Bayesian statistics and modelling

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

Abstract | Bayesian statistics is an approach to data analysis based on Bayes' theorem, where available knowledge about parameters in a statistical model is updated with the information in observed data. The background knowledge is expressed as a prior distribution and combined with observational data in the form of a likelihood function to determine the posterior distribution. The posterior can also be used for making predictions about future events. This Primer describes the stages involved in Bayesian analysis, from specifying the prior and data models to deriving inference, model checking and refinement. We discuss the importance of prior and posterior predictive checking, selecting a proper technique for sampling from a posterior distribution, variational inference and variable selection. Examples of successful applications of Bayesian analysis across various research fields are provided, including in social sciences, ecology, genetics, medicine and more. We propose strategies for reproducibility and reporting standards, outlining an updated WAMBS (when to Worry and how to Avoid the Misuse of Bayesian Statistics) checklist. Finally, we outline the impact of Bayesian analysis on artificial intelligence, a major goal in the next decade.

Prior distribution
Beliefs held by researchers about the parameters in a statistical model before seeing the data, expressed as probability distributions.

Likelihood function
The conditional probability distribution of the given parameters of the data, defined up to a constant.

Posterior distribution
A way to summarize one's updated knowledge, balancing prior knowledge with observed data.

✉*e-mail:* a.g.j.vandeschoot@uu.nl

https://doi.org/10.1038/s43586-020-00001-2

Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' theorem. Unique for Bayesian statistics is that all observed and unobserved parameters in a statistical model are given a joint probability distribution, termed the prior and data distributions. The typical Bayesian workflow consists of three main steps (FIG. 1): capturing available knowledge about a given parameter in a statistical model via the prior distribution, which is typically determined before data collection; determining the likelihood function using the information about the parameters available in the observed data; and combining both the prior distribution and the likelihood function using Bayes' theorem in the form of the posterior distribution. The posterior distribution reflects one's updated knowledge, balancing prior knowledge with observed data, and is used to conduct inferences. Bayesian inferences are optimal when averaged over this joint probability distribution and inference for these quantities is based on their conditional distribution given the observed data.

The basis of Bayesian statistics was first described in a 1763 essay written by Reverend Thomas Bayes and published by Richard Price[1] on inverse probability, or how to determine the probability of a future event solely based on past events. It was not until 1825 that Pierre Simon Laplace[2] published the theorem we now know as Bayes' theorem (BOX 1). Although the ideas of inverse probability and Bayes' theorem are longstanding in mathematics, these tools became prominent in applied statistics in the past 50 years[3–10]. We describe many advantages and disadvantages throughout the Primer.

This Primer provides an overview of the current and future use of Bayesian statistics that is suitable for quantitative researchers working across a broad range of science-related areas that have at least some knowledge of regression modelling. We supply an overview of the literature that can be used for further study and illustrate how to implement a Bayesian model on real data. All of the data and code are available for teaching purposes. This Primer discusses the general framework of Bayesian statistics and introduces a Bayesian research cycle (FIG. 1). We first discuss formalizing of prior distributions, prior predictive checking and determining the likelihood distribution (Experimentation). We discuss relevant algorithms and model fitting, describe examples of variable selection and variational inference, and provide an example calculation with posterior predictive checking (Results). Then, we describe how Bayesian statistics are 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 on avoiding bias introduced by using incorrect models (Limitations and optimizations), and provide a look into the future with Bayesian artificial intelligence (Outlook).

Author addresses

[1]Department of Methods and Statistics, Utrecht University, Utrecht, Netherlands.
[2]Department of Quantitative Psychology, University of California Merced, Merced, CA, USA.
[3]School of Mathematics, University of Edinburgh, Edinburgh, UK.
[4]The Alan Turing Institute, British Library, London, UK.
[5]Utrecht University Library, Utrecht University, Utrecht, 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]Department of Statistics, Columbia University, New York, NY, USA.
[10]Division of Informatics, Imaging & Data Sciences, University of Manchester, Manchester, UK.

Experimentation

This section outlines the first two steps in the Bayesian workflow described in FIG. 1. Prior distributions, shortened to priors, are first determined. The selection of priors is often viewed as one of the more important choices that a researcher makes when implementing a Bayesian model as it can have a substantial impact on the final results. The appropriateness of the priors being implemented is ascertained using the prior predictive checking process. The likelihood function, shortened to likelihood, is then determined. The likelihood is combined with the prior to form the posterior distribution, or posterior (Results). Given the important roles that the prior and the likelihood have in determining the posterior, it is imperative that these steps be conducted with care. We provide example calculations throughout to demonstrate the process.

Empirical example 1: PhD delays

To illustrate many aspects of Bayesian statistics we provide an example based on real-life data. Consider an empirical example of a study predicting PhD delays[11] in which the researchers asked 333 PhD recipients in the Netherlands how long it had taken them to complete their doctoral thesis. Based on this information, the researchers computed the delay — defined as the difference between the planned and the actual project time in months (mean = 9.97, minimum/maximum = −31/91, standard deviation = 14.43). Suppose we are interested in predicting PhD delay (y) using a polynomial regression model, $y = \beta_{\text{intercept}} + \beta_{\text{age}} Age + \beta_{\text{age}^2} Age^2 + \varepsilon$, with β_{age} representing the linear effect of age (in years). We expect this relation to be quadratic, denoted by β_{age^2}. The model contains an intercept, $\beta_{\text{intercept}}$, and we assume the residuals, ε, are normally distributed with mean zero and with an unknown variance, σ_ε^2. Note that we have simplified the statistical model, and so the results are only meant for instructional purposes. Instructions for running the code are available for different software[12], including steps for data exploration[13]. We refer to this example throughout the following sections to illustrate key concepts.

Formalizing prior distributions

Prior distributions play a defining role in Bayesian statistics. Priors can come in many different distributional forms, such as a normal, uniform or Poisson distribution, among others. Priors can have different levels of informativeness; the information reflected in a prior distribution can be anywhere on a continuum from complete uncertainty to relative certainty. Although priors can fall anywhere along this continuum, there are three main classifications of priors that are used in the literature to categorize the degree of (un)certainty surrounding the population parameter value: informative, weakly informative and diffuse. These classifications can be made based on the researcher's personal judgement. For example, a normal distribution is defined by a mean and a variance, and the variance (or width) of the distribution is linked to the level of informativeness. A variance of 1,000 may be considered diffuse in one research setting and informative in another, depending on the likelihood function as well as the scaling for the parameter.

The relationship between the likelihood, prior and posterior for different prior settings for β_{age} from our example calculation predicting PhD delays is shown in FIG. 2. The first column represents the prior, which has a normal distribution for the sake of this example. The five different rows of priors represent the different prior settings based on the level of informativeness and variance from the mean. The likelihood, based on the data, is represented by a single distribution. The prior and the likelihood are combined together to create the posterior according to Bayes' rule. The resulting posterior is dependent on the informativeness (or variance) of the prior, as well as the observed data. We demonstrate how to obtain the posterior in the Results section.

The individual parameters that control the amount of uncertainty in the priors are called hyperparameters. Take a normal prior as an example. This distribution is defined by a mean and a variance that are the hyperparameters for the normal prior, and we can write this distribution as $N(\mu_0, \sigma_0^2)$, where μ_0 represents the mean and σ_0^2 represents the variance. A larger variance represents a greater amount of uncertainty surrounding the mean, and vice versa. For example, FIG. 2 illustrates five prior settings with different values for μ_0 and σ_0^2. The diffuse and weakly informative priors show more spread than the informative priors, owing to their larger variances. The mean hyperparameter can be seen as the peak in the distribution.

Prior elicitation. Prior elicitation is the process by which a suitable prior distribution is constructed. Strategies for prior elicitation include asking an expert or a panel of experts to provide values for the hyperparameters of the prior distribution[14–17]. MATCH[18] is a generic expert elicitation tool, but many methods that can be used to elicit information from experts require custom elicitation procedures and tools. For examples of elicitation procedures designed for specific models, see REFS[19–23]. For an abundance of elicitation examples and methods, we refer the reader to the TU Delft expert judgement database of more than 67,000 elicited judgements[24] (see also[14,25,26]). Also, the results of a previous publication or meta-analysis can be used[27,28], or any combination[29] or variation of such strategies.

Prior elicitation can also involve implementing data-based priors. Then, the hyperparameters for the prior are derived from the sample data using methods such as maximum likelihood[30–33] or sample statistics[34–36].

Informativeness
Priors can have different levels of informativeness and can be anywhere on a continuum from complete uncertainty to relative certainty, but we distinguish between diffuse, weakly and informative priors.

Hyperparameters
Parameters that define the prior distribution, such as mean and variance for a normal prior.

Prior elicitation
The process by which background information is translated into a suitable prior distribution.

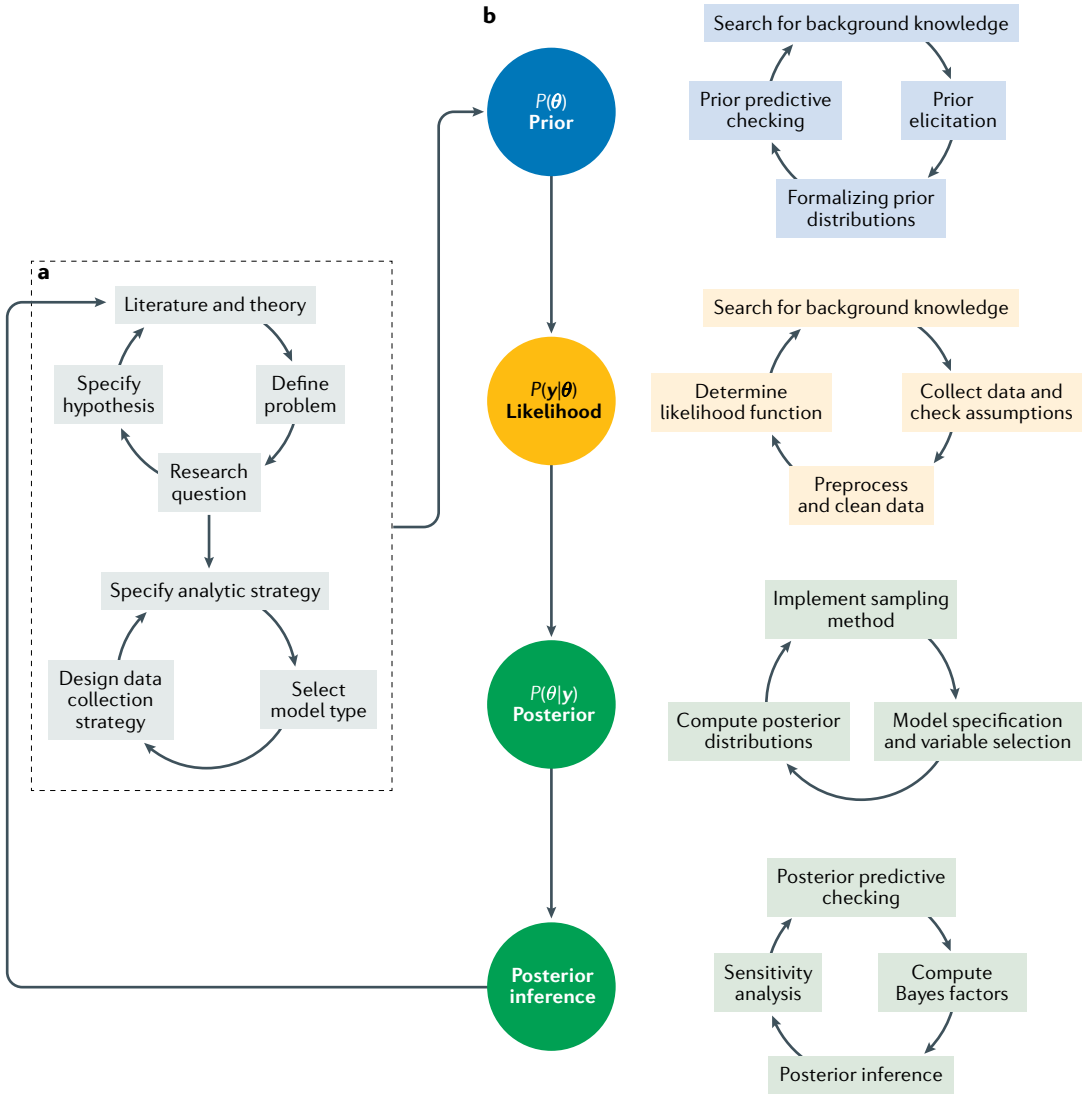

Fig. 1 | **The Bayesian research cycle.** The steps needed for a research cycle using Bayesian statistics include those of a standard research cycle and a Bayesian-specific workflow. **a** | Standard research cycle involves reading literature, defining a problem and specifying the research question and hypothesis[248,249]. The analytic strategy can be pre-registered to enhance transparency. **b** | Bayesian-specific workflow 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 the likelihood function[134,250]. After obtaining the posterior results, inferences can be made that can then be used to start a new research cycle. $\boldsymbol{\theta}$, unknown parameter; $P(.)$, probability distribution; $\mathbf{y}$, data.

These procedures lead to double-dipping, as the same sample data set is used to derive prior distributions and to obtain the posterior. Although databased priors are relatively common, we do not recommend the use of double-dipping procedures. Instead, a hierarchical modelling strategy can be implemented, where priors depend on hyperparameter values that are data-driven — for example, sample statistics pulled from the sample data — which avoids the direct problems linked to double-dipping. We refer the reader elsewhere[34] for more details on double-dipping.

Prior (un)certainty. An informative prior is one that reflects a high degree of certainty about the model parameters being estimated. For example, an informative normal prior would be expected to have a very small variance. A researcher may want to use an informative prior when existing information suggests restrictions on the possible range of a particular parameter, or a relationship between parameters, such as a positive but imperfect relationship between susceptibility to various medical conditions[37,38]. 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. An arbitrary example of an informative prior for our empirical example is $\beta_{age} \sim N(2.5, 5)$, with a prior mean for the linear relation of age with PhD delay of 2.5

Informative prior
A reflection of a high degree of certainty or knowledge surrounding the population parameters. Hyperparameters are specified to express particular information reflecting a greater degree of certainty about the model parameters being estimated.

Box 1 | **Bayes' theorem**

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:

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

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 $p(B \cap A) = p(A \cap B)$, where the notation $\cap$ represents an intersection. Similarly, we can write:

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

which, based on Eq. 1, can be reworked as:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \quad (3)$$

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

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

which is often simplified to:

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

The term $p(\boldsymbol{\theta}|\mathbf{y})$ represents a conditional probability, where the probability of the model parameters ($\boldsymbol{\theta}$) is computed conditional on the data ($\mathbf{y}$), representing the posterior distribution. The term $p(\mathbf{y}|\boldsymbol{\theta})$ represents the conditional probability of the data given the model parameters, and this term represents the likelihood function. Finally, the term $p(\boldsymbol{\theta})$ represents the probability of particular model parameter values existing in the population, also known as the prior distribution. The term $p(\mathbf{y})$ is a normalizing factor and can be dropped from the equation as it does not depend on $\boldsymbol{\theta}$. Thus, the posterior distribution is proportional to the likelihood function multiplied by the prior distribution.

and a prior variance of 5. A ShinyApp was developed specifically for the PhD example containing a visualization of how the different priors for all parameters in the regression model interact[39].

A weakly informative prior has a middling amount of certainty, being neither too diffuse nor too restrictive. A weakly informative normal prior would have a larger variance hyperparameter than an informative prior. Such priors will have a relatively smaller impact on the posterior compared with an informative prior, depending on the scale of the variables, and the posterior results are weighted more by the data observations as expressed in the likelihood.

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. In FIG. 2, the two examples of weakly informative normal priors 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. Weakly informative priors supply more information than diffuse priors, but they typically do not represent specific information like an informative prior[40,41]. When constructing a weakly informative prior, it is typical to specify a plausible parameter space, which captures a range of plausible parameter values — those within a reasonable range of values for the select parameter (for an example, see the ShinyApp we developed for the PhD example[39]) — and make improbable values unlikely by placing a limited density mass over them. For example, if a regression coefficient is known to be near 0, then a weakly informative prior can be specified to reduce the plausible range to, for example, ±5. This prior would reduce the probability of observing out-of-bound values (for example, a regression coefficient of 100) without being too informative.

