a function of both the specified prior and likelihood function[135,251]. To probe the consequences of the specified model, it is important to perform robustness checks along the way and after. All concepts are briefly discussed in the primer with references for the interested user.

Figure 2. Illustration of the Key Ingredients of Bayes' Theorem.

This figure displays how the likelihood and prior work together to form the posterior distribution. Notice that the likelihood remains constant across all rows. Each row only differs in the prior distribution specified. Priors are typically deemed to be informative, weakly informative, or diffuse, each defined through different degrees of (un)certainty—in this case, through the variance (or spread) of the prior. The posterior distribution is a compromise between the prior and the likelihood.

Figure 3: Prior Predictive Checks.

Prior predictive checks for the PhD-delay example, computed via Stan[98] – the scripts are available at the Open Science Framework: https://osf.io/ja859/ - DOI 10.17605/OSF.IO/JA859 (A) displays a scenario in which precision was mistakenly used instead of variance for β_{age} and displays an unexpected pattern for the prior predictive distribution. Note, in dark blue the observed mean and SD are presented, in light blue samples of the prior predictive distribution. (B) shows the prior predictive distribution for the correct implementation of the hyperparameters. The prior predictive checks for the correct implementation of the priors seem reasonable given the data. Additionally, in panel C, a kernel density estimate of the

observed data is displayed (y - in dark blue), and kernel density estimates for the simulated data (y_{rep} - in light blue)[67]. [83]As can be seen the priors cover the entire plausible parameter space with the observed data in the center.

Figure 4. Posterior mean and SD estimation using MCMC

In panel (A) trace plots (iteration number against parameter value) for the PhD delay data, computed in Stan[98] of four independent MCMC algorithms are shown for exploring the same posterior distribution of $\beta_{intercept}$, with the first part omitted for constructing the posterior distribution (i.e, warm-up phase); In panel (B) the associated $\hat{R}$ statistic is shown which appears to settle down around the value of 1 after approximately 2,000 iterations; and (C, D and E) prior and posterior distributions for the in the model, the intercept (Panel C, $\beta_{intercept}$), the linear effect of age on PhD delay (Panel D, β_{age}), and the quadratic effect of age on PhD delay (Panel E, β_{age^2}). For each chain, the first 2,000 iterations are discarded as warm-up. The scripts are available at the Open Science Framework: https://osf.io/ja859/ - DOI 10.17605/OSF.IO/JA859.

Figure 5. Examples of shrinkage priors for Bayesian variable selection.

In Panel A, the discrete spike-and-slab prior for β_j (solid blue line) is specified as a mixture of a point mass at 0 (spike; dashed black line) and a flat prior (slab; dotted red line). In panel B, the continuous spike-and-slab prior for β_j (solid blue line) is specified as a mixture of two normal distributions, one peaked around 0 (dashed black line) and the other with a large variance (dotted red line). In panel C, the Bayesian lasso specifies a conditional Laplace prior, which can be obtained as a scale mixture of normal distributions with an exponential mixing density. This prior does not offer enough flexibility to allow simultaneously a lot of mass around zero and heavy tails. In panel D, the horseshoe prior falls in the class of global-local shrinkage priors, which are characterized by a high concentration around zero to shrink small coefficients and heavy tails to avoid excessive shrinkage of large coefficients.

Figure 6. Posterior Predictive Checking

Wikipedia page views for the premier league as obtained using the 'wikipediatrend'[141] R package and analyzed with the 'prophet'[143] R package. The scripts are available at the Open Science Framework: https://osf.io/7yrud/ - DOI 10.17605/OSF.IO/7YRUD. Panels show posterior means for the following

parameters along with 95% Cis for non-periodic changes (A), holiday effects (B), weekly seasonality (C), and yearly seasonality effects (D). In panels E and F posterior predictive distributions at each time point can be seen. The posterior predictive distributions for the time points that fall in the observed data interval on which the posterior distribution is conditioned, are displayed in light red (50% CI) and dark-red (95% CI). The corresponding observations are marked as black dots. Additionally, the posterior predictive distributions for future data are presented in light blue (50% CI) and dark-blue (95% CI). The actual realisations of these dates are marked as black triangles (F).

Figure 7. Elements of reproducibility in the research workflow

The figure shows good research practices across the research workflow that can contribute to reproducibility and demonstrates where the Bayesian research cycle (see Figure 1) and the WAMBS checklist (see Box 4) fit in the wider context of transparency in research. Not all elements are applicable to all types of research, e.g. preregistration is typically used for hypothesis-driven research but the specification of the prior and likelihood may be pre-registered. There can be legitimate reasons why not all data can be shared openly, but all scripts for running the Bayesian models could be shared on a data repository. *Note that part of the figure is based on a figure originally used in the Utrecht University Summerschool on Open Science and Scholarship 2019*[252] *(licensed CC-BY).*

Boxes

Box 1 | Bayes' Theorem

In Bayesian statistics, all observed and unobserved quantities in a system are given a joint probability distribution, and inference for unobserved quantities is based on their conditional distribution given the observed data. By construction, Bayesian inferences are optimal when averaged over this joint distribution; in Bayesian terminology, the prior and data distributions. Rényi's axiom of probability[253] lends itself to examining conditional probabilities, where the probabilities of Event A and Event B occurring are dependent, or conditional. The basic conditional probability can be written as:

$$p(B|A) = \frac{p(B \cap A)}{p(A)}, \qquad (1)$$

where the probability of Event B occurring is conditional on Event A. Equation 1 sets the foundation for Bayes' rule, which is a mathematical expression of Bayes' theorem that recognizes $p(B|A) \neq p(A|B)$ but $p(B \cap A) = p(A \cap B)$. Bayes' rule can be written as:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}, \qquad (2)$$

which, based on Equation 1, can be reworked as:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \qquad (3 – \text{Bayes' rule})$$

These principles can be extended to the situation of data and model parameters. With dataset $\boldsymbol{y}$ and model parameters $\boldsymbol{\theta}$, Equation 3 (Bayes'rule) can be written as follows:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \qquad (4)$$

which is often simplified to:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \qquad (5)$$

The term $p(\boldsymbol{\theta}|\boldsymbol{y})$ represents a conditional probability, where the probability of the model parameters ($\boldsymbol{\theta}$) is computed conditional upon the data ($\boldsymbol{y}$), and this term is also known as the *posterior*. The term $p(\boldsymbol{y}|\boldsymbol{\theta})$ represents the conditional probability of the data given the model parameters, and this term represents the *data likelihood*. Finally, the term $p(\boldsymbol{\theta})$ represents the probability of particular model parameter values existing in the population. This term is called a *prior.* The term $p(\boldsymbol{y})$ is often viewed as a normalizing factor across all outcomes $\boldsymbol{y}$, which can be removed from the equation because $\boldsymbol{\theta}$ does not depend on $\boldsymbol{y}$ or $p(\boldsymbol{y})$. Given that $p(\boldsymbol{y})$ is not needed for the posterior, it can be removed, and we say that the posterior is

proportional to ($\propto$) the likelihood times the prior.. Figure 2 illustrates the relationship between the likelihood, prior, and posterior.