Finally, a diffuse prior reflects a great deal of uncertainty about the model parameter. This prior form represents a relatively flat density and does not include specific knowledge of the parameter (FIG. 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 for diffuse[42]. We refrain from using this term because we argue that even a completely flat prior, such as the Jeffreys prior[43], still provides information about the degree of uncertainty[44]. Therefore, no prior is truly 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[45]. For example, diffuse priors can have an adverse impact on parameter estimates via the posterior when sample sizes are small, especially in complex modelling situations involving meta-analytic models[46], logistic regression models[44] or mixture models[47]. In addition, improper priors are sometimes used with the intention of using them as diffuse priors. Although improper priors are common and can be implemented with relative ease within various 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 before analyses of the same or subsequent data are conducted with more informative priors.

Impact of 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 or biased results can be obtained with relatively flat priors[47]. Likewise, an informative prior that does not overlap well with the likelihood can shift the posterior away from the likelihood, indicating that inferences will be aligned more with the prior than the likelihood. Regardless of the informativeness of the prior, it is always important to conduct a prior sensitivity analysis to fully understand the influence that the prior settings have on posterior estimates[48,49]. When the sample size is small, Bayesian estimation with mildly informative priors is often used[9,50,51], but

Weakly informative prior
A prior incorporating some information about the population parameter but that is less certain than an informative prior.

Diffuse priors
Reflections of complete uncertainty about population parameters.

Improper priors
Prior distributions that integrate to infinity.

the prior specification might have a huge effect on the posterior results.

When priors do not conform with the likelihood, this is not necessarily evidence that the prior is not appropriate. It may be that the likelihood is at fault owing to a mis-specified model or biased data. The difference between the prior and the likelihood may also 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, by examining different forms of the model, for example, to assess how the priors and the likelihood align.

The subjectivity of priors is highlighted by critics as a potential drawback of Bayesian methods. We argue two distinct points here. First, many elements of the estimation process are subjective, aside from prior selection, including the model itself and the error assumptions. To place the notion of subjectivity solely on the priors is a misleading distraction from the other elements in the process that are inherently subjective. 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 towards a likely high-density region and improve estimation efficiency.

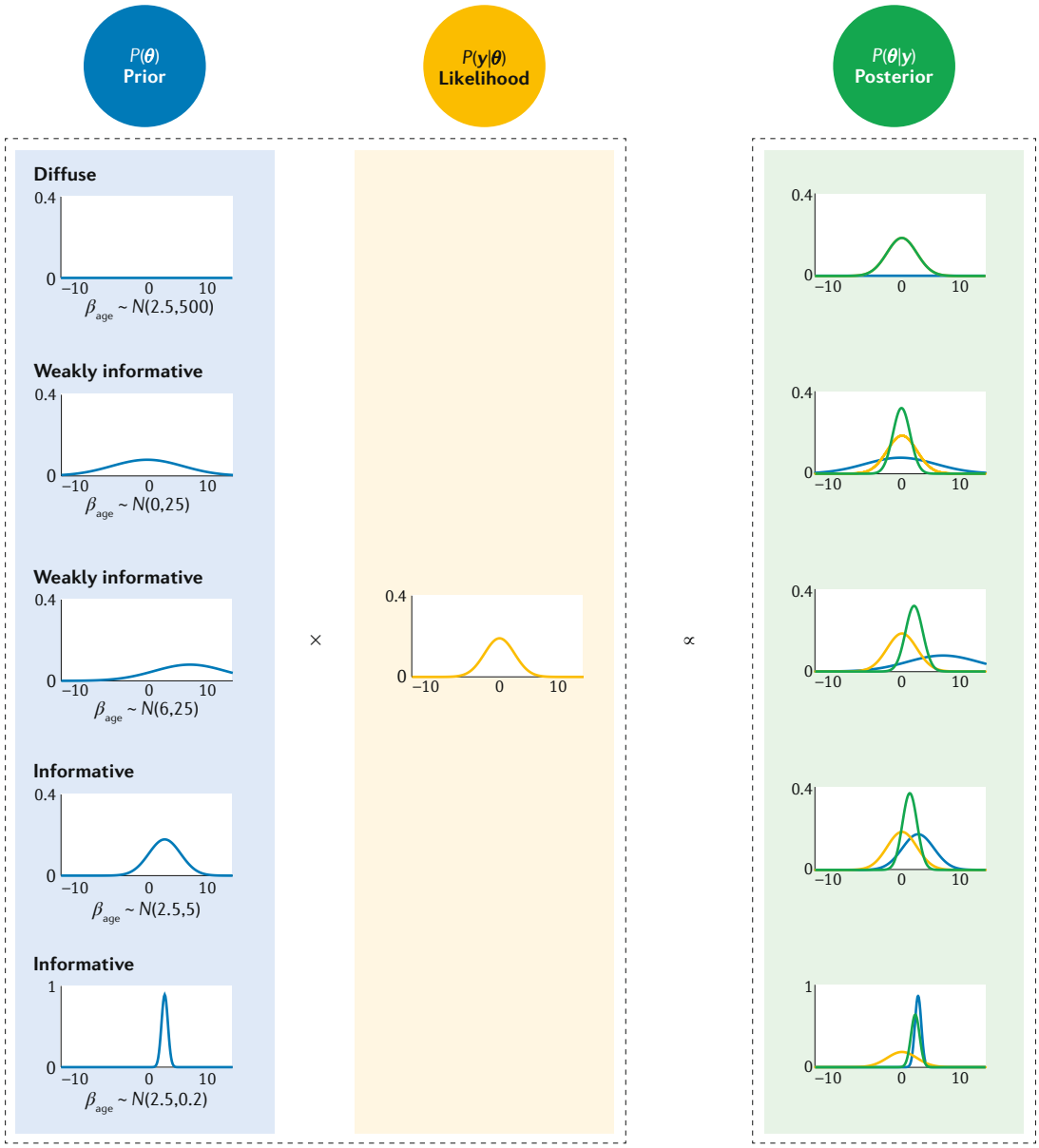

Fig. 2 | **Illustration of the key components of Bayes' theorem.** The prior distribution (blue) and the likelihood function (yellow) are combined in Bayes' theorem to obtain the posterior distribution (green) in our calculation of PhD delays. Five example priors are provided: one diffuse, two weakly informative with different means but the same variance and two informative with the same mean but different variances. The likelihood remains constant as it is determined by the observed data. The posterior distribution is a compromise between the prior and the likelihood. In this example, the posterior distribution is most strongly affected by the type of prior: diffuse, weakly informative or informative. β_{age}, linear effect of age (years); $\boldsymbol{\theta}$, unknown parameter; $P(.)$, probability distribution; $\boldsymbol{y}$, data.

Box 2 | **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 is the ratio of the posterior odds to the prior odds of distinct hypotheses[43,64]. For two hypotheses, H_0 and H_1, and observed data $\mathbf{y}$, the Bayes factor in favour of H_1, denoted BF_{10}, is given by:

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

where the prior probabilities are $p(H_0)$ and $p(H_1) = 1 - p(H_0)$. A larger value of BF_{10} provides stronger evidence against H_0 (REF.[64]). The posterior probability of hypothesis H_j, $p(H_j|\mathbf{y})$, for $j = 0$ or 1, is obtained using Bayes theorem:

$$p(H_j|\mathbf{y}) = \frac{p(\mathbf{y}|H_j)p(H_j)}{p(\mathbf{y})}. \quad (7)$$

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

$$BF_{10} = \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)}. \quad (8)$$

The competing hypotheses can take various forms and could be, for example, two non-nested regression models. If H_0 and H_1 are simple hypotheses in which the parameters are fixed (for example, 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 or there are additional unknown parameters, the marginal likelihood $p(\mathbf{y}|H_j)$ is obtained by integrating over the parameters $\boldsymbol{\theta}_j$ with prior densities $p(\boldsymbol{\theta}_j|H_j)$. This integral is often intractable and must be computed by numerical methods. If $p(\boldsymbol{\theta}_j|H_j)$ is improper (that is, $\int p(\boldsymbol{\theta}_j|H_j)d\boldsymbol{\theta}_j = \infty$), then $p(\mathbf{y}|H_j)$ 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 favours 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 $p(\bar{y}|H_1) = \int p(\mathbf{y}|\mu, H_1)p(\mu|H_1)\mathrm{d}\mu$ reduces to $\bar{y}|H_1 \sim N(\mu_0, \tau^2 + \sigma^2/n)$. Thus, the Bayes factor in favour of H_1 is:

$$BF_{10} = \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)} = \frac{(\tau^2 + \sigma^2/n)^{-1/2}\exp\left\{-\frac{(\bar{y}-\mu_0)^2}{2(\tau^2+\sigma^2/n)}\right\}}{(\sigma^2/n)^{-1/2}\exp\left\{-\frac{(\bar{y}-\mu_0)^2}{2(\sigma^2/n)}\right\}} \quad (9)$$

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.

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 quantifiable, 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.

This section on informative, weakly informative and diffuse priors was written in a general sense, with these terms applicable to both univariate and multivariate priors. The majority of the discussion in this Primer pertains to univariate priors placed on individual model parameters; however, these concepts can be extended to the multivariate setting, where priors are placed on an entire covariance matrix rather than a single element from a matrix, for example. For more information on multivariate priors, see REFS[52,53].

Prior predictive checking

Because 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[54,55]. This is partly done by means of a process known as prior predictive checking. Priors are based on background knowledge and cannot be inherently wrong if the prior elicitation procedure is valid, that is, if the background knowledge is correctly expressed in probability statements. However, even in the case of a valid prior elicitation procedure, it is extremely important to understand the exact probabilistic specification of the priors. This is especially true for complex models with smaller sample sizes[9]. Because smaller sample sizes usually convey less information, priors, in comparison, will exhibit a strong influence on the posteriors. Prior predictive checking is an exercise to improve the understanding of the implications of the specified priors on possible observations. It is not a method for changing the original prior, unless this prior explicitly generates incorrect data.

Box[56] suggested deriving a prior predictive distribution 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[54]. Prior predictive checking compares the observed data, or statistics of the observed data, with the prior predictive distribution, or statistics of the predictive distribution, and checks their compatibility[55]. For instance, values are drawn from the prior distributions. Using kernel density estimation, a non-parametric smoothing approach used to approximate a probability density function[57], the original sample and the samples from the predictive distribution can be compared[58]. Alternatively, the compatibility can be summarized by a prior predictive *p*-value, describing how far the characteristics of the observed data lie in the tails of the reference prior predictive distribution[59]. Evans and Moshonov[60,61] suggested restricting Box's approach to minimal sufficient statistics, that is, statistics that are as efficient as possible in relaying information about the value of a certain parameter from a sample[62].

Young and Pettit[63] argued that measures based on the tail area of the prior predictive distribution, such as the approaches of Box and Evans and Moshonov, do not favour the more precise prior in cases where two priors are both specified at the correct value. Instead, they propose using a Bayes factor[64] to compare two priors (BOX 2). The Bayes factor would favour the more precise prior. These three approaches leave the determination of prior–data conflict subjective, depending on an arbitrary cut-off value. The data agreement criterion[65] tries to resolve the prior–data conflict determination issue by introducing a clear classification, removing the subjective element of this decision[66]. This is done at the expense of selecting an arbitrary divergence-based criterion.

Prior predictive checking
The process of checking whether the priors make sense by generating data according to the prior in order to assess whether the results are within the plausible parameter space.

Prior predictive distribution
All possible samples that could occur if the model is true based on the priors.

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

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

An alternative criterion has been developed[67] that computes whether the distance between the prior and the data is unexpected. For a comparison of both criteria, we direct the reader to Lek and van de Schoot[68].

Empirical example 1 continued. Prior predictive checking can help prevent mistakes in the formalization of the priors. For instance, various software packages can notate the same distribution differently. The normal distribution of the prior can be specified by the hyperparameters mean and variance, mean and standard deviation or mean and precision (the inverse of the variance). To illustrate the impact of a data input error, for the informative prior with a mean of 2.5 and variance of 5 shown in FIG. 2, we also show a prior for which we have intentionally mis-specified the variance as the precision value (0.2), shown as $\beta_{age} \sim N(2.5, 0.2)$. If a user is not aware of differences between variance and precision, a prior that was intended to be weakly informative can easily turn into an informative prior distribution. Note that in this example the prior predictive distribution and the data are compared on the mean and standard deviation of the data, as these are commonly used to check prior predictive performance. The comparison statistics can, however, be chosen to reflect important characteristics of the data, such as skewness.

The prior predictive checks shown in FIG. 3 help to avoid mis-specification, for instance, when comparing the prior predictive distribution when precision is mistakenly used instead of variance (FIG. 3a) with the distribution based on the correct hyperparameters (FIG. 3b). We also show the kernel density estimate[57], or the estimate of the probability density function, of the observed data versus simulated data (FIG. 3c). Because of the combinations of uncertainty in the priors, the prior predictive kernel density estimates can be quite different from the observed data, and so it is also important to check that the prior predictive kernel distributions are not orders of magnitude different from the observed data.

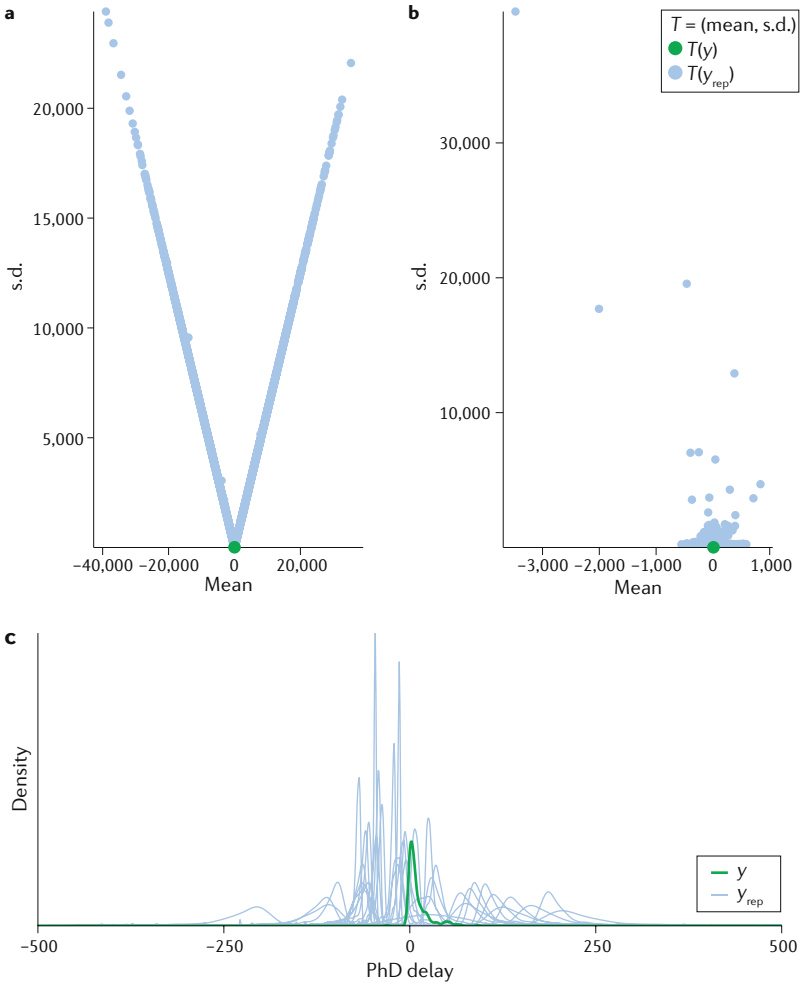

Fig. 3 | **Prior predictive checking for the PhD delay example. a** | In this scenario, precision is mistakenly used instead of variance (0.2 versus 5) for the linear effect of age (years), β_{age}, and the prior predictive distribution displays an unexpected pattern. The test statistic, *T*, of interest is the combination of the mean and standard deviation (s.d.): *T*(mean, s.d.). The observed combination of the mean and s.d., *T*(*y*) (green), is shown along with the same combination but now obtained from samples of the prior predictive distribution, $T(y_{rep})$ (blue). **b** | In this scenario, the prior predictive distribution is shown using the correct implementation of variance. The prior predictive checks for the correct implementation of the priors seem reasonable given the data. **c** | A kernel density estimate of the observed data is displayed (*y*; green), and kernel density estimates for the samples of the prior predictive distribution (y_{rep}; blue)[57,58]. The priors cover the entire plausible parameter space with the observed data in the centre. Computed via Stan[98] — the scripts are available at the Open Science Framework[141].

Determining the likelihood function

The likelihood is used in both Bayesian and frequentist inference[69]. In both inference paradigms, its role is to quantify the strength of support the observed data lends to possible value(s) for the unknown parameter(s). The key difference between Bayesian and frequentist inference is that frequentists do not consider probability statements about the unknown parameters to be useful. Instead, the unknown parameters are considered to be fixed; the likelihood is the conditional probability distribution $p(\boldsymbol{y}|\boldsymbol{\theta})$ of the data ($\boldsymbol{y}$), given fixed parameters ($\boldsymbol{\theta}$). In Bayesian inference, unknown parameters are referred to as random variables in order to make probability statements about them. The (observed) data are treated as fixed, whereas the parameter values are varied; the likelihood is a function of $\boldsymbol{\theta}$ for the fixed data $\boldsymbol{y}$. Therefore, the likelihood function summarizes the following elements: a statistical model that stochastically generates all of the data, a range of possible values for $\boldsymbol{\theta}$ and the observed data $\boldsymbol{y}$.

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, we direct the interested reader to a recent tutorial[70] describing likelihood in common frequentist and Bayesian statistical methods. For a complete mathematical explanation on this topic, see REF.[71]. Much of the discussion surrounding Bayesian inference focuses on the choice of priors, and there is a vast literature on potential default priors[72,73]. The inclusion of available knowledge into a prior is the most noticeable difference between frequentist and Bayesian methods, and a source of controversy. The importance of the likelihood is often omitted from the discussion, even though the specified model for the data — represented by the likelihood function — is the foundation for the analysis[74]. The posterior distribution is the result of the prior distribution in interaction

Bayes factor
The ratio of the posterior odds to the prior odds of two competing hypotheses, also calculated as the ratio of the marginal likelihoods under the two hypotheses. It can be used, for example, to compare candidate models, where each model would correspond to a hypothesis.

Credible interval
An interval that contains a parameter with a specified probability. The bounds of the interval are the upper and lower percentiles of the parameter's posterior distribution. For example, a 95% credible interval has the upper and lower 2.5% percentiles of the posterior distribution as its bounds.

Closed form
A mathematical expression that can be written using a finite number of standard operations.

Marginal posterior distribution
Probability distribution of a parameter or subset of parameters within the posterior distribution, irrespective of the values of other model parameters. It is obtained by integrating out the other model parameters from the joint posterior distribution.

with the assumed probability model for the data in the context of the observed data[72]. Without the context of the likelihood in which it will be paired, the prior is often impossible to interpret.

In some cases, specifying a likelihood function can be very straightforward (BOX 3). However, in practice, the underlying data-generating model is not always known. Researchers often naively choose a certain data-generating model out of habit or because they cannot easily change it in the software. Although based on background knowledge, the choice of the statistical data-generating model is subjective and should therefore be well understood, clearly documented and available to the reader. Robustness checks should be performed on the selected likelihood function to verify its influence on the posterior estimates[73]. Although most research on Bayesian robustness focuses on the sensitivity of the posterior results to the specification of the prior, a few contributions have focused on the sensitivity of the posterior results to the specification of the likelihood function[75–77].

Results

After specifying the prior and the likelihood, and collecting the data, the posterior distribution can be obtained. Here, we explain how a model can be fitted to data to obtain a posterior distribution, how to select variables and why posterior predictive checking is needed. Model building is an iterative process; any Bayesian model can be viewed as a placeholder that can be improved in response to new data or lack of fit to existing data, or simply through a process of model refinement. Box[56], Rubin[78] and Gelman et al.[74] discuss the fluidity of Bayesian model building, inference, diagnostics and model improvement.

Model fitting

Once the statistical model has been defined and the associated likelihood function derived, the next step is to fit the model to the observed data to estimate the unknown parameters of the model. Although statistical models are a simplification of reality, they aim to capture the primary factors of the underlying system that we wish to improve our understanding of and which lead to the data that we observe. Models may differ substantially in their complexity, taking into account the many different possible factors or mechanisms that act on the underlying system, and sources of stochasticity and variability resulting in the given data that we observe. 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.

The frequentist framework for model fitting focuses on the expected long-term outcomes of an experiment with the intent of producing a single point estimate for model parameters such as the maximum likelihood estimate and associated confidence interval. Within the Bayesian framework for model fitting, probabilities are assigned to the model parameters, describing the associated uncertainties. In Bayesian statistics, the focus is on estimating the entire posterior distribution of the model parameters. This posterior distribution is often summarized with associated point estimates, such as the posterior mean or median, and a credible interval. Direct inference on the posterior distribution is typically not possible, as the mathematical equation describing the posterior distribution is usually both very complicated and high-dimensional, with the number of dimensions equal to the number of parameters. The expression for the posterior distribution is typically only known up to a constant of proportionality, a constant term in the posterior distribution that is not a function of the parameters and, in general, cannot be explicitly calculated. In particular, the denominator of the expression for the posterior distribution is a function of only the data, where this function is not available in closed form but expressible only as an analytically intractable integral. This means that we cannot evaluate the posterior distribution exactly, and so cannot calculate, for example, associated summary statistics of interest directly. Further, the high dimensionality exacerbates these problems, so that calculating the marginal posterior distribution may also not be tractable, and expressible only in integral form. We note that this intractability of the posterior distribution was the primary practical reason why Bayesian statistics was discarded by many scientists in favour of frequentist statistics. The seminal article by Gelfand and Smith[79] described how Markov chain Monte Carlo (MCMC), a technique for sampling from a probability distribution, can be used to fit models to data within the Bayesian paradigm[80]. In particular, the MCMC algorithm only requires the probability distribution of interest to be specified up to a constant of proportionality and is scalable to high dimensions.

Markov chain Monte Carlo. MCMC is able to indirectly obtain inference on the posterior distribution using computer simulations[80]. MCMC permits a set of sampled parameter values of arbitrary size to be obtained from the posterior distribution, despite the posterior distribution being high-dimensional and only known up to a constant of proportionality. These sampled parameter values are used to obtain empirical estimates of the

Box 3 | **The likelihood function for a coin experiment**

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

$$p(y|\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}, \quad 0 \le \theta \le 1 \tag{10}$$

When y is kept fixed and θ is varied, $p(y|\theta)$ becomes a continuous function of θ, called the binomial likelihood function[254].

Suppose we flipped the coin 10 times and observed 4 heads, then the likelihood function of θ is defined by:

$$p(y|\theta) = \binom{10}{4}\theta^{4}(1-\theta)^{6}, \quad 0 \le \theta \le 1. \tag{11}$$

Table 1 | **An overview of MCMC-based and non-MCMC-based sampling techniques**

Approach	Short description
MCMC-based methods	
Metropolis–Hastings	An algorithm used for obtaining random samples from a probability distribution. Uses a general proposal distribution, with an associated accept/reject step for the proposed parameter value(s)[85,86]
Reversible jump MCMC	An extension of the Metropolis–Hastings algorithm. Permits simulation of trans-dimensional moves within parameter space[35,225]
Hamiltonian Monte Carlo	A Metropolis–Hastings algorithm based on Hamiltonian dynamics[87]. This algorithm is useful if direct sampling is difficult, if the sample size is small or when autocorrelation is high. The algorithm avoids the random walk of Metropolis–Hastings and sensitivity by taking a series of steps informed by first-order gradient information. The No-U-Turn Sampler[226] is an extension and is often faster because it often avoids the need for tuning the model
Gibbs sampler	A Metropolis–Hastings algorithm where the proposal distribution is the corresponding posterior conditional distribution, with an associated acceptance probability of 1 (REF.[84])
Particle MCMC	A combined sequential Monte Carlo algorithm and MCMC used when the likelihood is analytically intractable[177]
Evolutionary Monte Carlo	An MCMC algorithm that incorporates features of genetic algorithms and simulated annealing[227]. It allows the Markov chain to effectively and efficiently explore the parameter space and avoid getting trapped at local modes of the posterior distribution. It is particularly useful when the target distribution function is high-dimensional or multimodal
Non-MCMC-based methods	
Sequential Monte Carlo	An algorithm based on multiple importance sampling steps for each observed data point. Often used for online or real-time processing of data arrivals[228]
Approximate Bayesian computation	An approximate approach, typically used when the likelihood function is analytically intractable or very computationally expensive[229]
Integrated nested Laplace approximations	An approximate approach developed for the large class of latent Gaussian models, which includes generalized additive spline models, Gaussian Markov processes and random fields[230]
Variational Bayes	Variational inference describes a technique to approximate posterior distributions via simpler approximating distributions. The popular mean-field approximation assigns an approximating variational distribution to each parameter independently. Gradient descent is then used to optimize the variational parameters to minimize a loss function known as the evidence lower bound[99]

MCMC, Markov chain Monte Carlo.

posterior distribution of interest. This posterior distribution, and associated summary statistics of interest, can be estimated up to the desired accuracy by increasing the number of sampled parameter values, if necessary. We note that owing to the high dimensionality of the posterior distribution, it is often useful to focus on the marginal posterior distribution of each parameter, defined by integrating out over the other parameters. Marginal distributions are useful for focusing on individual parameters but, by definition, do not provide any information on the relationship between the parameters.

Here, we focus on MCMC for posterior inference. MCMC combines two concepts: obtaining a set of parameter values from the posterior distribution using the Markov chain; and obtaining a distributional estimate of the posterior and associated statistics with sampled parameters using Monte Carlo integration. Although MCMC is the most common class of algorithm used in Bayesian analyses, there are other model-fitting algorithms (TABLE 1). Other available estimators can be found elsewhere[81,82].

In general, Monte Carlo integration is a technique for estimating integrals using computer simulations of sampled values from a given distribution. Given these sampled parameter values, Monte Carlo integration permits estimation of this distribution using associated empirical estimates[83]. For example, for distributional summary statistics, such as the mean, variance or symmetric 95% credible interval of a parameter, we estimate these summary statistics using the corresponding sample mean, sample variance, and 2.5% and 97.5% quantile parameter values, respectively. Similarly, probability statements — such as the probability that a parameter is positive or negative, or that it lies in a range $[a,b]$ — can be estimated as the proportion of the sampled values that satisfy the given statement. The marginal posterior distribution of any given parameter can be obtained by kernel density estimation, which uses a non-parametric approach for estimating the associated density from which sampled values have been drawn[58].

It is not possible to directly and independently sample parameter values from the posterior distribution. This leads to the use of the Markov chain. The idea is to obtain a set of sampled parameter values from the posterior distribution of interest by constructing a Markov chain with a specified first-order transition kernel, such that the resulting stationary distribution of the Markov chain is equal to this posterior distribution of interest.

Markov chain Monte Carlo (MCMC). A method to indirectly obtain inference on the posterior distribution by simulation. The Markov chain is constructed such that its corresponding stationary distribution is the posterior distribution of interest. Once the chain has reached the stationary distribution, realizations can be regarded as a dependent set of sampled parameter values from the posterior distribution. These sampled parameter values can then be used to obtain empirical estimates of the posterior distribution, and associated summary statistics of interest, using Monte Carlo integration.

Markov chain
An iterative process whereby the values of the Markov chain at time $t+1$ are only dependent on the values of the chain at time t.

Monte Carlo
A stochastic algorithm for approximating integrals using the simulation of random numbers from a given distribution. In particular, for sampled values from a distribution, the associated empirical value of a given statistic is an estimate of the corresponding summary statistic of the distribution.

Transition kernel
The updating procedure of the parameter values within a Markov chain.

Auxiliary variables
Additional variables entered in a model such that the joint distribution is available in closed form and quick to evaluate.

If the Markov chain is run sufficiently long to reach its stationary distribution, subsequent realizations of the chain can be regarded as a dependent sample from the posterior distribution, and can be used to obtain the corresponding Monte Carlo estimates (FIG. 4a). We emphasize that the sampled parameter values obtained from the Markov chain are autocorrelated — they are dependent on their previous values in the chain — and are generated by the first-order Markov chain. The Markov chain is defined by the specification of the initial parameter values and transition kernel. The Gibbs sampler[84], the Metropolis–Hastings algorithm[85,86] and Hamiltonian Monte Carlo[87] are standard approaches for defining the transition kernel so that the corresponding stationary distribution is the correct posterior distribution.

MCMC technical aspects. Obtaining posterior inference by fitting models to observed data can be complicated owing 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 of the random effect terms or latent variables. Alternatively, the likelihood may be available in closed form, but may be multimodal — for example, for a finite mixture model or a discrete latent variable model. This, in turn, can lead to poor performance of the algorithm with one (or more) mode(s) not explored by the algorithm. In such circumstances, data augmentation is often used[88], where we define additional variables, or auxiliary variables, such that the

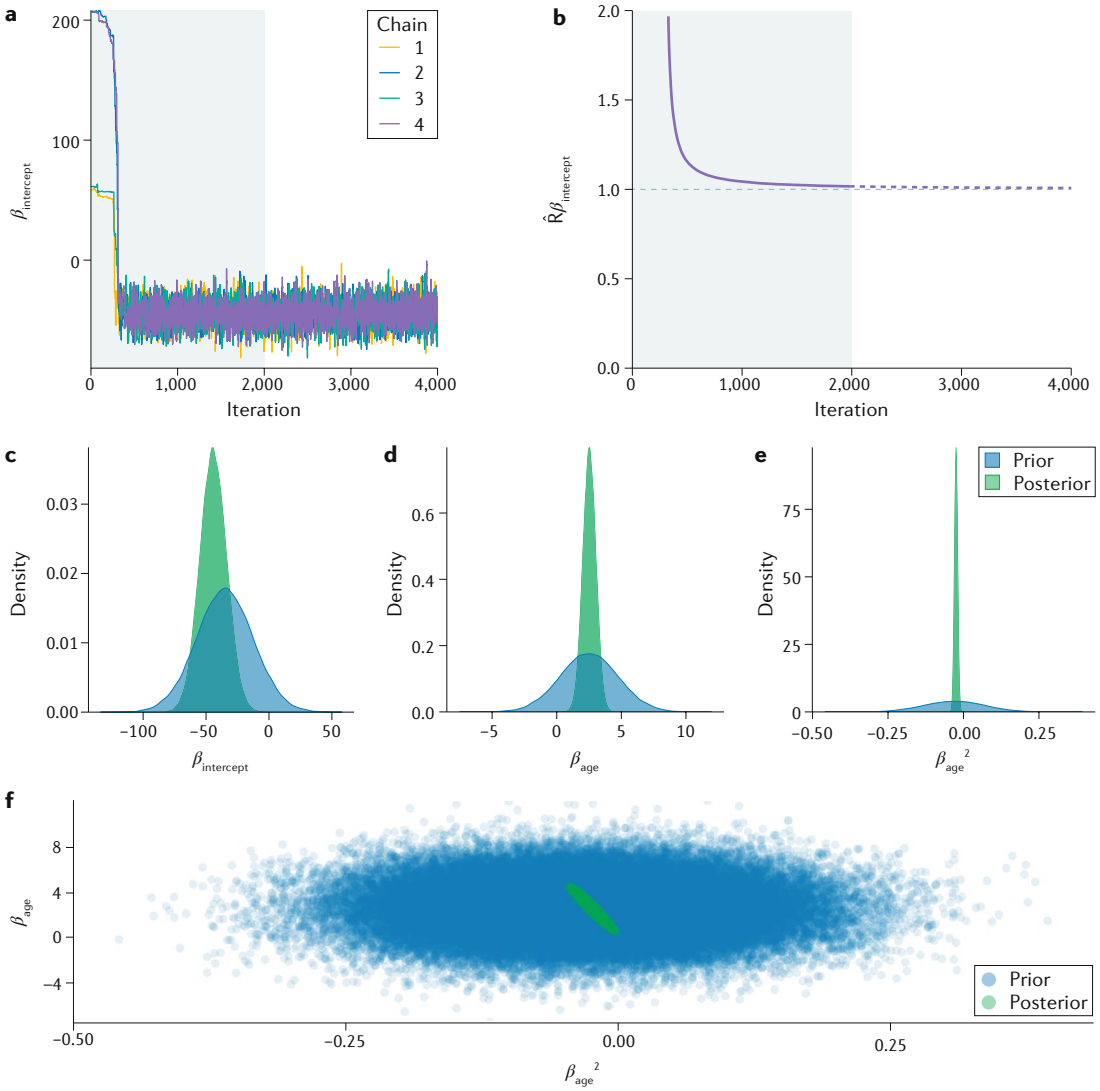

Fig. 4 | **Posterior estimation using MCMC for the PhD-delays example. a** | Trace plots showing the iteration number against the parameter value for the PhD delay data of four independent chains of the Markov chain Monte Carlo (MCMC) algorithms for exploring the posterior distribution of an intercept, $\beta_{intercept}$. The shaded section represents the warm-up phase and was omitted for constructing the posterior distribution. **b** | The associated $\hat{R}$ statistic for $\beta_{intercept}$, which appears to converge on 1 after approximately 2,000 iterations (shaded). **c**–**e** | The prior and posterior distributions for $\beta_{intercept}$ (part **c**), the linear effect of age (years), β_{age} (part **d**) and β_{age^2} (part **e**) are shown. For each chain, the first 2,000 iterations are discarded as warm-up. **f** | To illustrate the interrelation between two parameters, the prior (blue) and posterior (green) marginal densities for β_{age} and β_{age^2} are shown. All results are computed in Stan[98] — the scripts are available at the Open Science Framework[141].