Box 2 | The likelihood function for a coin experiment

Consider the following textbook example: we are given a coin and want to know what the probability of obtaining "heads" (θ) is. To examine this, we toss the coin a number of times and count the number of heads. Let the outcome of the ith flip be denoted by $h_i = 1$ for heads and $h_i = 0$ for tails. The total experiment yields a sample of n independent binary observations $\{h_1, \dots, h_n\} = \boldsymbol{h}$ with y as the total number of heads; $y = \sum_{i=1}^{n} h_i$. We can assume that the probability to obtain heads remains constant over the experiment, i.e. $p(h_i) = \theta, (\mathrm{i} = 1, \dots, \mathrm{n})$. Therefore the probability of the observed number of heads is expressed by the binomial distribution, given by

$$P(\mathrm{y}|\theta) = \binom{n}{h}\theta^h(1-\theta)^{n-h}\,,\ 0 \leq \theta \leq 1 \qquad (1)$$

[254].

When y is kept fixed and θ is varying, $P(\mathrm{y}|\theta)$ becomes a continuous function of $\boldsymbol{\theta}$, called the binomial likelihood function[254].

Suppose we flipped the coin 10 times and observed 4 heads, the likelihood function of $\boldsymbol{\theta}$ is defined by

$$f(\mathrm{y}|\theta) = \binom{10}{4}\theta^4(1-\theta)^6\,,\ 0 \leq \theta \leq 1. \qquad (2)$$

.

Box 3 | Bayes Factors

Hypothesis testing consists of using data to evaluate the evidence for competing claims or hypotheses. In the Bayesian framework, this can be accomplished using the Bayes factor, which corresponds to the ratio of the posterior odds to the prior odds of distinct hypotheses[38,62]. For two hypotheses, H_0 and H_1, and observed data $\boldsymbol{y}$, the Bayes factor in favor of H_1 is given by

$$BF_{10} = \frac{p(H_1|\boldsymbol{y})/p(H_0|\boldsymbol{y})}{p(H_1)/p(H_0)}, \qquad (6)$$

where $p(H_0)$ and $p(H_1) = 1 - p(H_0)$ are the prior probabilities. A larger value of BF_{10} provides stronger evidence against H_0 [62]. The posterior probability $p(H_j|\boldsymbol{y})$ is obtained using Bayes theorem

$$p(H_j|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|H_j)p(H_j)}{f(\boldsymbol{y})}, j = 0,1. \qquad (7)$$

Thus, the Bayes factor can equivalently be written as the ratio of the marginal likelihoods of the observed data under the two hypotheses

$$BF_{10} = \frac{f(\boldsymbol{y}|H_1)}{f(\boldsymbol{y}|H_0)}. \qquad (8)$$

The competing hypotheses can take various forms and could be, for example, two non-nested regression models (see Variable Selection subsection). If H_0 and H_1 are simple hypotheses in which the parameters are fixed (e.g., $H_0: \mu = \mu_0$ versus $H_1: \mu = \mu_1$), the Bayes factor is identical to the likelihood ratio test. When either or both hypotheses are composite (i.e., not simple) or there are additional unknown parameters, the marginal likelihood $f(\boldsymbol{y}|H_j)$ is obtained by integrating over the parameters $\boldsymbol{\theta}_j$ with prior densities $p(\boldsymbol{\theta_j}|H_j)$

$$f(\boldsymbol{y}|H_j) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_j, H_j)\, p(\boldsymbol{\theta_j}|H_j)\, d\boldsymbol{\theta}_j. \qquad (9)$$

This integral is often intractable and must be computed by numerical methods. If $p(\boldsymbol{\theta_j}|H_j)$ is improper (i.e., $\int p(\boldsymbol{\theta_j}|H_j)\, d\boldsymbol{\theta}_j = \infty$) then $f(\boldsymbol{y}|H_j)$ will be improper and the Bayes factor will not be uniquely defined. Overly diffuse priors should also be avoided, as they result in a Bayes factor that favors H_0 regardless of the information in the data[104]. As a simple illustrative example, suppose one collects n random samples from a normally distributed population with an unknown mean μ and a known variance

σ^2, and wishes to test H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$. Let $\bar{y}$ be the sample mean. H_0 is a simple hypothesis with a point mass at μ_0 , so $\bar{y}|H_0 \sim N(\mu_0, \sigma^2/n)$. Under H_1, $\bar{y}|\mu, H_1 \sim N(\mu, \sigma^2/n)$ and assuming $\mu|H_1 \sim N(\mu_0, \tau^2)$ with τ^2 fixed, then $f(\bar{y}|H_1) = \int f(\boldsymbol{y}|\mu, H_1)\, p(\mu|H_1)\, d\mu$ reduces to $\bar{y}|H_1 \sim N(\mu_0, \tau^2 + \sigma^2/n)$. Thus, the Bayes factor in favor of H_1 is

$$BF_{10} = \frac{f(\boldsymbol{y}|H_1)}{f(\boldsymbol{y}|H_0)} = \frac{(\tau^2 + \sigma^2/n)^{-1/2} \exp\left\{-\frac{(\bar{y} - \mu_0)^2}{2(\tau^2 + \sigma^2/n)}\right\}}{(\sigma^2/n)^{-1/2} \exp\left\{-\frac{(\bar{y} - \mu_0)^2}{2(\sigma^2/n)}\right\}} \quad (10)$$

For example, for $n = 20$, $\bar{y} = 5.8$, $\mu_0 = 5$, $\sigma^2 = 1$ and $\tau^2 = 1$, the Bayes factor is $BF_{10} = 96.83$, which provides strong evidence that the mean μ is not 5.

Box 4 | WAMBS-Checklist

[bH1] The 10 checklist points of WAMBS-v2

[b1] Ensure the prior distributions and the model (or likelihood) are well understood and described in detail in the text, including the hyperparameter settings and all details surrounding the model. In addition, prior-predictive checking can help identify any prior-data conflict.

[b1] Assess each parameter for convergence. Use multiple convergence diagnostics if possible. This may involve examining trace-plots or ensuring diagnostics (e.g., $\hat{R}$ or effective sample size) are being met for each parameter. For example, $\hat{R}$ values smaller than 1.05 are typically recommended. Likewise, effective sample sizes of 10,000 or more are recommended as a general rule of thumb.

[b1] Sometimes convergence diagnostics can fail at detecting non-convergence within the chain. Subsequent measures, such as the split-$\hat{R}$ can be used to identify such situations. The split-$\hat{R}$ can detect trends that are missed if the chains have similar marginal distributions (the $\hat{R}$ may miss these trends).

[b1] Ensure that there were sufficient chain iterations to construct a meaningful posterior distribution. The posterior distribution should consist of enough samples to visually examine the shape, scale, and central tendency of the distribution. Without enough samples, there is an incomplete picture of the full distribution.

[b1] Check all parameters for strong degrees of autocorrelation (e.g., through examining the effective sample size for parameters), which may be a sign of model or prior misspecification.

[b1] Visually examine the marginal posteriors distribution for each model parameter to ensure that they make substantive sense. Posterior predictive distributions can be used to aid in examining the posteriors.

[b1] Fully examine multivariate priors through a sensitivity analysis. These priors can be particularly influential on the posterior, even with slight modifications to the hyperparameters.