Trace plots
Plots describing the posterior parameter value at each iteration of the Markov chain (on the *y* axis) against the iteration number (on the *x* axis).

$\hat{R}$ statistic
The ratio of within-chain and between-chain variability. Values close to one for all parameters and quantities of interest suggest the Markov chain Monte Carlo algorithm has sufficiently converged to the stationary distribution.

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, in a random effects model, 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 to which each observation belongs. A new joint posterior distribution is then constructed over both the model parameters and the auxiliary variables; this posterior distribution is defined to be proportional to the complete data likelihood and associated prior distributions specified on the parameters. A standard MCMC algorithm can then be applied that obtains a set of sampled parameter values over both the model parameters and auxiliary variables. Discarding the auxiliary variables and considering only the values of the model parameters of interest within the Markov chain provides a sample from the original posterior distribution of the model parameters conditional on the observed data. Auxiliary variables may themselves be of interest in some cases, for example where they represent missing data values or some tangible concept such as a homogeneous subgroup (for a mixture model) or true underlying state (as for a state space model), and inference on these can be easily obtained using 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 straightforward rules. The transition kernel is typically defined using some predefined proposal distribution: a set of new parameter values is proposed from this proposal distribution, and these values are subsequently either accepted or rejected based on a given acceptance probability, which is a function of the proposal distribution. If the proposed values are accepted, the Markov chain moves to this new state; whereas if the values are rejected, the Markov chain remains in the same state at the next iteration. We note that the transition kernel is non-unique, in that there are many possible choices for this proposal distribution that will lead to the correct stationary distribution. Common proposal distributions include: the posterior conditional distribution, leading to the Gibbs sampler where the acceptance probability in the updating step is equal to one; the Metropolis–Hastings random walk sampler, which randomly perturbs the parameter values from their current values; the slice sampler; and the No-U-Turn Sampler, among 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 using an MCMC approach. For further discussion, see, for example, REFS[74,89,90].

Assessing performance. The choice of transition kernel defines the performance of the MCMC algorithm by determining how long the Markov chain needs to be run to obtain reliable inference on the posterior distribution. Trace plots can display the values of the parameters over many iterations. One-dimensional trace plots are most commonly used; they describe the value of a parameter at each iteration of the Markov chain on the *y* axis against the iteration number on the *x* axis and are often a useful exploratory tool (FIG. 4a). In particular, trace plots provide a visualization of the chain in terms of how each parameter is exploring the parameter space — referred to as mixing. If this mixing is poor, in that the chain takes a long time to explore the posterior parameter space, changes to the specified transition kernel may be required. For example, poor mixing may be due to only very small parameter value changes between successive iterations or if there is a high rejection rate of the proposed parameter values, so that the parameter values remain the same across many successive iterations of the MCMC algorithm. These plots are also informally used for identifying when the Markov chain has reached its stationary distribution. Realizations of the chain prior to convergence to its stationary distribution are discarded; this process is commonly known as burn-in, although we prefer the term warm-up and refer to this process thus in this Primer[91].

The most common technique for assessing the convergence of a Markov chain to its stationary distribution is the $\hat{R}$ statistic, which is defined as the ratio of within-chain to between-chain variability[92,93]. In order to apply this approach, multiple independent runs of the MCMC algorithm need to be run (FIG. 4b). Ideally, each of the Markov chains should start from different starting values and using different random seeds in order to provide greater initial variability across the Markov chains, and to make it more likely that non-convergence of the chain to the stationary distribution will be identified. This non-convergence could happen, for example, if different sub-modes of the posterior distribution are being explored. Values close to one for all parameters and quantities of interest suggest that the chain has sufficiently converged to the stationary distribution, so that future realizations can be regarded as a sample from the posterior distribution (FIG. 4b). When the stationary distribution is reached, the number of iterations needed to obtain reliable, low-error Monte Carlo estimates can be determined. To assess the required number of iterations, the sampled values are often batched, which involves subdividing 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[94].

The effective sample size of the sampled parameter values may be obtained to provide an indication of the efficiency of the algorithm. The effective sample size 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 using a first-order Markov chain. Here, the effective sample size does not refer to the sample size of the data; rather, it is the effective length of the MCMC chain. Low sampling efficiency is related to high autocorrelation (and poor mixing) — so that the variability

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

Software package	Summary
General-purpose Bayesian inference software	
BUGS[231,232]	The original general-purpose Bayesian inference engine, in different incarnations. These use Gibbs and Metropolis sampling. Windows-based software (WinBUGS[233]) with a user-specified model and a black-box MCMC algorithm. Developments include an open-source version (OpenBUGS[234]) also available on Linux and Mac
JAGS[235]	An open-source variation of BUGS that can run cross-platform and can run from R via rjags[236]
PyMC3[237]	An open-source framework for Bayesian modelling and inference entirely within Python; includes Gibbs sampling and Hamiltonian Monte Carlo
Stan[98]	An open-source, general-purpose Bayesian inference engine using Hamiltonian Monte Carlo; can be run from R, Python, Julia, MATLAB and Stata
NIMBLE[238]	Generalization of the BUGS language in R; includes sequential Monte Carlo as well as MCMC. Open-source R package using BUGS/JAGS-model language to develop a model; different algorithms for model fitting including MCMC and sequential Monte Carlo approaches. Includes the ability to write novel algorithms
Programming languages that can be used for Bayesian inference	
TensorFlow Probability[239,240]	A Python library for probabilistic modelling built on Tensorflow[203] from Google
Pyro[241]	A probabilistic programming language built on Python and PyTorch[204]
Julia[242]	A general-purpose language for mathematical computation. In addition to Stan, numerous other probabilistic programming libraries are available for the Julia programming language, including Turing.jl[243] and Mamba.jl[244]
Specialized software doing Bayesian inference for particular classes of models	
JASP[245]	A user-friendly. higher-level interface offering Bayesian analysis. Open source and relies on a collection of open-source R packages
R-INLA[230]	An open-source R package for implementing INLA[246]. Fast inference in R for a certain set of hierarchical models using nested Laplace approximations
GPstuff[247]	Fast approximate Bayesian inference for Gaussian processes using expectation propagation; runs in MATLAB, Octave and R

MCMC, Markov chain Monte Carlo.

of the parameter values is small over successive iterations — and non-smooth histograms of posteriors. In these circumstances, longer simulations are typically required to obtain reliable estimates of the posterior distribution and sufficiently small Monte Carlo error in the estimated posterior summary statistics. The latter issue of a small effective sample size, in turn, could point towards potential problems in the model estimation or weak identifiability of the parameters[21]. Therefore, when problems occur in obtaining reliable Monte Carlo estimates, a good starting point is to sort all variables based on effective sample size and investigate those with the lowest effective sample size first. Effective sample size is also useful for diagnosing the sampling efficiency for a large number of variables[95].

Computer software. There are now many standard computational packages for implementing Bayesian analyses (TABLE 2), which have subsequently led to the growth of Bayesian inference across many scientific fields. Many of the available packages perform the MCMC algorithm as a black box — although often with options to change the default settings — permitting the analyst to focus on the prior and model specification, and avoid any technical coding. There are many additional packages that make it easier to work with the sometimes heavily code-based software, such as the packages BRMS[96] and Blavaan[97] in R for simplifying the use of the probabilistic programming language Stan[98].

Empirical example 1 continued. The priors for the PhD delay example were updated with the data, and posteriors were computed in Stan[98]. The trace plot of four independent runs of the MCMC algorithms for $\beta_{\text{intercept}}$ is shown in FIG. 4a, displaying stability post warm-up. The associated $\hat{R}$ statistic stabilizes after approximately 2,000 iterations (FIG. 4b). The prior and posterior distributions are displayed in FIG. 4c–e. As can be seen, the priors and posteriors are very close to each other, indicating that our prior knowledge is 'confirmed' by the newly collected data. Also, it can be seen that the uncertainty has decreased (for example, the posterior variances are smaller compared with the prior variances), indicating that we have updated our knowledge. To illustrate how easy it is to compute parameter interrelations, we also plotted the prior and posterior marginal densities between β_{age} and β_{age^2} (FIG. 4f).

Variational inference. As we have outlined, Bayesian analysis consists of numerous stages including detailed model development, specifying the prior and data models, the derivation of exact inference approaches based on MCMC, and model checking and refinement. Each of these stages 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 the gold standard for Bayesian inference. Monte Carlo methods for Bayesian

Variational inference
A technique to build approximations to the true Bayesian posterior distribution using combinations of simpler distributions whose parameters are optimized to make the approximation as close as possible to the actual posterior.

Approximating distribution
In the context of posterior inference, replacing a potentially complicated posterior distribution with a simpler distribution that is easy to evaluate and sample from. For example, in variational inference, it is common to approximate the true posterior with a Gaussian distribution.

Stochastic gradient descent
An algorithm that uses a randomly chosen subset of data points to estimate the gradient of a loss function with respect to parameters, providing computational savings in optimization problems involving many data points.

Multicollinearity
A situation that arises in a regression model when a predictor can be linearly predicted with high accuracy from the other predictors in the model. This causes numerical instability in the estimation of parameters.

Shrinkage priors
Prior distributions for a parameter that shrink its posterior estimate towards a particular value.

Sparsity
A situation where most parameter values are zero and only a few are non-zero.

Spike-and-slab prior
A shrinkage prior distribution used for variable selection specified as a mixture of two distributions, one peaked around zero (spike) and the other with a large variance (slab).

inference adopt a simulation-based strategy for approximating posterior distributions. An alternative approach is to produce functional approximations of the posterior using techniques including variational inference[99] or expectation propagation[100]. Here, we describe variational inference, also known as variational methods or variational Bayes, owing to its popularity and prevalence of use in machine learning.

Variational inference begins by constructing an approximating distribution to estimate the desired — but intractable — posterior distribution. Typically, the approximating distribution chosen is from a family of standard probability distributions, for example multivariate normal distributions, and further assumes that some of the dependencies between the variables in our model are broken to make subsequent computations tractable. In the case where the approximating distribution assumes all variables are independent, this gives us the mean-field approximation. The approximating distribution will be specified up to a set of variational parameters that we optimize to find the best posterior approximation by minimizing the Kullback–Leibler divergence from the true posterior. As a consequence, variational inference reframes Bayesian inference problems as optimization problems rather than as sampling problems, allowing them to be solved using numerical optimization. When combined with subsampling-based optimization techniques such as stochastic gradient descent, variational inference makes approximate Bayesian inference possible for complex large-scale problems[101–103].

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 are available. The inclusion of unnecessary variables in a model has several disadvantages, such as increasing the risk of multicollinearity, insufficient samples to estimate all model parameters, overfitting the current data leading to poor predictive performance on new data and making 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 markers 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 probabilities, whereas penalized parameter estimation approaches specify shrinkage priors that induce sparsity. 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. On the other hand, penalization methods fit a single model and are able to scale up to high-dimensional data.

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, $\boldsymbol{y}$, is 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 coefficient $\boldsymbol{\beta}$ captures the effect of the covariates on the response variable and $\boldsymbol{\varepsilon}$ represents the residuals assumed to follow a normal distribution with mean zero and variance σ^2.

Bayes factors and posterior model probabilities. Bayes factors[64] (BOX 2) can be used to compare and choose between candidate models, where each candidate model corresponds 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 predictors[104,105]. These 2^p possible models can be indexed by a binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, 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_γ be the model that includes the X_j values with $\gamma_j = 1$. Prior distributions for each model, $p(M_\gamma)$, and for the parameters under each model, $p(\beta_\gamma, \sigma^2|M_\gamma)$, are specified, and Bayes factors $BF_{\gamma b}$ are evaluated to compare each model M_γ with one of the models taken as a baseline, M_b. The posterior probability, $p(M_\gamma|\boldsymbol{y})$, for each model can be expressed in terms of the Bayes factors as:

$$p(M_\gamma|\boldsymbol{y}) = \frac{BF_{\gamma b}\, p(M_\gamma)}{\sum_{\gamma'} BF_{\gamma' b}\, p(M_{\gamma'})} \tag{12}$$

where the denominator sums over all considered models $M_{\gamma'}$. The models with the largest posterior probabilities would correspond to models with the highest amount of evidence in their favour among those under consideration. When p is relatively small (for example, <20), all 2^p variable subsets and their posterior probabilities can be evaluated. The model with the highest posterior probability may be selected as the one most supported by the data. Alternatively, the covariates with high marginal posterior inclusion probabilities, $p(\gamma_j = 1|\boldsymbol{y}) = \sum_{X_j \in M_\gamma} p(M_\gamma|\boldsymbol{y})$, may be selected. For a 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.