[b1] To fully understand the impact of subjective priors, compare the posterior results to an analysis using diffuse (or objective) priors. This comparison can facilitate a deeper understanding of the impact the subjective priors (i.e., the theory being implemented) are having on findings. Next, conduct a full

sensitivity analysis of all priors to gain a clearer understanding of the robustness of the results to different prior settings.

[b1] Given the subjectivity of the model, it is also important to conduct a sensitivity analysis of the model (or likelihood) to help uncover how robust results are to deviations in the model.

[b1] Report findings by including Bayesian interpretations. Take advantage of explaining and capturing the entire posterior rather than simply a point estimate. For example, it may be helpful to examine the density at different quantiles to fully capture and understand the posterior distribution.

Glossary Terms

Prior distribution: Beliefs held by researchers about the parameters in a statistical model BEFORE seeing the data.

Hyperparameters: Hyperparameters are the parameters that define the prior distribution. For example, the normal distribution is defined through a mean and variance, and these are referred to as the hyperparameters.

Informative prior: Informative priors reflect a high degree of certainty or knowledge surrounding the population parameters and the hyperparameters are specified to express particular information reflecting a greater degree of certainty about the model parameters being estimated

Weakly informative prior: The weakly informative prior incorporates some information about the population parameter but are not as restrictive as an informative prior.; some researchers find this to be a nice middle ground regarding the informativeness of the prior

Diffuse priors: Diffuse priors reflect complete uncertainty about population parameters.

Shrinkage priors A specific prior that shrinks the posterior estimate towards a particular value.

Spike-and-slab prior A specific shrinkage prior distribution used for variable selection that corresponds to a mixture of two distributions, one spiked around 0 and the other with a large variance corresponding to the slab component.

Horseshoe prior A prior for variable selection that uses a half-Cauchy scale mixture of normal distribution. This prior is characterized by a high concentration around zero to shrink small coefficients and heavy tails to avoid excessive shrinkage of large parameters.

Prior predictive distribution All possible samples that could occur if the model is true based on the priors. In theory, a "correct" prior provides a prior predictive distribution similar to the true data generating distribution

Prior predictive p-value An estimate to indicate how unlikely the observed data is to be generated by the model based on the prior predictive distribution

likelihood function The conditional probability distribution $p(y|\theta)$ of the data y given parameters θ.

posterior distribution The posterior distribution reflects one's updated knowledge, balancing prior knowledge with observed data.

Markov chain Monte Carlo (MCMC) = A method to indirectly obtain inference on the posterior distribution via simulation which combines two concepts: (i) obtain a set of parameter values from the posterior distribution (using the Markov chain, or the first "MC"); and (ii) given sampled parameter values obtain a distributional estimate of the posterior and associated posterior statistics of interest (using Monte Carlo, or the second "MC").

Trace plots A plot describing the posterior parameter value at each iteration of the Markov chain (on the y-axis) against iteration number (on the x-axis)

$\hat{R}$ statistic $\hat{R}$ is defined to be the ratio of the within- and between-chain variability. Values close to 1 for all parameters and quantities of interest suggest the chain has sufficiently converged to the stationary distribution

Bayes factor Bayes factors (Box 3) can be used to compare and choose between candidate models, where each candidate model would correspond to a hypothesis

kernel density estimation A kernel density estimation is a non-parametric approach used to estimate a probability density function for the observed data.

transition kernel determines the performance of the MCMC algorithm in terms of how long the Markov chain needs to be run to obtain reliable inference on the posterior distribution of interest.

auxiliary variables additional variables entered in the model to improve the missing data model.

sparsity: indicates that most parameter values are zero and only a few are non-zero.

Stochastic Gradient Descent (SGD) algorithm. SGD algorithms use a randomly chosen subset of data points to estimate the gradient of a loss function with respect to parameters. This can provide radical computational savings in optimisation problems involving many data points.

Variational Inference (VI). Variational methods refers to a class of approximate inference techniques in which deterministic posterior approximations are constructed from a family of predefined distributions. These approximations contain variational parameters which are optimised to match the approximating distribution as closely as possible to the true posterior. They are popular methods for achieving scalable but approximate Bayesian inference in large data scenarios where MCMC sampling-based inference would be prohibitive.

Highlighted Refences

1. **O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J.,..... & Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.**
   This book is a great collection of information with respect to prior elicitation. It includes elicitation techniques, summarizes potential pitfalls, and describes examples across a wide variety of disciplines.

2. **E.J. George and R.E. McCulloch (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88: 881-889.**
   This is the paper that popularized the use of spike-and-slab priors for Bayesian variable selection and introduced MCMC techniques to explore the model space.

3. **N.G. Polson and J.G. Scott (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. Bayesian Statistics 9, 9: 501-538.**
   This paper provides a unified framework for continuous shrinkage priors, which allow global sparsity while controlling the amount of regularization for each regression coefficient.

4. **M.G. Tadesse and M. Vannucci (2020). Handbook of Mixture Analysis. CRC Press, Chapman \& Hall/CRC Handbooks of Modern Statistical Methods, in preparation.**
   This is a forthcoming edited book that presents a comprehensive review of Bayesian variable selection methods and highlights recent developments.

5. **Gelfand, A. E. & Smith, A. F. M. Sampling-Based Approaches to Calculating Marginal Densities. *J. Am. Stat. Assoc.* 85, 398-409, doi:10.1080/01621459.1990.10476213 (1990).**
   Seminal paper that identified Markov chain Monte Carlo as a practical approach for Bayesian inference.

6. **Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10, 325–337**
   Provided an early user-friendly and freely-available black-box MCMC sampler opening up Bayesian inference to the wider scientific community.

7. **Brooks, S. P., Gelman, A., Jones, G., Meng, X. (Eds) (2011) *Handbook of Markov chain Monte Carlo*. CRC Press.**
   Comprehensive review of Markov chain Monte Carlo and its use in many different applications.

8. **Depaoli, S., and van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods, 22,* 240-261.**
This paper goes through, in a step-by-step manner, the various points that need to be checked when estimating a model via Bayesian statistics. It can be used as a guide for implementing Bayesian methods.

9. **Kass, R.E., Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90: 773-795.**
This paper provides an extensive discussion of Bayes factors with several examples.

10. **Blei, D. M., et al. (2017). "Variational inference: A review for statisticians." Journal of the American statistical Association 112(518): 859–877.**
Recent review of variational inference methods, including stochastic variants, which underpin popular approximate Bayesian inference methods for large data or complex modelling problems where computation using MCMC stochastic simulation would be prohibitively costly.

11. **Kingma, D. P. and M. Welling (2019). An Introduction to Variational Autoencoders.**
Recent review of variational autoencoders, encompassing deep generative models, the reparameterisation trick and current inference methods. These are an important class of models in modern Bayesian machine learning that combines the use of Bayesian modelling with deep neural networks for flexible function parameterisation.

12. **Neal, R. M. (1996). Priors for Infinite Networks. Bayesian Learning for Neural Networks. R. M. Neal. New York, NY, Springer New York: 29-53.**
A classic text highlighting the connection between neural networks and Gaussian processes and the application of Bayesian approaches for fitting neural networks.

13. **Berger, J. (2006). The case for objective Bayesian analysis. Bayesian Analysis, 1(3), 385-402.**
"A discussion of objective Bayesian analysis, including criticisms of the approach and a personal perspective on the debate on the value of objective Bayesian versus subjective Bayesian analysis."

14. **Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. 1098 Journal of the American statistical Association, 82(398), 528-540.**
"In this article the authors explain how to use data augmentation when direct computation of the posterior density of the parameters of interest is not possible."

1442

References

1 Bayes, M. & Price, M. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London* **53**, 370-418, doi:10.1098/rstl.1763.0053 (1997).

2 Laplace, P. S. *Essai Philosophique sur les Probabilities*. (Courcier, 1814).

3 König, C. & van de Schoot, R. Bayesian statistics in educational research: a look at the current state of affairs. *Educational Review*, 1-24 (2017).

4 van de Schoot, R., Winter, S., Zondervan-Zwijnenburg, M., Ryan, O. & Depaoli, S. A systematic review of Bayesian applications in psychology: The last 25 years. *Psychological Methods* **22**, 217-239 (2017).

5 Ashby, D. Bayesian statistics in medicine: a 25 year review. *Stat Med* **25**, 3589-3631, doi:10.1002/sim.2672 (2006).

6 Rietbergen, C., Debray, T. P. A., Klugkist, I., Janssen, K. J. M. & Moons, K. G. M. Reporting of Bayesian analysis in epidemiologic research should become more transparent. *J Clin Epidemiol* **86**, 51-58 e52, doi:10.1016/j.jclinepi.2017.04.008 (2017).

7 Spiegelhalter, D. J., Myles, J. P., Jones, D. R. & Abrams, K. R. Bayesian methods in health technology assessment: a review. *Health Technol Assess* **4**, 1-130, doi:10.3310/hta4380 (2000).

8 Kruschke, J. K., Aguinis, H. & Joo, H. The Time Has Come Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods* **15**, 722-752 (2012).

9 Smid, S. C., McNeish, D., Miočević, M. & van de Schoot, R. Bayesian Versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review. *Structural Equation Modeling: A Multidisciplinary Journal* **27**, 131-161, doi:10.1080/10705511.2019.1577140 (2019).

10 Rupp, A. A., Dey, D. K. & Zumbo, B. D. To bayes or not to bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling* **11**, 424-451 (2004).

11 Depaoli, S. & Clifton, J. P. A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal* **22**, 327-351 (2015).

12 Kim, S.-Y., Suh, Y., Kim, J.-S., Albanese, M. A. & Langer, M. M. Single and multiple ability estimation in the SEM framework: A noninformative Bayesian estimation approach. *Multivariate behavioral research* **48**, 563-591 (2013).

13 Depaoli, S. Mixture class recovery in GMM under varying degrees of class separation: frequentist versus Bayesian estimation. *Psychol Methods* **18**, 186-219, doi:10.1037/a0031609 (2013).

14 Blaxter, L. *How to research*. (McGraw-Hill Education (UK), 2010).

15 Neuman, W. L. *Understanding research*. (Pearson, 2016).

16 Heo, I. & Van de Schoot, R. Tutorial: Advanced Bayesian regression in JASP. (2020).

17 Van de Schoot, R., Yerkes, M. A., Mouw, J. M. & Sonneveld, H. What took them so long? Explaining PhD delays among doctoral candidates. *PloS one* **8**, e68839 (2013).

18 Muthen, B. & Asparouhov, T. Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol Methods* **17**, 313-335, doi:10.1037/a0026802 (2012).

19 van de Schoot, R. *et al.* Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front Psychol* **4**, 770, doi:10.3389/fpsyg.2013.00770 (2013).

20 O'Hagan, A. *et al. Uncertain judgements: eliciting experts' probabilities*. (John Wiley & Sons, 2006).
21 Howard, G. S., Maxwell, S. E. & Fleming, K. J. The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychol Methods* **5**, 315-332, doi:10.1037/1082-989x.5.3.315 (2000).
22 Veen, D., Stoel, D., Zondervan-Zwijnenburg, M. & van de Schoot, R. Proposal for a Five-Step Method to Elicit Expert Judgement. *Frontiers in psychology* **8**, 2110 (2017).
23 Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T. & Feldman, B. M. Methods to elicit beliefs for Bayesian priors: a systematic review. *J Clin Epidemiol* **63**, 355-369, doi:10.1016/j.jclinepi.2009.06.003 (2010).
24 Ibrahim, J. G., Chen, M. H., Gwon, Y. & Chen, F. The power prior: theory and applications. *Stat Med* **34**, 3724-3749, doi:10.1002/sim.6728 (2015).
25 Rietbergen, C., Klugkist, I., Janssen, K. J., Moons, K. G. & Hoijtink, H. J. Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemp Clin Trials* **32**, 848-855, doi:10.1016/j.cct.2011.06.002 (2011).
26 van de Schoot, R. *et al.* Bayesian PTSD-Trajectory Analysis with Informed Priors Based on a Systematic Literature Search and Expert Elicitation. *Multivariate Behav Res* **53**, 267-291, doi:10.1080/00273171.2017.1412293 (2018).
27 Smeets, L. & van de Schoot, R. Code for the ShinyApp to Determine the Plausible Parameter Space for the PhD-delay Data (Version v1.0). . (2020).
28 Berger, J. The case for objective Bayesian analysis. *Bayesian analysis* **1**, 385-402 (2006).
29 Brown, L. D. In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, 113-152 (2008).
30 Candel, M. J. & Winkens, B. Performance of empirical Bayes estimators of level-2 random parameters in multilevel analysis: A Monte Carlo study for longitudinal designs. *Journal of Educational and Behavioral Statistics* **28**, 169-194 (2003).
31 van der Linden, W. J. Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics* **33**, 5-20 (2008).
32 Darnieder, W. F. *Bayesian methods for data-dependent priors*, The Ohio State University, (2011).
33 Richardson, S. & Green, P. J. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 731-792, doi:10.1111/1467-9868.00095 (1997).
34 Wasserman, L. Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 159-180, doi:10.1111/1467-9868.00226 (2000).
35 Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J. & Dorie, V. Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics* **40**, 136-157 (2015).
36 Gelman, A., Jakulin, A., Pittau, M. G. & Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics* **2**, 1360-1383 (2008).
37 Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian data analysis*. Vol. 2 (Chapman&HallCRC, 2004).
38 Jeffreys, H. *Theory of probability*. Vol. 3 (Clarendon Press, 1961).
39 Seaman III, J. W., Seaman Jr, J. W. & Stamey, J. D. Hidden dangers of specifying noninformative priors. *The American Statistician* **66**, 77-84, doi:https://doi.org/10.1080/00031305.2012.695938 (2012).
40 Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis* **1**, 515-534 (2006).