Shrinkage priors. Various shrinkage priors have been proposed over the years. A widely used shrinkage prior is the spike-and-slab prior, 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) and the other a diffuse distribution (slab)[106,107]. The spike component identifies the zero elements whereas the slab component captures the non-zero coefficients. The discrete spike-and-slab formulation[106] uses a mixture of a point mass at zero and a diffuse prior (FIG. 5a), whereas the continuous spike-and-slab

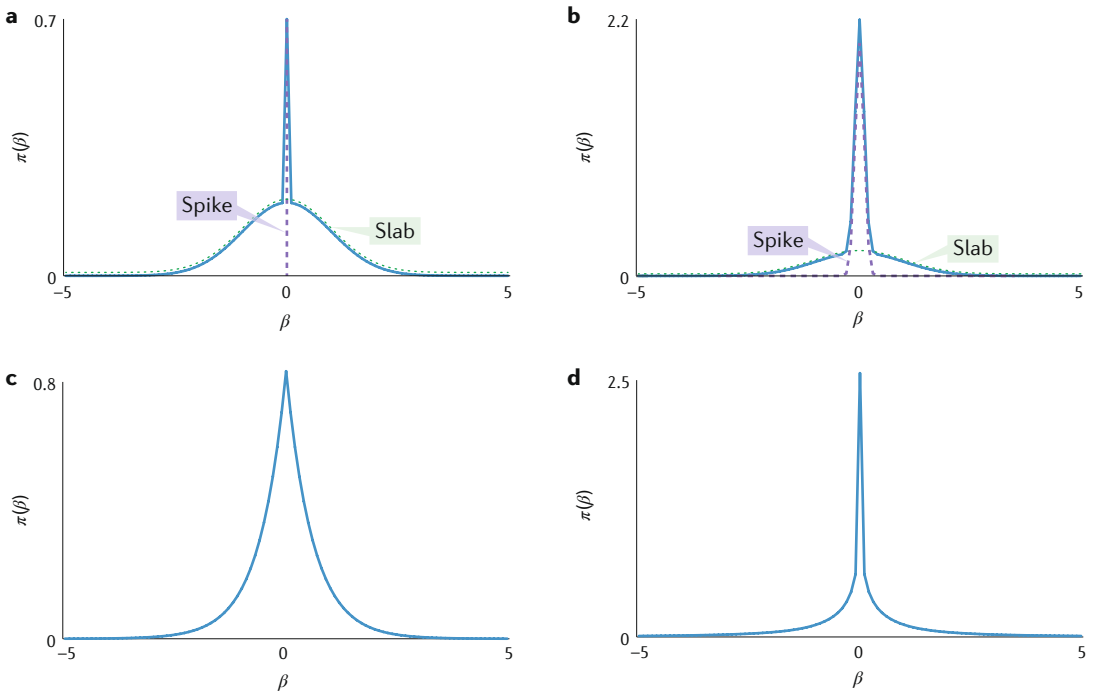

Fig. 5 | **Examples of shrinkage priors for Bayesian variable selection.** Prior density $\pi(\beta)$ versus β values. **a** | The discrete spike-and-slab prior for β (solid line) is specified as a mixture of a point mass at zero (spike; dashed line) and a diffuse prior (slab; dotted line). **b** | The continuous spike-and-slab prior for β (solid line) is specified as a mixture of two normal distributions, one peaked around zero (dashed line) and the other with a large variance (dotted line). **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 simultaneously allow a lot of probability mass around zero and heavy tails. **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.

prior[107] uses a mixture of two continuous distributions (FIG. 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 β and σ^2 can accelerate the MCMC implementation, 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] (TABLE 1).

Another class of penalization 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 the shrinkage of small regression coefficients towards zero, similar 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, or lasso[114], 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 estimate that maximizes the posterior distribution under independent Laplace distribution priors. Motivated by this connection, the Bayesian lasso[111] specifies conditional Laplace priors on $\beta_j|\sigma^2$. Unlike the frequentist lasso method, Bayesian penalization methods do not shrink regression coefficients to be exactly zero. 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 parameterized as a scale mixture of normal distributions, which facilitates their implementation in MCMC methods. For example, the Laplace prior in the Bayesian lasso can be formulated as a scale mixture of normal distributions with an exponential mixing density for the scale parameter. The exponential mixing distribution has a single hyperparameter, which limits its flexibility in differentially shrinking small and large effects (FIG. 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, an example of a global–local shrinkage prior, achieves the tight peak around zero and the heavy tails by specifying a normal distribution for the regression coefficient β_j, conditional on its scale parameters, which themselves follow half-Cauchy distributions[112] (FIG. 5d). A comprehensive review and thorough comparison of the

Continuous shrinkage prior
A unimodal prior distribution for a parameter that promotes shrinkage of its posterior estimate towards zero.

Global–local shrinkage prior
A continuous shrinkage prior distribution characterized by a high concentration around zero to shrink small parameter values to zero and heavy tails to prevent excessive shrinkage of large parameter values.

Horseshoe prior
An example of a global–local shrinkage prior for variable selection that uses a half-Cauchy scale mixture of normal distributions.

characteristics and performance of different shrinkage priors can be found in REF.[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 response variables[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 effect and time-varying coefficient models[118,119], mixture models for unsupervised clustering[120] and estimation of single and multiple Gaussian graphical models[121,122].

Variable selection in biomedicine. Variable selection priors for linear models have found important applications in biomedical studies. 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 significant predictors. In Bayesian approaches, additional knowledge about correlations 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 of gene-to-gene interaction networks have been employed to aid the identification of predictive genes[123], as well as the identification of both relevant pathways and subsets of genes[124]. Bayesian variable selection priors have been successfully applied in genome-wide association studies, where hundreds of thousands of single-nucleotide polymorphisms 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[125,126].

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, which can be factored into Bayesian models of air pollution (for a resource on Bayesian hierarchical models for spatial data we refer readers to REF.[127]). Spatially varying coefficient models with spike-and-slab priors inducing spatial correlation have been proposed to identify pollutants associated with adverse health outcomes, either over a whole region or within separate subregions[128]. Over the past couple of decades, numerous omics studies have been conducted to investigate the effects of exposure to air pollution 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 between markers have been proposed to identify and estimate the effect of pollutants on DNA methylation outcomes[117].

In neuroscience, neuroimaging studies often employ functional MRI, 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 across multiple subjects. Task-based experiments use functional MRI to scan the brain dynamically while the subject is presented to different external stimuli. The data are analysed with the goal of identifying brain regions that are activated by these stimuli. Bayesian general linear models with spatial priors, which allow flexible modelling of the correlation structure in these data, have been successfully applied[129]. 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 activation patterns[130,131]. Another application of functional MRI is in brain connectivity studies, where data are measured on subjects at rest with the aim of understanding how brain regions interact with each other. Among other approaches, multivariate vector autoregressive linear models have been investigated as a way to infer effective connectivity. Continuous shrinkage priors and structured spike-and-slab prior constructions have been employed for the selection of the active connections[132,133]. Bayesian variable selection methods have been successfully applied to numerous other biomedical data sets, including longitudinal data, functional data, survival outcome data and case–control studies.

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 that might be helpful to assess whether the model provides valid predictions so that these can be used for extrapolating to future events. Those simulations can be used for several purposes. They can be used to check whether the simulated data from the model resemble the observed data by comparing kernel density estimates of the observed data with density estimates for the simulated data[57]. A more formal posterior predictive checking approach can be taken to evaluate whether the model can be considered a good fit with the data-generating mechanism[57,78,134–136]. Any parameter-dependent statistic or discrepancy can be used for posterior predictive checking[135]. This is similar to how prior predictive checking can be used, but much more stringent in the comparison between the observed and simulated data[57]. The sensitivity of the posterior predictive checks is useful because if realistic models are used, the expectation is that the results are well calibrated in the long-term average[78]. These two uses of posterior predictive checking should be used with care; there is a risk of over-adjusting and over-refining models to the details of a specific data set. Posterior predictive distributions can further be used to extrapolate beyond the observed data and make predictions, for example extrapolating data from a time series. Based on the posterior distributions for a particular model of interest, posterior predictive distributions can be simulated for observed and future data, naturally becoming more uncertain as they predict further into the future owing to accumulated uncertainty. It is important to be aware

that in temporal models there are some challenges in terms of posterior inference that are inherent to spatial and/or temporal dependencies, such as autocorrelation of parameters over time[52,137–139].

Empirical example 2: Wikipedia page views. To illustrate the use of posterior predictive distributions, we present a second example. Suppose that it is of interest to know how many page views a webpage has, and what time-related factors might be relevant to page views. Consider the page views for the Wikipedia page on the English Premier League — the highest level of the English professional football league — obtained using the wikipediatrend[140] R package. The scripts are available at the Open Science Framework[141]. The decomposable time-series model[142], implemented in the prophet[143] R package, allows the estimation of trends with non-periodic changes, holiday effects, weekly seasonality and yearly seasonality effects (FIG. 6). 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 September 2011 — the wedding day of Prince William and Catherine Middleton. Additionally, a decrease in page views occurs each Christmas day and notable increases occur on Boxing day and at the start of the year, when matches are played during the Christmas holiday season. The model is estimated using observed data in the period between 1 January 2010 and 1 January 2018. Based on the posterior distributions for the particular model, posterior predictive distributions can be simulated for observed and future data (FIG. 6e,f). 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 into the future owing 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 that might relate to the final stage of the FIFA World Cup, which was played at that time.

Applications

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

Social and behavioural sciences

A recent systematic review examining the use of Bayesian statistics reported that the social and behavioural sciences — psychology, sociology and political sciences — have experienced an increase in empirical Bayesian work[4]. Specifically, there have been two parallel uses of Bayesian methods that have been increasing in popularity within the social and behavioural sciences: theory development and as a tool for model estimation.

Bayes' rule has been used as an underlying theory for understanding reasoning, decision-making, cognition and theories of mind, and has been particularly prevalent in developmental psychology and related fields. 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. Gigerenzer and Hoffrage[145] discuss the use of frequencies, rather than probabilities, as a method to improve on Bayesian reasoning. In another seminal article, Slovic and Lichtenstein[146] discuss 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 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 distributions. Informative priors are useful in complex modelling situations, which are common in the social sciences, as well as in cases of small sample sizes. Certain models that are used to explore education outcomes and standardized tests, such as some multidimensional item response theory models, are intractable using frequentist statistics and require the use of Bayesian methods.

The number of publications concerning Bayesian statistics has been steadily rising since 2004, with a more notable increase in the last decade. In part, this focus on Bayesian methods is owing to the development of more accessible software, as well as a focus on publishing tutorials targeting applied social and behavioural scientists. A systematic review of Bayesian methods in the field of psychology uncovered 740 eligible regression-based articles using Bayesian methods. Of these, 100 articles (13.5%) were tutorials for implementing Bayesian methods, and an additional 225 articles (30.4%) were either technical papers or commentaries on Bayesian statistics (BOX 4). Methodologists have been attempting to guide applied researchers towards using Bayesian methods within the social and behavioural sciences, although the implementation has been relatively slow to catch on. For example, the systematic review found that only 167 regression-based Bayesian articles (22.6%) were applications using human samples. Nevertheless, some subfields are regularly publishing work implementing Bayesian methods.

The field has gained many interesting insights into psychological and social behaviour through Bayesian methods, and the substantive areas in which this work has been conducted are quite diverse. For example, Bayesian statistics has helped to uncover the role that craving suppression has in smoking cessation[147], to make population forecasts based on expert opinions[148], to examine the role that stress related to infant care has in divorce[149], to examine the impact of the President of the USA's ideology on US Supreme Court rulings[150] and to predict behaviours that limit the intake of free sugars in one's diet[151]. 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], that are expository[152,153], that focus on how Bayesian reasoning can inform

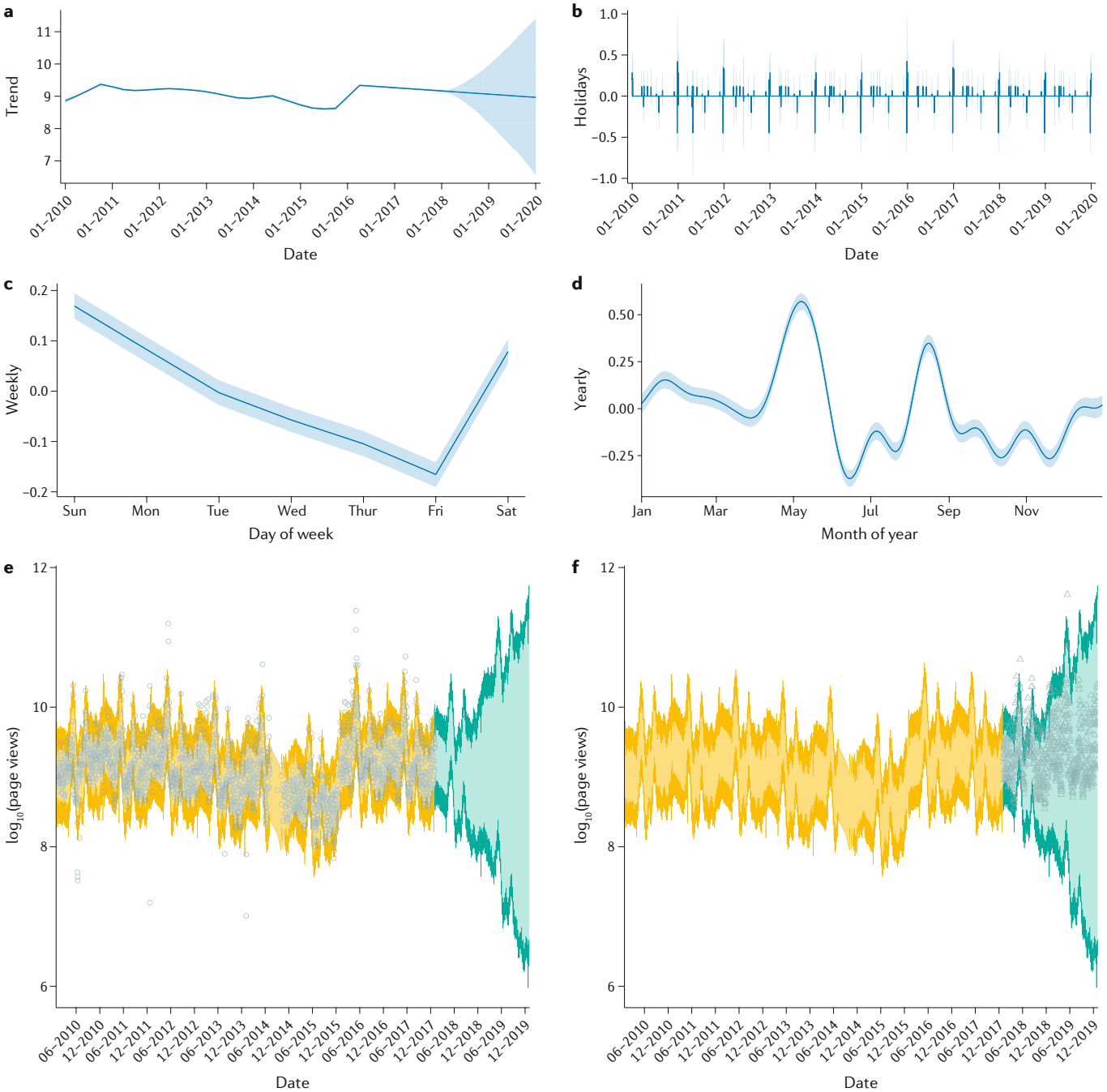

Fig. 6 | **Posterior predictive checking and predicted future page views based on current observations. a–d** | Posterior means along with 95% credible intervals (CIs) for non-periodic changes (part **a**), holiday effects (part **b**), weekly seasonality (part **c**) and yearly seasonality effects (part **d**). Displayed as how date-specific characteristics contribute to expected $\log_{10}$(page views). **e**,**f** | Posterior predictive distributions at each time point. 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 yellow (50% CI) and dark yellow (95% CI), whereas posterior predictive distributions for future data are presented in light green (50% CI) and dark green (95% CI). Page view observations are marked as grey circles (part **e**). Actual page view observations for the predicted time period are marked as grey triangles, overlaid on the posterior predictive distribution (part **f**). Page views for the Wikipedia page of the English Premier League as obtained using the wikipediatrend[140] R package and analysed with the prophet[143] R package — the scripts are available at the Open Science Framework[251].

theory through use of Bayesian inference[154] and that use Bayesian modelling to extract findings that would have been difficult to derive using frequentist methods[147]. 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

Box 4 | **Bayesian methods in the social and behavioural sciences**

Hoijtink et al.[255] discuss the use of Bayes factors for informative hypotheses within cognitive diagnostic assessment, illustrating how Bayesian evaluation of informative diagnostics hypotheses can be used as an alternative approach to traditional diagnostic methods. There is added flexibility with the Bayesian approach as informative diagnostic hypotheses can be evaluated using the Bayes factor utilizing only data from the individual being diagnosed. Lee[154] present an overview of the application of Bayes' theorem in the field of cognitive psychology, discussing how Bayesian methods can be used to develop more complete theories of cognitive psychology. Bayesian methods can also 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.[152] show how Bayesian methods can benefit health-based research being conducted in psychology by highlighting how informative priors elicited with 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 owing to the cost of data collection and the population being difficult to access for sampling. Finally, Kruschke[153] present the simplest example using a *t*-test geared towards 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.

results are replicable and transparent. 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.