41 Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R. & Jones, D. R. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med* **24**, 2401-2428, doi:10.1002/sim.2112 (2005).
42 McNeish, D. On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal* **23**, 750-773 (2016).
43 van de Schoot, R. & Miocević, M. *Small sample size solutions: A guide for applied researchers and practitioners*. (Taylor & Francis, 2020).
44 Schuurman, N. K., Grasman, R. P. & Hamaker, E. L. A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models. *Multivariate Behav Res* **51**, 185-206, doi:10.1080/00273171.2015.1065398 (2016).
45 Liu, H., Zhang, Z. & Grimm, K. J. Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal* **23**, 354-367 (2016).
46 Ranganath, R. & Blei, D. M. Population predictive checks. *arXiv preprint arXiv:1908.00882* (2019).
47 Daimon, T. Predictive checking for Bayesian interim analyses in clinical trials. *Contemp Clin Trials* **29**, 740-750, doi:10.1016/j.cct.2008.05.005 (2008).
48 Morris, D. E., Oakley, J. E. & Crowe, J. A. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software* **52**, 1-4 (2014).
49 Garthwaite, P. H., Al-Awadhi, S. A., Elfadaly, F. G. & Jenkinson, D. J. Prior distribution elicitation for generalized linear and piecewise-linear models. *Journal of Applied Statistics* **40**, 59-75 (2013).
50 Elfadaly, F. G. & Garthwaite, P. H. Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models. *Statistics and Computing* **27**, 449-467 (2017).
51 Veen, D., Egberts, M. R., van Loey, N. E. E. & van de Schoot, R. Expert Elicitation for Latent Growth Curve Models: The Case of Posttraumatic Stress Symptoms Development in Children With Burn Injuries. *Front Psychol* **11**, 1197, doi:10.3389/fpsyg.2020.01197 (2020).
52 Runge, A. K., Scherbaum, F., Curtis, A. & Riggelsen, C. An interactive tool for the elicitation of subjective probabilities in probabilistic seismic-hazard analysis. *Bulletin of the Seismological Society of America* **103**, 2862-2874 (2013).
53 Zondervan-Zwijnenburg, M., van de Schoot-Hubeek, W., Lek, K., Hoijtink, H. & van de Schoot, R. Application and Evaluation of an Expert Judgment Elicitation Procedure for Correlations. *Front Psychol* **8**, 90, doi:10.3389/fpsyg.2017.00090 (2017).
54 Cooke, R. M. & Goossens, L. H. J. TU Delft expert judgment data base. *Reliability Engineering & System Safety* **93**, 657-674 (2008).
55 Hanea, A. M., Nane, G. F., Bedford, T. & French, S. *Expert Judgment in Risk and Decision Analysis*. (Springer, 2020).
56 Dias, L. C., Morton, A. & Quigley, J. *Elicitation*. (Springer, 2018).
57 Box, G. E. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)* **143**, 383-404 (1980).
58 Nott, D. J., Drovandi, C. C., Mengersen, K. & Evans, M. Approximation of Bayesian Predictive p-Values with Regression ABC. *Bayesian Analysis* **13**, 59-83 (2018).
59 Evans, M. & Moshonov, H. Checking for prior-data conflict with hierarchically specified priors. *Bayesian statistics and its applications*, 145-159 (2007).
60 Evans, M. & Jang, G. H. A limit result for the prior predictive applied to checking for prior-data conflict. *Statistics & Probability Letters* **81**, 1034-1038, doi:10.1016/j.spl.2011.02.025 (2011).
61 Young, K. & Pettit, L. Measuring discordancy between prior and data. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 679-689 (1996).
62 Kass, R. E. & Raftery, A. E. Bayes factors. *Journal of the american statistical association* **90**, 773-795 (1995).

63 Bousquet, N. Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics* **35**, 1011-1029, doi:https://doi.org/10.1080/02664760802192981 (2008).

64 Veen, D., Stoel, D., Schalken, N., Mulder, K. & van de Schoot, R. Using the Data Agreement Criterion to Rank Experts' Beliefs. *Entropy* **20**, 592, doi:10.3390/e20080592 (2018).

65 Nott, D. J., Xueou, W., Evans, M. & Englert, B. Checking for prior-data conflict using prior to posterior divergences. *arXiv preprint arXiv:1611.00113* (2016).

66 Lek & Van De, S. How the Choice of Distance Measure Influences the Detection of Prior-Data Conflict. *Entropy* **21**, 446, doi:10.3390/e21050446 (2019).

67 Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **182**, 389-402 (2019).

68 O'Hagan, A. Bayesian statistics: principles and benefits. *Frontis*, 31-45 (2004).

69 Etz, A. Introduction to the Concept of Likelihood and Its Applications. *Advances in Methods and Practices in Psychological Science* **1**, 60-69, doi:10.1177/2515245917744314 (2018).

70 Pawitan, Y. *In all likelihood: statistical modelling and inference using likelihood*. (Oxford University Press, 2001).

71 Gelman, A., Simpson, D. & Betancourt, M. The prior can often only be understood in the context of the likelihood. *Entropy* **19**, 555, doi:https://doi.org/10.3390/e19100555s (2017).

72 Aczel, B. *et al.* Discussion points for Bayesian inference. *Nat Hum Behav* **4**, 561-563, doi:10.1038/s41562-019-0807-z (2020).

73 Gelman, A. *et al. Bayesian data analysis*. (CRC press, 2013).

74 Greco, L., Racugno, W. & Ventura, L. Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference* **138**, 1258-1270, doi:10.1016/j.jspi.2007.05.001 (2008).

75 Shyamalkumar, N. D. in *Robust Bayesian Analysis Lecture Notes in Statistics* Ch. Chapter 7, 127-143 (Springer, 2000).

76 Agostinelli, C. & Greco, L. A weighted strategy to handle likelihood uncertainty in Bayesian inference. *Computational Statistics* **28**, 319-339 (2013).

77 Rubin, D. B. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 1151-1172 (1984).

78 Gelfand, A. E. & Smith, A. F. M. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **85**, 398-409, doi:10.1080/01621459.1990.10476213 (1990).

79 Geyer, C. J. Markov chain Monte Carlo maximum likelihood. 156-163 (1991).

80 Van de Schoot, R., Veen, D., Smeets, L., Winter, S. D. & Depaoli, S. A tutorial on using the WAMBS checklist to avoid the misuse of Bayesian statistics. *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners; van de Schoot, R., Miocevic, M., Eds*, 30-49 (2020).

81 Veen, D. & Egberts, M. The Importance of Collaboration in Bayesian Analyses with Small Samples. *SMALL SAMPLE SIZE SOLUTIONS*, 50 (2020).

82 Robert, C. & Casella, G. *Monte Carlo statistical methods*. (Springer Science & Business Media, 2013).

83 Silverman, B. W. *Density estimation for statistics and data analysis*. Vol. 26 (CRC press, 1986).

84 Geman, S. & Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* **6**, 721-741, doi:10.1109/tpami.1984.4767596 (1984).

85 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**, 1087-1092 (1953).

86 Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109 (1970).

87 Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. Hybrid monte carlo. *Physics letters B* **195**, 216-222 (1987).
88 Tanner, M. A. & Wong, W. H. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association* **82**, 528-540 (1987).
89 Gelman, A. *Burn-in for MCMC, why we prefer the term warm-up*, <https://statmodeling.stat.columbia.edu/2017/12/15/burn-vs-warm-iterative-simulation-algorithms/> (2017).
90 Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457-511 (1992).
91 Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434-455 (1998).
92 Roberts, G. O. Markov chain concepts related to sampling algorithms. *Markov chain Monte Carlo in practice* **57**, 45-58 (1996).
93 Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P. (2020).
94 Gamerman, D. & Lopes, H. F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. (CRC Press, 2006).
95 Brooks, S. P., Gelman, A., Jones, G. & Meng, X.-L. *Handbook of markov chain monte carlo*. (CRC press, 2011).
96 Bürkner, P.-C. Advanced Bayesian multilevel modeling with the R package brms. *arXiv preprint arXiv:1705.11123* (2017).
97 Merkle, E. C. & Rosseel, Y. blavaan: Bayesian structural equation models via parameter expansion. *arXiv preprint arXiv:1511.05604* (2015).
98 Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76**, doi:10.18637/jss.v076.i01 (2017).
99 Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association* **112**, 859–877 (2017).
100 Minka, T. P. 362–369.
101 Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research* **14**, 1303–1347 (2013).
102 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
103 Li, Y., Hernández-Lobato, J. M. & Turner, R. E. 2323–2331.
104 Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**, 410-423 (2008).
105 Forte, A., Garcia-Donato, G. & Steel, M. Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *International Statistical Review* **86**, 237-258 (2018).
106 Mitchell, T. J. & Beauchamp, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical association* **83**, 1023-1032 (1988).
107 George, E. J. & McCulloch, R. E. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889-889 (1993).
108 Ishwaran, H. & Rao, J. S. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* **33**, 730-773 (2005).
109 Bottolo, L. & Richardson, S. Evolutionary stochastic search. *Bayesian Analysis* **5**, 583-618 (2010).
110 Ročková, V. & George, E. I. EMVS: the EM approach to Bayesian variable selection. *Journal of the American Statistical Association* **109**, 828-846 (2014).
111 Park, T. & Casella, G. The Bayesian lasso. *Journal of the American Statistical Association* **103**, 681-686 (2008).

112 Carvalho, C. M., Polson, N. G. & Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika* **97**, 465?480-465?480 (2010).

113 Polson, N. G. & Scott, J. G. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian statistics* **9**, 105 (2010).

114 Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267-288 (1996).

115 Van Erp, S., Oberski, D. L. & Mulder, J. Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology* **89**, 31-50 (2019).

116 Brown, P. J., Vannucci, M. & Fearn, T. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60**, 627-641 (1998).

117 Lee, K. H., Tadesse, M. G., Baccarelli, A. A., Schwartz, J. & Coull, B. A. Multivariate Bayesian variable selection exploiting dependence structure among outcomes: Application to air pollution effects on DNA methylation. *Biometrics* **73**, 232-241 (2017).

118 Frühwirth-Schnatter, S. & Wagner, H. Stochastic model specification search for Gaussian and partially non-Gaussian state space models. *Journal of Econometrics* **154**, 85-100 (2010).

119 Scheipl, F., Fahrmeir, L. & Kneib, T. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association* **107**, 1518-1532 (2012).

120 Tadesse, M. G., Sha, N. & Vannucci, M. Bayesian variable selection in clustering high dimensional data. *Journal of the American Statistical Association*, 602-617 (2005).

121 Wang, H. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis* **10**, 351-377 (2015).

122 Peterson, C. B., Stingo, F. C. & Vannucci, M. Bayesian Inference of Multiple Gaussian Graphical Models. *J Am Stat Assoc* **110**, 159-174, doi:10.1080/01621459.2014.896806 (2015).

123 Tadesse, M. G. & Vannucci, M. *Handbook of Bayesian variable selection*. (CRC Press, 2020).

124 Li, F. & Zhang, N. R. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical association* **105**, 1978-2002 (2010).

125 Stingo, F., Chen, Y., Tadesse, M. G. & Vannucci, M. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Annals of Applied Statistics* **5**, 1202-1214 (2011).

126 Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics* **5**, 1780-1815 (2011).

127 Bottolo, L. *et al.* GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genetics* **9(8)**, e1003657-e1003657 (2013).

128 Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical modeling and analysis for spatial data*. (CRC press, 2014).

129 Vock, L. F. B., Reich, B. J., Fuentes, M. & Dominici, F. Spatial variable selection methods for investigating acute health effects of fine particulate matter components. *Biometrics* **71**, 167-177 (2015).

130 Penny, W. D., Trujillo-Barreto, N. J. & Friston, K. J. Bayesian fMRI time series analysis with spatial priors. *Neuroimage* **24**, 350-362, doi:10.1016/j.neuroimage.2004.08.034 (2005).

131 Smith, M., Pütz, B., Auer, D. & Fahrmeir, L. Assessing brain activity through spatial Bayesian variable selection. *Neuroimage* **20**, 802-815 (2003).

132 Zhang, L., Guindani, M., Versace, F. & Vannucci, M. A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. *Neuroimage* **95**, 162-175, doi:10.1016/j.neuroimage.2014.03.024 (2014).

133 Gorrostieta, C., Fiecas, M., Ombao, H., Burke, E. & Cramer, S. Hierarchical vector auto-regressive models and their applications to multi-subject effective connectivity. *Frontiers on Computational Neurosciences* **7**, 159-159 (2013).
134 Chiang, S. *et al.* Bayesian vector autoregressive model for multi-subject effective connectivity inference using multi-modal neuroimaging data. *Human Brain Mapping* **38**, 1311-1332 (2017).
135 Schad, D. J., Betancourt, M. & Vasishth, S. Toward a principled Bayesian workflow in cognitive science. *arXiv preprint arXiv:1904.12765* (2019).
136 Gelman, A., Meng, X.-L. & Stern, H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733-760 (1996).
137 Meng, X.-L. Posterior predictive p-values. *The annals of statistics* **22**, 1142-1160 (1994).
138 Asparouhov, T., Hamaker, E. L. & Muthén, B. Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal* **25**, 359-388 (2018).
139 Zhang, Z., Hamaker, E. L. & Nesselroade, J. R. Comparisons of four methods for estimating a dynamic factor model. *Structural Equation Modeling: A Multidisciplinary Journal* **15**, 377-402 (2008).
140 Hamaker, E., Ceulemans, E., Grasman, R. & Tuerlinckx, F. Modeling affect dynamics: State of the art and future challenges. *Emotion Review* **7**, 316-322 (2015).
141 Meissner, P. *wikipediatrend: Public Subject Attention via Wikipedia Page View Statistics*. (2019).
142 Harvey, A. C. & Peters, S. Estimation procedures for structural time series models. *Journal of Forecasting* **9**, 89-108 (1990).
143 Taylor, S. J. & Letham, B. Forecasting at scale. *The American Statistician* **72**, 37-45 (2018).
144 Gopnik, A. & Bonawitz, E. Bayesian models of child development. *Wiley Interdiscip Rev Cogn Sci* **6**, 75-86, doi:10.1002/wcs.1330 (2015).
145 Gigerenzer, G. & Hoffrage, U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review* **102**, 684 (1995).
146 Slovic, P. & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational behavior and human performance* **6**, 649-744 (1971).
147 Hoijtink, H., Beland, S. & Vermeulen, J. A. Cognitive diagnostic assessment via Bayesian evaluation of informative diagnostic hypotheses. *Psychol Methods* **19**, 21-38, doi:10.1037/a0034176 (2014).
148 Lee, M. D. How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology* **55**, 1-7 (2011).
149 Depaoli, S., Rus, H. M., Clifton, J. P., van de Schoot, R. & Tiemensma, J. An introduction to Bayesian statistics in health psychology. *Health Psychol Rev* **11**, 248-264, doi:10.1080/17437199.2017.1343676 (2017).
150 Kruschke, J. K. Bayesian estimation supersedes the t test. *J Exp Psychol Gen* **142**, 573-603, doi:10.1037/a0029146 (2013).
151 Bolt, D. M., Piper, M. E., Theobald, W. E. & Baker, T. B. Why two smoking cessation agents work better than one: Role of craving suppression. *Journal of Consulting and Clinical Psychology* **80**, 54-65 (2012).
152 Billari, F. C., Graziani, R. & Melilli, E. Stochastic population forecasting based on combinations of expert evaluations within the Bayesian paradigm. *Demography* **51**, 1933-1954 (2014).
153 Fallesen, P. & Breen, R. Temporary Life Changes and the Timing of Divorce. *Demography* **53**, 1377-1398, doi:10.1007/s13524-016-0498-2 (2016).
154 Hansford, T. G., Depaoli, S. & Canelo, K. S. Locating U.S. Solicitors General in the Supreme Court_s policy space. *Presidential Studies Quarterly* **49**, 855-869 (2019).
155 Phipps, D. J., Hagger, M. S. & Hamilton, K. Predicting limiting _free sugar_ consumption using an integrated model of health behavior. *Appetite* (2020).