Ecology

The application of Bayesian analyses to answer ecological questions has become increasingly widespread owing to both philosophical arguments, particularly in terms of subjective versus objective reasoning, and practical model-fitting advantages. This is combined with readily available software (TABLE 2) and numerous publications describing Bayesian ecological applications using these software packages (see REFS[155–161] for examples). The underlying Bayesian philosophy is attractive in many ways within ecology[162] as it permits the incorporation of external, independent prior information either from previous studies on the same/similar species or inherent knowledge of the biological processes within a rigorous framework[163,164]. Further, the Bayesian approach also permits both direct probabilistic statements to be made on parameters of interest, such as survival probabilities, reproductive rates, population sizes and future predictions[157], and the calculation of relative probabilities of competing models — such as the presence or absence of density dependence or environmental factors in driving the dynamics of the ecosystem — that in turn permits 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.[165] 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 ecological research 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. 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 an ecosystem level that includes understanding the population dynamics of the given system[166], modelling spatial point pattern data[167], investigating population genetics, estimating abundance[168] and assessing conservation management[169].

Ecological data collection processes generally come from observational studies, where a sample is observed from the population of interest using some data survey protocol. The survey should be carefully designed, taking into account the ecological question(s) of interest and minimizing the complexity of the model required to fit the data to provide reliable inference. Nevertheless, associated model-fitting challenges may still arise owing to data collection problems, such as those resulting from equipment failure or poor weather conditions. There may also be inherent data collection problems in some data surveys, such as the inability to record individual-level information. Such model-fitting challenges may include — but are far from limited to — irregularly spaced observations in time owing to equipment failure or experimental design, measurement error due to imperfect data observations, missing information at a range of different levels, from the individual level to the global environmental level, and challenges associated with multiscaled studies where different aspects of data are recorded at different temporal scales — for example, from hourly location data of individuals to daily and monthly collections of environmental data. The data complexities that arise, combined with the associated modelling choices, may lead to a range of model-fitting challenges that can often be addressed using standard techniques within the Bayesian paradigm.

For a given ecological study, separating out the individual processes acting on the ecosystem is an attractive mechanism for simplifying model specification[166]. For example, state space models provide a general and flexible modelling framework that describes two distinct types of process: the system process and the observation process. The system process describes the true underlying state of the system and how this changes over time. This state may be univariate or multivariate, such as population size or location data, respectively. The system process may also describe multiple processes acting on the system, such as birth, reproduction, dispersal and death. We are typically not able to observe these true underlying system states without some associated error and the observation process describes how the observed data relate to the true unknown states. These general state space models span many applications, including

animal movement[170], population count data[171], capture–recapture-type data[165], fisheries stock assessment[172] and biodiversity[173]. For a review of these topics and further applications, we direct the reader elsewhere[166,174,175]. Bayesian model-fitting tools, such as MCMC with data augmentation[176], sequential Monte Carlo or particle MCMC[177–179], permit general state space models to be fitted to the observed data without the need to specify further restrictions — such as distributional assumptions — on the model specification, or to make additional likelihood approximations.

The process of collecting data continues to evolve with advances in technology. For example, the use of GPS geolocation tags and associated additional accelerometers, remote sensing, the use of drones for localized aerial photographs, unmanned underwater vehicles and motion-sensor camera traps are increasingly used within ecological research. The use of these technological devices and the growth of crowdsourced science have led to new forms of data collected in great quantities and associated model-fitting challenges, providing a fertile ground for Bayesian analyses.

Genetics

Genetics and genomics studies have made extensive use of Bayesian methods. In genome-wide association studies, Bayesian approaches have provided a powerful alternative to frequentist approaches for assessing associations between genetic variants and a phenotype of interest in a population[180]. These include statistical models for incorporating genetic admixture[181], fine-mapping to identify causal genetic variants[182], imputation of genetic markers not directly measured using reference populations[183] and meta-analysis for combining information across studies. These applications further benefit from the use of marginalization to account for modelling uncertainties when drawing inferences. More recently, large cohort studies such as the UK Biobank[184] have expanded the methodological requirements for identifying genetic associations with complex (sub)phenotypes by collating genetic information alongside heterogeneous data sets including imaging, lifestyle and routinely collected health data. A Bayesian analysis framework known as TreeWAS[185] 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, such as information from the latest version of the International Classification of Diseases (ICD-10).

The availability of multiple molecular data types in multi-omics data sets 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 uncharacterized biological processes across different molecular scales. Multi-omics factor analysis[186] uses a Bayesian linear factor model to disentangle sources of heterogeneity that are common across multiple data modalities from those patterns that are specific to only a single data modality.

In recent years, high-throughput molecular profiling technologies have advanced to allow the routine multi-omics analysis of individual cells[187]. This has led to the development of many novel approaches for modelling single-cell measurement noise, cell-to-cell heterogeneity, high dimensionality, large sample sizes and interventional effects from, for example, genome editing[188]. Cellular heterogeneity lends itself naturally to Bayesian hierarchical modelling and formal uncertainty propagation and quantification owing to the layers of variability induced by tissue-specific activity, heterogeneous cellular phenotypes within a given tissue and stochastic molecular expression at the level of the single cell. In the integrated Bayesian hierarchical model BASiCS[189], this approach is used to account for cell-specific normalization constants and technical variability to decompose total gene expression variability into technical and biological components.

Deep neural networks (DNNs) have also been utilized to specify flexible, non-linear conditional dependencies within hierarchical models for single-cell omics. SAVER-X[190] couples a Bayesian hierarchical model with a pretrainable deep autoencoder to extract transferable gene–gene relationships across data sets from different laboratories, variable experimental conditions and divergent species to de-noise novel target data sets. In scVI[191], hierarchical modelling is used to pool information across similar cells and genes to learn models of the distribution of observed expression values. Both SAVER-X and scVI perform approximate Bayesian inference using mini-batch stochastic gradient descent, the latter within a variational setting — a standard technique in DNNs — that allow these models to be fitted to hundreds of thousands or even millions of cells.

Bayesian approaches have also been popular in large-scale cancer genomic data sets[192] and have enabled a data-driven approach to identifying novel molecular changes that drive cancer initiation and progression. Bayesian network models[193] have been developed to identify the interactions between mutated genes and capture mutational signatures that highlight key genetic interactions with the potential to allow for genomic-based patient stratification in both clinical trials and the personalized use of therapeutics. Bayesian methods have also 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 their ancestral relationships through the analysis of single-cell and bulk tissue-sequencing data[194]. These models therefore consider the joint problem of learning a mixture model and graph inference through considering the number and identity of the subpopulations and deriving the phylogenetic tree.

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[195]. A workflow incorporating good research practices that encourage reproducibility

Autoencoder
A particular type of multilayer neural network used for unsupervised learning consisting of two components: an encoder and a decoder. The encoder compresses the input information into low-dimensional summaries of the inputs. The decoder takes these summaries and attempts to recreate the inputs from these. By training the encoder and decoder simultaneously, the hope is that the autoencoder learns low-dimensional, but highly informative, representations of the data.

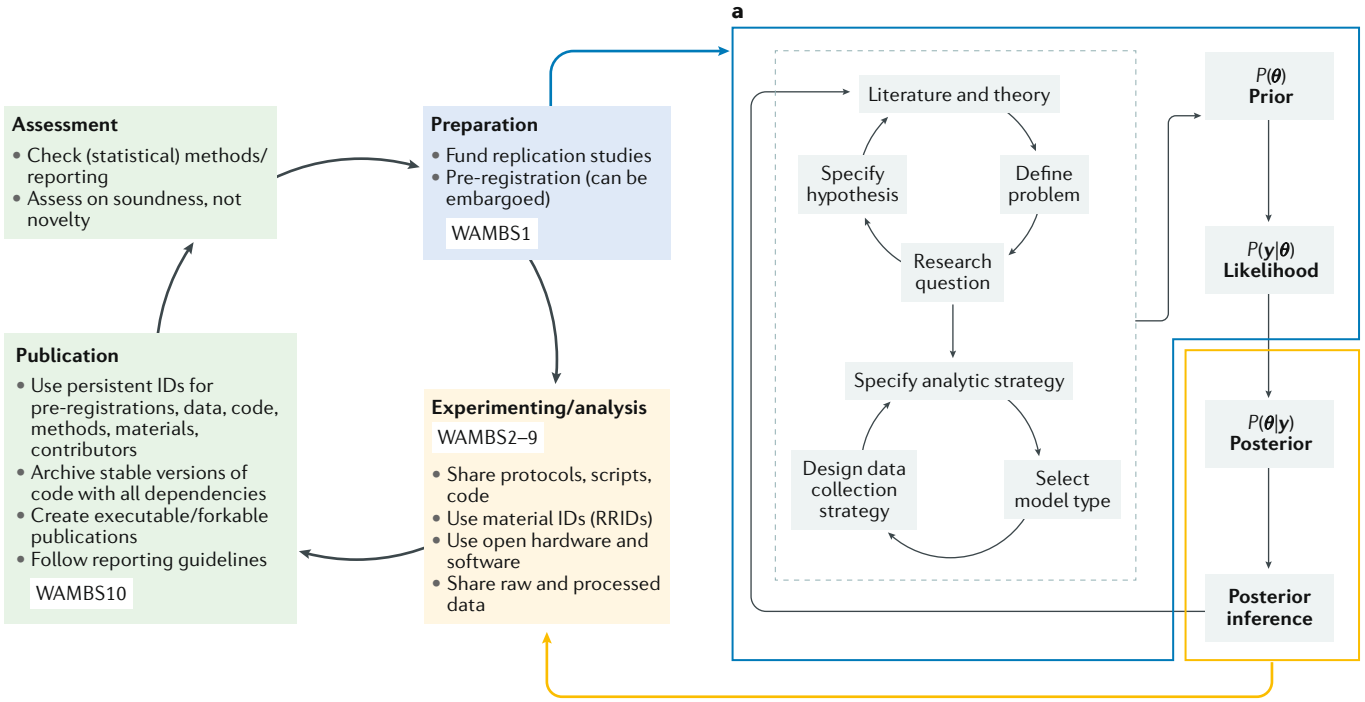

Fig. 7 | **Elements of reproducibility in the research workflow.** Good research practices across the research workflow that can contribute to reproducibility, demonstrating where the Bayesian research cycle (part **a**) and the WAMBS (when to Worry and how to Avoid the Misuse of Bayesian Statistics) checklist fit in the wider context of transparency in research[252]. Not all elements are applicable to all types of research — for example, pre-registration is typically used for hypothesis-driven research — but the specification of the prior and the likelihood may be pre-registered. There may be legitimate reasons why data cannot be shared openly, but all scripts for running the Bayesian models could be shared on a data repository. $\boldsymbol{\theta}$, unknown parameter; RRID, Research Resource Identifier; $P(.)$, probability distribution; $\boldsymbol{y}$, data.

in the Bayesian research cycle is displayed in FIG. 7. We demonstrate where the Bayesian research cycle (FIG. 1) and the WAMBS (when to Worry and how to Avoid the Misuse of Bayesian Statistics) checklist[48,81] fit in the wider context of transparency in research, and we offer an updated version of the WAMBS checklist (BOX 5). In this section, we highlight some important aspects of reproducibility and the deposition of data and scripts.

Allowing others to assess the statistical methods and underlying data used in a study (by transparent reporting and making code and data available) can help with interpreting the study results, the assessment of the suitability of the parameters used and the detection and fixing of errors. Reporting practices are not yet consistent across fields or even journals in individual fields. A systematic review on Bayesian statistics in psychology[4] found huge discrepancies in reporting practices and standards across the social sciences; for example, of the 167 regression-based Bayesian articles 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.

Not reporting any information on the priors is problematic for any Bayesian paper. There are many dangers in naively using priors and, we argue, one might want to pre-register the specification of the priors and the likelihood when possible. Moreover, 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, with Markov chains that appear to be converged and posteriors that appear appropriate 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} that was mis-specified, using precision instead of variance (FIG. 3).

To enable reproducibility and allow others to run Bayesian statistics on the same data with different parameters, priors, models or likelihood functions for sensitivity analyses[49], it is important that the underlying data and code used are properly documented and shared following the FAIR principles[196,197]: findability, accessibility, interoperability and reusability. Preferably, data and code should be shared in a trusted repository (Registry of Research Data Repositories) with their own persistent identifier (such as a DOI), and tagged with metadata describing the data set or codebase. This also allows the data set and the code to be recognized as separate research outputs and allows others to cite them accordingly[198]. Repositories can be general, such as Zenodo; language-specific, such as CRAN for R packages and `PyPI` for Python code; or

Split-$\hat{R}$
To detect non-stationarity within individual Markov chain Monte Carlo chains (for example, if the first part shows gradually increasing values whereas the second part involves gradually decreasing values), each chain is split into two parts for which the $\hat{R}$ statistic is computed and compared.

domain-specific[198]. As data and code require different licence options and metadata, data are generally best stored in dedicated data repositories, which can be general or discipline-specific[199]. Some journals, such as *Scientific Data*, have their own list of recommended data repositories. To make depositing data and code easier for researchers, two repositories (Zenodo and Dryad) are exploring collaboration to allow the deposition of code and data through one interface, with data stored in Dryad and code stored in Zenodo[200]. Many scientific journals adhere to transparency and openness promotion guidelines[201], which specify requirements for code and data sharing.

Verification and reproducibility require access to both the data and the code used in Bayesian modelling, ideally replicating the original environment in which the code was run, with all dependencies documented either in a dependency file accompanying the code or by creating a static container image that provides a virtual environment in which to run the code[199]. 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 themselves. However, open-source software is only truly accessible with proper documentation, which includes listing dependencies and configuration instructions in Readme files, commenting on code to explain functionality and including a comprehensive reference manual when releasing packages.

Box 5 | **The ten checklist points of WAMBS-v2**

WAMBS-v2, an updated version of the WAMBS (when to Worry and how to Avoid the Misuse of Bayesian Statistics) checklist.

- Ensure the prior distributions and the model or likelihood are well understood and described in detail in the text. Prior-predictive checking can help identify any prior–data conflict.
- Assess each parameter for convergence, using multiple convergence diagnostics if possible. This may involve examining trace plots or ensuring diagnostics ($\hat{R}$ statistic or effective sample size) are being met for each parameter.
- Sometimes convergence diagnostics such as the $\hat{R}$ statistic can fail at detecting non-stationarity within a chain. Use a subsequent measure, such as the split-$\hat{R}$, to detect trends that are missed if parts of a chain are non-stationary but, on average, appear to have reached diagnostic thresholds.
- 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.
- Examine the effective sample size for all parameters, checking for strong degrees of autocorrelation, which may be a sign of model or prior mis-specification.
- Visually examine the marginal posterior distribution for each model parameter to ensure that they do not have irregularities that could have resulted from misfit or non-convergence. Posterior predictive distributions can be used to aid in examining the posteriors.
- Fully examine multivariate priors through a sensitivity analysis. These priors can be particularly influential on the posterior, even with slight modifications to the hyperparameters.
- To fully understand the impact of subjective priors, compare the posterior results with an analysis using diffuse priors. This comparison can facilitate a deeper understanding of the impact the subjective priors have 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.
- 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.
- Report findings, including Bayesian interpretations. Take advantage of explaining and capturing the entire posterior rather than simply a point estimate. It may be helpful to examine the density at different quantiles to fully capture and understand the posterior distribution.

Limitations and optimizations

The optimization of Bayesian inference is conditional on the assumed model. Bayesian posterior probabilities are calibrated as long-term averages if parameters are drawn from the prior distribution and data are drawn from the model of the data given these parameters. Events with a stated probability occur at that frequency in the long term, when averaging over the generative model. In practice, our models are never correct. There are two ways we would like to overcome this limitation: by identifying and fixing problems with the model; and by demonstrating that certain inferences are robust to reasonable departures from the model.

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 that represents the effect of some treatment. If this estimate z is normally distributed with standard error s, we can write $z \sim N(\theta, s^2)$, a normal distribution parameterized by its location and scale parameter. Suppose that θ has a flat uniform prior distribution, then the posterior distribution is $\theta \sim N(z, s^2)$. Now suppose we observe $z = s$; that is, the estimate of θ is one standard error from zero. 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 was zero. But the Bayesian calculation gives a posterior probability $\Pr(\theta > 0|z) = 0.84$. This makes the calibration of the probability questionable (calibrated inferences or predictions are correct on average, conditional on the prediction).

In this example, the probability is calibrated if you average over the prior. It is mathematically impossible to average over a uniform distribution on an infinite range, but we could consider a very diffuse prior, for example $\theta \sim N(0, 1{,}000^2)$, where we are assuming that s is roughly on a unit scale, that is, is a dimensionless parameter that is expected to take on a value not far from one in absolute value. 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 does not appear 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 1 or 2 standard errors from 0, but they are rarely 5, 10 or 100 standard errors away. In this example, Bayesian inference, if taken literally, would lead to over-certainty: an 84% posterior probability. However, a positive way to look at this is that the evident problem with the posterior allowed us to recognize that prior information was available that we had not included in our model, in this case, prior information that it would be unlikely to see very large values of θ. Moreover, a weakly informative

prior such as $\theta \sim N(0, s^2)$ does not have a large impact on the posterior, as then the posterior becomes normal:

$$N\left(\frac{s}{2}, \frac{s^2}{2}\right),$$

so $\Pr(\theta > 0|z) = 0.76$, compared with 0.84 from our previous example. 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. The important thing about this example is not the specific numbers, which will depend on the context, but the idea that any statistical method should be evaluated over the range of problems to which it will be applied.

More generally, Bayesian models can be checked by comparing posterior predictive simulations with data[135] and by estimating the out-of-sample predictive error[202]. There is a benefit to strong prior distributions that 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 to tell a plausible story, and all models come from researcher decisions. Any choice of model has implications; the flat prior is weak, providing no shrinkage of the estimate, but can lead to a strong, possibly inappropriate, level of certainty about θ.

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 allow the creation of bespoke models and computational approaches for particular problems. 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.

An ongoing challenge for Bayesian statistics is the ever-growing demands posed by increasingly complex real-world applications, which are often associated with issues such as large data sets 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 recent years, the revision and popularization of the term artificial intelligence to encompass a broad range of ideas including statistics and computation has blurred the traditional boundaries between these disciplines. This has been hugely successful in popularizing probabilistic modelling and Bayesian concepts outside their 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 at the innovative forefront of research in artificial intelligence.

Driven by the need to support large-scale applications involving data sets of increasing dimensionality and sample numbers, Bayesian concepts have exploited the growth of new technologies centred on deep learning. This includes deep learning programming frameworks (TensorFlow[203], PyTorch[204]), which simplify the use of DNNs, permitting the construction of more expressive, data-driven models that are immediately amenable to inference techniques using off-the-shelf optimization algorithms and state-of-the-art hardware. In addition to providing a powerful tool to specify flexible and modular generative models, DNNs have been employed to develop new approaches for approximate inference and stimulated a new paradigm for Bayesian practice that sees the integration of statistical modelling and computation at its core.

An archetypal example is the variational autoencoder[205], which has been successfully used in various applications, including single-cell genomics[190,191], providing a general modelling framework that has led to numerous extensions including latent factor disentanglement[206–208]. The underlying statistical model is a simple Bayesian hierarchical latent variable model, which maps high-dimensional observations to low-dimensional latent variables assumed to be normally distributed through functions defined by DNNs. Variational inference is used to approximate the posterior distribution over the latent variables. However, in standard variational inference 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. Variational autoencoders use a further approximation process known as amortization to replace inference over the many individual variational parameters with a single global set of parameters — known as a recognition network — that are used to parameterize a DNN that outputs the local variational parameters for each data point.

Remarkably, when the model and inference are combined and interpreted together, the variational autoencoder 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 variational autoencoder, demonstrating the increasingly blurry boundary between principled Bayesian modelling and algorithmic deep learning techniques. Other recent examples include the use of DNNs to construct probabilistic models that define distributions over possible functions[209–211], build complex probability distributions by applying a sequence of invertible transformations[212,213] and define models for exchangeable sequence data[214].

The expressive power of DNNs and their utility within model construction and inference algorithms come with compromises that will require Bayesian research. The trend towards entangling models and inference

Amortization
A technique used in variational inference to reduce the number of free parameters to be estimated in a variational posterior approximation by replacing the free parameters with a trainable prediction function that can instead predict the values of these parameters.

has popularized these techniques for large-scale data problems; however, fundamental Bayesian concepts remain to be fully incorporated within this paradigm. Integrating-out, model-averaging decision theoretic approaches rely on accurate posterior characterization, which remains elusive owing to the challenge posed by high-dimensional neural network parameter spaces[215]. Although Bayesian approaches to neural network learning have been around for decades[216–219], further investigation into prior specifications for modern Bayesian deep learning models that involve complex network structures is required to understand how priors translate to specific functional properties[220].

Recent debates within the field of artificial intelligence have questioned the requirement for Bayesian approaches and highlighted potential alternatives. For instance, deep ensembles[221] 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[222]. Similarly, dropout is a regularization approach popularized for use in the training of DNNs to improve robustness by randomly dropping out nodes during the training of the network[223]. 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[224]. Although the full extent of Bayesian principles has not yet been generalized to all recent developments in artificial intelligence, it is nonetheless a success that Bayesian thinking is deeply embedded and crucial to numerous innovations that have arisen. The next decade is sure to bring a new wave of exciting innovative developments for Bayesian intelligence.

Published online: 14 January 2021

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. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **53**, 370–418 (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. *Educ. Rev.* https://doi.org/10.1080/00131911.2017.1350636 (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. *Psychol. Methods* **22**, 217–239 (2017).
5. Ashby, D. Bayesian statistics in medicine: a 25 year review. *Stat. Med.* **25**, 3589–3631 (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.* https://doi.org/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.* https://doi.org/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. *Organ. Res. 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. *Struct. Equ. Modeling* **27**, 131–161 (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. *Struct. Equ. Modeling* **11**, 424–451 (2004).
11. 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).
12. van de Schoot, R. Online stats training. *Zenodo* https://zenodo.org/communities/stats_training (2020).
13. Heo, I. & van de Schoot, R. Tutorial: advanced Bayesian regression in JASP. *Zenodo* https://doi.org/10.5281/zenodo.3991325 (2020).
14. O'Hagan, A. et al. *Uncertain Judgements: Eliciting Experts' Probabilities* (Wiley, 2006).
This book presents a great collection of information with respect to prior elicitation, and includes elicitation techniques, summarizes potential pitfalls and describes examples across a wide variety of disciplines.
15. 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 (2000).
16. Veen, D., Stoel, D., Zondervan-Zwijnenburg, M. & van de Schoot, R. Proposal for a five-step method to elicit expert judgement. *Front. Psychol.* **8**, 2110 (2017).
17. 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 (2010).
18. Morris, D. E., Oakley, J. E. & Crowe, J. A. A web-based tool for eliciting probability distributions from experts. *Environ. Model. Softw.* https://doi.org/10.1016/j.envsoft.2013.10.010 (2014).
19. Garthwaite, P. H., Al-Awadhi, S. A., Elfadaly, F. G. & Jenkinson, D. J. Prior distribution elicitation for generalized linear and piecewise-linear models. *J. Appl. Stat.* **40**, 59–75 (2013).
20. Elfadaly, F. G. & Garthwaite, P. H. Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models. *Stat. Comput.* **27**, 449–467 (2017).
21. 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 (2020).
22. Runge, A. K., Scherbaum, F., Curtis, A. & Riggelsen, C. An interactive tool for the elicitation of subjective probabilities in probabilistic seismic-hazard analysis. *Bull. Seismol. Soc. Am.* **103**, 2862–2874 (2013).
23. 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.* https://doi.org/10.3389/fpsyg.2017.00090 (2017).
24. Cooke, R. M. & Goossens, L. H. J. TU Delft expert judgment data base. *Reliab. Eng. Syst. Saf.* **93**, 657–674 (2008).
25. Hanea, A. M., Nane, G. F., Bedford, T. & French, S. *Expert Judgment in Risk and Decision Analysis* (Springer, 2020).
26. Dias, L. C., Morton, A. & Quigley, J. *Elicitation* (Springer, 2018).
27. Ibrahim, J. G., Chen, M. H., Gwon, Y. & Chen, F. The power prior: theory and applications. *Stat. Med.* **34**, 3724–3749 (2015).
28. 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 (2011).
29. 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 (2018).
30. Berger, J. The case for objective Bayesian analysis. *Bayesian Anal.* **1**, 385–402 (2006).
This discussion of objective Bayesian analysis includes criticisms of the approach and a personal perspective on the debate on the value of objective Bayesian versus subjective Bayesian analysis.
31. Brown, L. D. In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Stat.* https://doi.org/10.1214/07-AOAS138 (2008).
32. 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. *J. Educ. Behav. Stat.* **28**, 169–194 (2003).
33. van der Linden, W. J. Using response times for item selection in adaptive testing. *J. Educ. Behav. Stat.* **33**, 5–20 (2008).
34. Darnieder, W. F. *Bayesian Methods for Data-Dependent Priors* (The Ohio State Univ., 2011).
35. Richardson, S. & Green, P. J. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Stat. Soc. Series B* **59**, 731–792 (1997).
36. Wasserman, L. Asymptotic inference for mixture models by using data-dependent priors. *J. R. Stat. Soc. Series B* **62**, 159–180 (2000).
37. Muthen, B. & Asparouhov, T. Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* **17**, 313–335 (2012).
38. 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 (2013).
39. Smeets, L. & van de Schoot, R. Code for the ShinyApp to determine the plausible parameter space for the PhD-delay data (version v1.0). *Zenodo* https://doi.org/10.5281/zenodo.3999424 (2020).
40. Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J. & Dorie, V. Weakly informative prior for point estimation of covariance matrices in hierarchical models. *J. Educ. Behav. Stat.* **40**, 136–157 (2015).
41. Gelman, A., Jakulin, A., Pittau, M. G. & Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2**, 1360–1383 (2008).
42. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* Vol. 2 (Chapman&HallCRC, 2004).
43. Jeffreys, H. *Theory of Probability* Vol. 3 (Clarendon, 1961).
44. Seaman III, J. W., Seaman Jr, J. W. & Stamey, J. D. Hidden dangers of specifying noninformative priors. *Am. Stat.* **66**, 77–84 (2012).
45. Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1**, 515–534 (2006).
46. 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 (2005).
47. Depaoli, S. Mixture class recovery in GMM under varying degrees of class separation: frequentist versus Bayesian estimation. *Psychol. Methods* **18**, 186–219 (2013).

48. Depaoli, S. & van de Schoot, R. Improving transparency and replication in Bayesian statistics: the WAMBS-Checklist. *Psychol. Methods* **22**, 240 (2017).
This article describes, in a step-by-step manner, the various points that need to be checked when estimating a model using Bayesian statistics. It can be used as a guide for implementing Bayesian methods.
49. van Erp, S., Mulder, J. & Oberski, D. L. Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychol. Methods* **23**, 363–388 (2018).
50. McNeish, D. On using Bayesian methods to address small sample problems. *Struct. Equ. Modeling* **23**, 750–773 (2016).
51. van de Schoot, R. & Miocević, M. *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners* (Taylor & Francis, 2020).
52. 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 (2016).
53. 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. *Struct. Equ. Modeling* **23**, 354–367 (2016).
54. Ranganath, R. & Blei, D. M. Population predictive checks. Preprint at https://arxiv.org/abs/1908.00882 (2019).
55. Daimon, T. Predictive checking for Bayesian interim analyses in clinical trials. *Contemp. Clin. Trials* **29**, 740–750 (2008).
56. Box, G. E. Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Stat. Soc. Ser. A* **143**, 383–404 (1980).
57. Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. Visualization in Bayesian workflow. *J. R. Stat. Soc. Ser. A* **182**, 389–402 (2019).
58. Silverman, B. W. *Density Estimation for Statistics and Data Analysis* Vol. 26 (CRC, 1986).
59. Nott, D. J., Drovandi, C. C., Mengersen, K. & Evans, M. Approximation of Bayesian predictive *p*-values with regression ABC. *Bayesian Anal.* **13**, 59–83 (2018).
60. Evans, M. & Moshonov, H. in *Bayesian Statistics and its Applications* 145–159 (Univ. of Toronto, 2007).
61. Evans, M. & Moshonov, H. Checking for prior–data conflict. *Bayesian Anal.* **1**, 893–914 (2006).
62. Evans, M. & Jang, G. H. A limit result for the prior predictive applied to checking for prior–data conflict. *Stat. Probab. Lett.* **81**, 1034–1038 (2011).
63. Young, K. & Pettit, L. Measuring discordancy between prior and data. *J. R. Stat. Soc. Series B Methodol.* **58**, 679–689 (1996).
64. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
This article provides an extensive discussion of Bayes factors with several examples.
65. Bousquet, N. Diagnostics of prior–data agreement in applied Bayesian analysis. *J. Appl. Stat.* **35**, 1011–1029 (2008).
66. Veen, D., Stoel, D., Schalken, N., Mulder, K. & van de Schoot, R. Using the data agreement criterion to rank experts' beliefs. *Entropy* **20**, 592 (2018).
67. Nott, D. J., Xueou, W., Evans, M. & Englert, B. Checking for prior–data conflict using prior to posterior divergences. Preprint at https://arxiv.org/abs/1611.00113 (2016).
68. Lek, K. & van de Schoot, R. How the choice of distance measure influences the detection of prior–data conflict. *Entropy* **21**, 446 (2019).
69. O'Hagan, A. Bayesian statistics: principles and benefits. *Frontis* **3**, 31–45 (2004).
70. Etz, A. Introduction to the concept of likelihood and its applications. *Adv. Methods Practices Psychol. Sci.* **1**, 60–69 (2018).
71. Pawitan, Y. *In All Likelihood: Statistical Modelling and Inference Using Likelihood* (Oxford Univ. Press, 2001).
72. Gelman, A., Simpson, D. & Betancourt, M. The prior can often only be understood in the context of the likelihood. *Entropy* **19**, 555 (2017).
73. Aczel, B. et al. Discussion points for Bayesian inference. *Nat. Hum. Behav.* **4**, 561–563 (2020).
74. Gelman, A. et al. *Bayesian Data Analysis* (CRC, 2013).
75. Greco, L., Racugno, W. & Ventura, L. Robust likelihood functions in Bayesian inference. *J. Stat. Plan. Inference* **138**, 1258–1270 (2008).
76. Shyamalkumar, N. D. in *Robust Bayesian Analysis Lecture Notes in Statistics* Ch. 7, 127–143 (Springer, 2000).
77. Agostinelli, C. & Greco, L. A weighted strategy to handle likelihood uncertainty in Bayesian inference. *Comput. Stat.* **28**, 319–339 (2013).
78. Rubin, D. B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* **12**, 1151–1172 (1984).
79. Gelfand, A. E. & Smith, A. F. M. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **85**, 398–409 (1990).
This seminal article identifies MCMC as a practical approach for Bayesian inference.
80. Geyer, C. J. Markov chain Monte Carlo maximum likelihood. *IFNA* http://hdl.handle.net/11299/58440 (1991).
81. van de Schoot, R., Veen, D., Smeets, L., Winter, S. D. & Depaoli, S. in *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners* Ch. 3 (eds van de Schoot, R. & Miocevic, M.) 30–49 (Routledge, 2020).
82. Veen, D. & Egberts, M. in *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners* Ch. 4 (eds van de Schoot, R. & Miocevic, M.) 50–70 (Routledge, 2020).
83. Robert, C. & Casella, G. *Monte Carlo Statistical Methods* (Springer Science & Business Media, 2013).
84. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984).
85. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **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. *Phys. Lett. B* **195**, 216–222 (1987).
88. Tanner, M. A. & Wong, W. H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **82**, 528–540 (1987).
This article explains how to use data augmentation when direct computation of the posterior density of the parameters of interest is not possible.
89. Gamerman, D. & Lopes, H. F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (CRC, 2006).
90. Brooks, S. P., Gelman, A., Jones, G. & Meng, X.-L. *Handbook of Markov Chain Monte Carlo* (CRC, 2011).
This book presents a comprehensive review of MCMC and its use in many different applications.
91. Gelman, A. Burn-in for MCMC, why we prefer the term warm-up. *Satistical Modeling, Causal Inference, and Social Science* https://statmodeling.stat.columbia.edu/2017/12/15/burn-vs-warm-iterative-simulation-algorithms/ (2017).
92. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–511 (1992).
93. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**, 434–455 (1998).
94. Roberts, G. O. Markov chain concepts related to sampling algorithms. *Markov Chain Monte Carlo in Practice* **57**, 45–58 (1996).
95. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P. Rank-normalization, folding, and localization: an improved $\tilde{R}$ for assessing convergence of MCMC. Preprint at https://arxiv.org/abs/1903.08008 (2020).
96. Bürkner, P.-C. Advanced Bayesian multilevel modeling with the R package brms. Preprint at https://arxiv.org/abs/1705.11123 (2017).
97. Merkle, E. C. & Rosseel, Y. blavaan: Bayesian structural equation models via parameter expansion. Preprint at https://arxiv.org/abs/1511.05604 (2015).
98. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* https://doi.org/10.18637/jss.v076.i01 (2017).
99. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
This recent review of variational inference methods includes stochastic variants that underpin popular approximate Bayesian inference methods for large data or complex modelling problems.
100. Minka, T. P. Expectation propagation for approximate Bayesian inference. Preprint at https://arxiv.org/abs/1301.2294 (2013).
101. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **14**, 1303–1347 (2013).
102. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
103. Li, Y., Hernández-Lobato, J. M. & Turner, R. E. Stochastic expectation propagation. *Adv. Neural Inf. Process. Syst.* **28**, 2323–2331 (2015).
104. Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. Mixtures of *g* priors for Bayesian variable selection. *J. Am. Stat. Assoc.* **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. *Int. Stat.Rev.* **86**, 237–258 (2018).
106. Mitchell, T. J. & Beauchamp, J. J. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* **83**, 1023–1032 (1988).
107. George, E. J. & McCulloch, R. E. Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **88**, 881–889 (1993).
This article popularizes the use of spike-and-slab priors for Bayesian variable selection and introduces MCMC techniques to explore the model space.
108. Ishwaran, H. & Rao, J. S. Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.* **33**, 730–773 (2005).
109. Bottolo, L. & Richardson, S. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal.* **5**, 583–618 (2010).
110. Ročková, V. & George, E. I. EMVS: the EM approach to Bayesian variable selection. *J. Am. Stat. Assoc.* **109**, 828–846 (2014).
111. Park, T. & Casella, G. The Bayesian lasso. *J. Am. Stat. Assoc.* **103**, 681–686 (2008).
112. Carvalho, C. M., Polson, N. G. & Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480 (2010).
113. Polson, N. G. & Scott, J. G. Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Stat.* **9**, 105 (2010).
This article provides a unified framework for continuous shrinkage priors, which allow global sparsity while controlling the amount of regularization for each regression coefficient.
114. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B* **58**, 267–288 (1996).
115. Van Erp, S., Oberski, D. L. & Mulder, J. Shrinkage priors for Bayesian penalized regression. *J. Math. Psychol.* **89**, 31–50 (2019).
116. Brown, P. J., Vannucci, M. & Fearn, T. Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. 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. *J. Econom.* **154**, 85–100 (2010).
119. Scheipl, F., Fahrmeir, L. & Kneib, T. Spike-and-slab priors for function selection in structured additive regression models. *J. Am. Stat. Assoc.* **107**, 1518–1532 (2012).
120. Tadesse, M. G., Sha, N. & Vannucci, M. Bayesian variable selection in clustering high dimensional data. *J. Am. Stat. Assoc.* https://doi.org/10.1198/016214504000001565 (2005).
121. Wang, H. Scaling it up: stochastic search structure learning in graphical models. *Bayesian Anal.* **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 (2015).
123. Li, F. & Zhang, N. R. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Am. Stat. Assoc.* **105**, 1978–2002 (2010).
124. 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. *Ann. Appl. Stat.* **5**, 1202–1214 (2011).
125. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815 (2011).