156 Royle, J. & Dorazio, R. Hierarchical Modeling and Inference in Ecology.,(Academic Press: Amsterdam.). (2008).
157 Gimenez, O. *et al.* in *Modeling demographic processes in marked populations* 883-915 (Springer, 2009).
158 King, R., Morgan, B., Gimenez, O. & Brooks, S. P. *Bayesian analysis for population ecology*. (CRC press, 2009).
159 Kéry, M. & Schaub, M. *Bayesian population analysis using WinBUGS: a hierarchical perspective*. (Academic Press, 2011).
160 McCarthy, M. (New York, New York: Cambridge University Press, 2012).
161 Korner-Nievergelt, F. *et al. Bayesian data analysis in ecology using linear models with R, BUGS, and Stan*. (Academic Press, 2015).
162 Monnahan, C. C., Thorson, J. T. & Branch, T. A. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution* **8**, 339-348 (2017).
163 Ellison, A. M. Bayesian inference in ecology. *Ecology letters* **7**, 509-520 (2004).
164 Choy, S. L., O'Leary, R. & Mengersen, K. Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology* **90**, 265-277 (2009).
165 Kuhnert, P. M., Martin, T. G. & Griffiths, S. P. A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecology letters* **13**, 900-914 (2010).
166 King, R., Brooks, S. P., Mazzetta, C., Freeman, S. N. & Morgan, B. J. Identifying and diagnosing population declines: a Bayesian assessment of lapwings in the UK. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **57**, 609-632 (2008).
167 Newman, K. *et al. Modelling population dynamics*. (Springer, 2014).
168 Bachl, F. E., Lindgren, F., Borchers, D. L. & Illian, J. B. inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution* **10**, 760-766 (2019).
169 King, R. & Brooks, S. P. On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. *Biometrics* **64**, 816-824, doi:10.1111/j.1541-0420.2007.00938.x (2008).
170 Saunders, S. P., Cuthbert, F. J. & Zipkin, E. F. Evaluating population viability and efficacy of conservation management using integrated population models. *Journal of Applied Ecology* **55**, 1380-1392 (2018).
171 McClintock, B. T. *et al.* A general discrete-time modeling framework for animal movement using multistate random walks. *Ecological Monographs* **82**, 335-349 (2012).
172 Dennis, B., Ponciano, J. M., Lele, S. R., Taper, M. L. & Staples, D. F. Estimating density dependence, process noise, and observation error. *Ecological Monographs* **76**, 323-341 (2006).
173 Aeberhard, W. H., Mills Flemming, J. & Nielsen, A. Review of state-space models for fisheries science. *Annual Review of Statistics and Its Application* **5**, 215-235 (2018).
174 Isaac, N. J. B. *et al.* Data Integration for Large-Scale Models of Species Distributions. *Trends Ecol Evol* **35**, 56-67, doi:10.1016/j.tree.2019.08.006 (2020).
175 McClintock, B. T. *et al.* Uncovering ecological state dynamics with hidden Markov models. *arXiv preprint arXiv:2002.10497* (2020).
176 King, R. Statistical ecology. *Annual Review of Statistics and its Application* **1**, 401-426 (2014).
177 Fearnhead, P. MCMC for state-space models. (2011).
178 Andrieu, C., Doucet, A. & Holenstein, R. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 269-342 (2010).
179 Knape, J. & de Valpine, P. Fitting complex population models by combining particle filters with Markov chain Monte Carlo. *Ecology* **93**, 256-263, doi:10.1890/11-0797.1 (2012).

180 Finke, A., King, R., Beskos, A. & Dellaportas, P. Efficient sequential Monte Carlo algorithms for integrated population models. *Journal of Agricultural, Biological and Environmental Statistics* **24**, 204-224 (2019).
181 Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* **10**, 681-690, doi:10.1038/nrg2615 (2009).
182 Mimno, D., Blei, D. M. & Engelhardt, B. E. Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proc Natl Acad Sci U S A* **112**, E3441-3450, doi:10.1073/pnas.1412301112 (2015).
183 Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504, doi:10.1038/s41576-018-0016-z (2018).
184 Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511, doi:10.1038/nrg2796 (2010).
185 Allen, N. E., Sudlow, C., Peakman, T., Collins, R. & Biobank, U. K. UK biobank data: come and get it. *Sci Transl Med* **6**, 224ed224, doi:10.1126/scitranslmed.3008601 (2014).
186 Cortes, A. *et al.* Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat Genet* **49**, 1311-1318, doi:10.1038/ng.3926 (2017).
187 Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, e8124, doi:10.15252/msb.20178124 (2018).
188 Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257-272, doi:10.1038/s41576-019-0093-7 (2019).
189 Yau, C. & Campbell, K. Bayesian statistical learning for big data biology. *Biophys Rev* **11**, 95-102, doi:10.1007/s12551-019-00499-1 (2019).
190 Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Comput Biol* **11**, e1004333, doi:10.1371/journal.pcbi.1004333 (2015).
191 Wang, J. *et al.* Data denoising with transfer learning in single-cell transcriptomics. *Nat Methods* **16**, 875-878, doi:10.1038/s41592-019-0537-1 (2019).
192 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058, doi:10.1038/s41592-018-0229-2 (2018).
193 Institute, N. C. & National Cancer, I. The Cancer Genome Atlas. *Definitions*, doi:10.32388/e1plqh (2020).
194 Kuipers, J. *et al.* Mutational interactions define novel cancer subgroups. *Nat Commun* **9**, 4353, doi:10.1038/s41467-018-06867-x (2018).
195 Schwartz, R. & Schaffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nat Rev Genet* **18**, 213-229, doi:10.1038/nrg.2016.170 (2017).
196 Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour* **1**, doi:10.1038/s41562-016-0021 (2017).
197 van Erp, S., Mulder, J. & Oberski, D. L. Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychol Methods* **23**, 363-388, doi:10.1037/met0000162 (2018).
198 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018, doi:10.1038/sdata.2016.18 (2016).
199 Lamprecht, A.-L. *et al.* Towards FAIR principles for research software. *Data Science* **3**, 37-59, doi:10.3233/ds-190026 (2020).
200 *re3data.org - Registry of Research Data Repositories*.
201 Smith, A. M., Katz, D. S. & Niemeyer, K. E. Software citation principles. *PeerJ Computer Science* **2**, e86, doi:10.7717/peerj-cs.86 (2016).
202 Clyburne-Sherin, A., Fei, X. & Green, S. A. Computational Reproducibility via Containers in Psychology. *Meta-Psychology* **3** (2019).

203 Nosek, B. A. *et al.* SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* **348**, 1422-1425, doi:10.1126/science.aab2374 (2015).

204 Vehtari, A. & Ojanen, J. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* **6**, 142-228 (2012).

205 Abadi, M. *et al.* in *USENIX symposium on operating systems design and implementation (OSDI'16).*12 edn 265-283.

206 Paszke, A. *et al.* in *Advances in neural information processing systems.* 8026-8037.

207 Kingma, D. P. & Welling, M. *An Introduction to Variational Autoencoders*. (2019).

208 Higgins, I. *et al.* beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr* **2**, 6 (2017).

209 Märtens, K. & Yau, C. BasisVAE: Translation-invariant feature-level clustering with Variational Autoencoders. *arXiv [stat.ML]* (2020).

210 Liu, Q., Allamanis, M., Brockschmidt, M. & Gaunt, A. in *Advances in Neural Information Processing Systems 31* (eds S. Bengio *et al.*) 7795-7804 (Curran Associates, Inc., 2018).

211 Louizos, C., Shi, X., Schutte, K. & Welling, M. in *Advances in Neural Information Processing Systems* 8743-8754 (2019).

212 Garnelo, M. *et al.* in *Proceedings of the 35th International Conference on Machine Learning* Vol. 80 (eds Jennifer Dy & Andreas Krause) 1704-1713 (PMLR, 2018).

213 Kim, H. *et al.* Attentive Neural Processes. *arXiv [cs.LG]* (2019).

214 Rezende, D. & Mohamed, S. in *Proceedings of the 32nd International Conference on Machine Learning* Vol. 37 (eds Francis Bach & David Blei) 1530-1538 (PMLR, 2015).

215 Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. & Lakshminarayanan, B. Normalizing Flows for Probabilistic Modeling and Inference. *arXiv [stat.ML]* (2019).

216 Korshunova, I. *et al.* in *Advances in Neural Information Processing Systems 31* (eds S. Bengio *et al.*) 7190-7198 (Curran Associates, Inc., 2018).

217 Zhang, R., Li, C., Zhang, J., Chen, C. & Wilson, A. G. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902. 03932* (2019).

218 Neal, R. M. *Bayesian Learning for Neural Networks*. (Springer Science & Business Media, 2012).

219 Neal, R. M. in *Bayesian Learning for Neural Networks Lecture Notes in Statistics* (ed Radford M. Neal) Ch. Chapter 2, 29-53 (Springer New York, 1996).

220 Williams, C. K. I. in *Advances in neural information processing systems* 295-301 (1997).

221 MacKay David, J. C. A practical bayesian framework for backprop networks. *Neural Comput.* (1992).

222 Sun, S., Zhang, G., Shi, J. & Grosse, R. in *International Conference on Learning Representations* (2019).

223 Lakshminarayanan, B., Pritzel, A. & Blundell, C. in *Advances in neural information processing systems* 6402-6413 (2017).

224 Wilson, A. G. The Case for Bayesian Deep Learning. *arXiv [cs.LG]* (2020).

225 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929–1958 (2014).

226 Gal, Y. & Ghahramani, Z. 1050–1059.

227 Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732 (1995).

228 Hoffman, M. D. & Gelman, A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593-1623 (2014).

229 Liang, F. & Wong, W. H. Evolutionary Monte Carlo: applications to C p model sampling and change point problem. *Statistica sinica*, 317-342 (2000).

230 Liu, J. S. & Chen, R. Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association* **93**, 1032-1044 (1998).
231 Sisson, S., Fan, Y. & Beaumont, M. *Handbook of approximate Bayesian computation*. (Chapman and Hall/CRC 2018).
232 Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series B (Statistical Methodology)* **71**, 319-392 (2009).
233 WinBUGS user manual (Citeseer, 2003).
234 Ntzoufras, I. *Bayesian modeling using WinBUGS*. Vol. 698 (John Wiley & Sons, 2011).
235 OpenBUGS user manual, version 3.0. 2 (2007).
236 Plummer, M. in *Proceedings of the 3rd international workshop on distributed statistical computing.* 1-10.
237 Goudie, R. J., Turner, R. M., De Angelis, D. & Thomas, A. MultiBUGS: A parallel implementation of the BUGS modelling framework for faster Bayesian inference. *arXiv preprint arXiv:1704.03216* (2017).
238 Sturtz, S., Ligges, U. & Gelman, A. E. R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software* **12**, 1-16 (2005).
239 Thomas, A., O'Hara, B., Ligges, U. & Sturtz, S. Making BUGS Open. *R News* **6** (2006).
240 Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2**, e55, doi:10.7717/peerj-cs.55 (2016).
241 de Valpine, P. *et al.* Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* **26**, 403-413 (2017).
242 Dillon, J. V. *et al.* Tensorflow distributions. *arXiv preprint arXiv:1711.10604* (2017).
243 Keydana, S. *tfprobability: R interface to TensorFlow Probability*, <https://rstudio.github.io/tfprobability/index.html> (2020).
244 Bingham, E. *et al.* Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research* **20**, 973-978 (2019).
245 Bezanson, J., Karpinski, S., Shah, V. B. & Edelman, A. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145* (2012).
246 Ge, H., Xu, K., Scibior, A. & Ghahramani, Z. in *Artificial Intelligence and Statistics.*
247 Smith, B. J. *et al.* brian-j-smith/Mamba.jl: v0.12.4. *Zenodo*, doi:http://doi.org/10.5281/zenodo.3740216 (2020).
248 JASP Team. JASP (Version 0.13.1)[Computer software]. (2020).
249 Lindgren, F. & Rue, H. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software* **63**, 1-25 (2015).
250 Vanhatalo, J. *et al.* GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research* **14**, 1175-1179 (2013).
251 Betancourt, M. *Towards A Principled Bayesian Workflow*, <https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html> (April 2020).
252 Kramer, B. & Bosman, J. Summerschool Open Science and Scholarship 2019 - Utrecht University. doi:10.5281/ZENODO.3925004 (2020).
253 Rényi, A. On a new axiomatic theory of probability. *Acta Mathematica Hungarica* **6**, 285-335 (1955).
254 Lesaffre, E. & Lawson, A. B. *Bayesian biostatistics*. (John Wiley & Sons, 2012).