126. Bottolo, L. et al. GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genetics* **9**, e1003657–e1003657 (2013).
127. Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical Modeling and Analysis for Spatial Data* (CRC, 2014).
128. 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).
129. Penny, W. D., Trujillo-Barreto, N. J. & Friston, K. J. Bayesian fMRI time series analysis with spatial priors. *Neuroimage* **24**, 350–362 (2005).
130. Smith, M., Pütz, B., Auer, D. & Fahrmeir, L. Assessing brain activity through spatial Bayesian variable selection. *Neuroimage* **20**, 802–815 (2003).
131. 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 (2014).
132. Gorrostieta, C., Fiecas, M., Ombao, H., Burke, E. & Cramer, S. Hierarchical vector auto-regressive models and their applications to multi-subject effective connectivity. *Front. Computat. Neurosci.* **7**, 159–159 (2013).
133. 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).
134. Schad, D. J., Betancourt, M. & Vasishth, S. Toward a principled Bayesian workflow in cognitive science. Preprint at https://arxiv.org/abs/1904.12765 (2019).
135. Gelman, A., Meng, X.-L. & Stern, H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sinica* **6**, 733–760 (1996).
136. Meng, X.-L. Posterior predictive *p*-values. *Ann. Stat.* **22**, 1142–1160 (1994).
137. Asparouhov, T., Hamaker, E. L. & Muthén, B. Dynamic structural equation models. *Struct. Equ. Modeling* **25**, 359–388 (2018).
138. Zhang, Z., Hamaker, E. L. & Nesselroade, J. R. Comparisons of four methods for estimating a dynamic factor model. *Struct. Equ. Modeling* **15**, 377–402 (2008).
139. Hamaker, E., Ceulemans, E., Grasman, R. & Tuerlinckx, F. Modeling affect dynamics: state of the art and future challenges. *Emot. Rev.* **7**, 316–322 (2015).
140. Meissner, P. wikipediatrend: Public Subject Attention via Wikipedia Page View Statistics. R package version 2.1.6. *Peter Meissner* https://CRAN.R-project.org/package=wikipediatrend (2020).
141. Veen, D. & van de Schoot, R. Bayesian analysis for PhD-delay dataset. *OSF* https://doi.org/10.17605/OSF.IO/JA859 (2020).
142. Harvey, A. C. & Peters, S. Estimation procedures for structural time series models. *J. Forecast.* **9**, 89–108 (1990).
143. Taylor, S. J. & Letham, B. Forecasting at scale. *Am. Stat.* **72**, 37–45 (2018).
144. Gopnik, A. & Bonawitz, E. Bayesian models of child development. *Wiley Interdiscip. Rev. Cogn. Sci.* **6**, 75–86 (2015).
145. Gigerenzer, G. & Hoffrage, U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* **102**, 684 (1995).
146. Slovic, P. & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organ. Behav. Hum. Perform.* **6**, 649–744 (1971).
147. 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. *J. Consult. Clin. Psychol.* **80**, 54–65 (2012).
148. 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).
149. Fallesen, P. & Breen, R. Temporary life changes and the timing of divorce. *Demography* **53**, 1377–1398 (2016).
150. Hansford, T. G., Depaoli, S. & Canelo, K. S. Locating U.S. Solicitors General in the Supreme Court's policy space. *Pres. Stud. Q.* **49**, 855–869 (2019).
151. Phipps, D. J., Hagger, M. S. & Hamilton, K. Predicting limiting 'free sugar' consumption using an integrated model of health behavior. *Appetite* **150**, 104668 (2020).
152. 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 (2017).
153. Kruschke, J. K. Bayesian estimation supersedes the *t* test. *J. Exp. Psychol. Gen.* **142**, 573–603 (2013).
154. Lee, M. D. How cognitive modeling can benefit from hierarchical Bayesian models. *J. Math. Psychol.* **55**, 1–7 (2011).
155. Royle, J. & Dorazio, R. *Hierarchical Modeling and Inference in Ecology* (Academic, 2008).
156. Gimenez, O. et al. in *Modeling Demographic Processes in Marked Populations* Vol. 3 (eds Thomson D. L., Cooch E. G. & Conroy M. J.) 883–915 (Springer, 2009).
157. King, R., Morgan, B., Gimenez, O. & Brooks, S. P. *Bayesian Analysis for Population Ecology* (CRC, 2009).
158. Kéry, M. & Schaub, M. *Bayesian Population Analysis using WinBUGS: A Hierarchical Perspective* (Academic, 2011).
159. McCarthy, M. *Bayesian Methods of Ecology* 5th edn (Cambridge Univ. Press, 2012).
160. Korner-Nievergelt, F. et al. *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and Stan* (Academic, 2015).
161. Monnahan, C. C., Thorson, J. T. & Branch, T. A. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods Ecol. Evol.* **8**, 339–348 (2017).
162. Ellison, A. M. Bayesian inference in ecology. *Ecol. Lett.* **7**, 509–520 (2004).
163. 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).
164. Kuhnert, P. M., Martin, T. G. & Griffiths, S. P. A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecol. Lett.* **13**, 900–914 (2010).
165. 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. *J. R. Stat. Soc. Series C* **57**, 609–632 (2008).
166. Newman, K. et al. *Modelling Population Dynamics* (Springer, 2014).
167. Bachl, F. E., Lindgren, F., Borchers, D. L. & Illian, J. B. inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods Ecol. Evol.* **10**, 760–766 (2019).
168. 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 (2008).
169. Saunders, S. P., Cuthbert, F. J. & Zipkin, E. F. Evaluating population viability and efficacy of conservation management using integrated population models. *J. Appl. Ecol.* **55**, 1380–1392 (2018).
170. McClintock, B. T. et al. A general discrete-time modeling framework for animal movement using multistate random walks. *Ecol. Monog.* **82**, 335–349 (2012).
171. Dennis, B., Ponciano, J. M., Lele, S. R., Taper, M. L. & Staples, D. F. Estimating density dependence, process noise, and observation error. *Ecol. Monog.* **76**, 323–341 (2006).
172. Aeberhard, W. H., Mills Flemming, J. & Nielsen, A. Review of state-space models for fisheries science. *Ann. Rev. Stat. Appl.* **5**, 215–235 (2018).
173. Isaac, N. J. B. et al. Data integration for large-scale models of species distributions. *Trends Ecol Evol* **35**, 56–67 (2020).
174. McClintock, B. T. et al. Uncovering ecological state dynamics with hidden Markov models. Preprint at https://arxiv.org/abs/2002.10497 (2020).
175. King, R. Statistical ecology. *Ann. Rev. Stat. Appl.* **1**, 401–426 (2014).
176. Fearnhead, P. in *Handbook of Markov Chain Monte Carlo* Ch. 21 (eds Brooks, S., Gelman, A., Jones, G.L. & Meng, X.L.) 513–529 (Chapman & Hall/CRC, 2011).
177. Andrieu, C., Doucet, A. & Holenstein, R. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Series B* **72**, 269–342 (2010).
178. Knape, J. & de Valpine, P. Fitting complex population models by combining particle filters with Markov chain Monte Carlo. *Ecology* **93**, 256–263 (2012).
179. Finke, A., King, R., Beskos, A. & Dellaportas, P. Efficient sequential Monte Carlo algorithms for integrated population models. *J. Agric. Biol. Environ. Stat.* **24**, 204–224 (2019).
180. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet* **10**, 681–690 (2009).
181. 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. USA* **112**, E3441–3450 (2015).
182. 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 (2018).
183. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
184. Allen, N. E., Sudlow, C., Peakman, T., Collins, R. & Biobank, U. K. UK Biobank data: come and get it. *Sci. Transl. Med.* **6**, 224ed224 (2014).
185. Cortes, A. et al. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat. Genet.* **49**, 1311–1318 (2017).
186. Argelaguet, R. et al. Multi-omics factor analysis — a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
187. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
188. Yau, C. & Campbell, K. Bayesian statistical learning for big data biology. *Biophys. Rev.* **11**, 95–102 (2019).
189. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* **11**, e1004333 (2015).
190. Wang, J. et al. Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16**, 875–878 (2019).
191. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
192. National Cancer Institute. The Cancer Genome Atlas. *Qeios* https://doi.org/10.32388/e1plqh (2020).
193. Kuipers, J. et al. Mutational interactions define novel cancer subgroups. *Nat. Commun.* **9**, 4353 (2018).
194. Schwartz, R. & Schaffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18**, 213–229 (2017).
195. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-016-0021 (2017).
196. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
197. Lamprecht, A.-L. et al. Towards FAIR principles for research software. *Data Sci.* **3**, 37–59 (2020).
198. Smith, A. M., Katz, D. S. & Niemeyer, K. E. Software citation principles. *PeerJ Comput. Sci.* **2**, e86 (2016).
199. Clyburne-Sherin, A., Fei, X. & Green, S. A. Computational reproducibility via containers in psychology. *Meta Psychol.* https://doi.org/10.15626/MP.2018.892 (2019).
200. Lowenberg, D. Dryad & Zenodo: our path ahead. *WordPress* https://blog.datadryad.org/2020/03/10/dryad-zenodo-our-path-ahead/ (2020).
201. Nosek, B. A. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
202. Vehtari, A. & Ojanen, J. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **6**, 142–228 (2012).
203. Abadi, M. et al. in *USENIX Symposium on Operating Systems Design and Implementation (OSDI'16)* 265–283 (USENIX Association, 2016).
204. Paszke, A. et al. in *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) 8026–8037 (Urran Associates, 2019).
205. Kingma, D. P. & Welling, M. An introduction to variational autoencoders. Preprint at https://arxiv.org/abs/1906.02691 (2019).
This recent review of variational autoencoders encompasses deep generative models, the re-parameterization trick and current inference methods.
206. Higgins, I. et al. beta-VAE: learning basic visual concepts with a constrained variational framework. *ICLR 2017* https://openreview.net/forum?id=Sy2fzU9gl (2017).
207. Märtens, K. & Yau, C. BasisVAE:tTranslation-invariant feature-level clustering with variational autoencoders. Preprint at https://arxiv.org/abs/2003.03462 (2020).
208. Liu, Q., Allamanis, M., Brockschmidt, M. & Gaunt, A. in *Advances in Neural Information Processing Systems 31* (eds Bengio, S. et al.) 7795–7804 (Curran Associates, 2018).
209. Louizos, C., Shi, X., Schutte, K. & Welling, M. in *Advances in Neural Information Processing Systems* 8743-8754 (MIT Press, 2019).
210. Garnelo, M. et al. in *Proceedings of the 35th International Conference on Machine Learning* Vol. 80 (eds Dy, J. & Krause, A.) 1704–1713 (PMLR, 2018).
211. Kim, H. et al. Attentive neural processes. Preprint at https://arxiv.org/abs/1901.05761 (2019).

212. Rezende, D. & Mohamed, S. in *Proceedings of the 32nd International Conference on Machine Learning* Vol. 37 (eds Bach, F. & Blei, D.) 1530–1538 (PMLR, 2015).
213. Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. & Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. Preprint at https://arxiv.org/abs/1912.02762 (2019).
214. Korshunova, I. et al. in *Advances in Neural Information Processing Systems 31* (eds Bengio, S.et al.) 7190–7198 (Curran Associates, 2018).
215. Zhang, R., Li, C., Zhang, J., Chen, C. & Wilson, A. G. Cyclical stochastic gradient MCMC for Bayesian deep learning. Preprint at https://arxiv.org/abs/1902.03932 (2019).
216. Neal, R. M. *Bayesian Learning for Neural Networks* (Springer Science & Business Media, 2012).
217. Neal, R. M. in *Bayesian Learning for Neural Networks Lecture Notes in Statistics* Ch 2 (ed Nea, R. M.) 29–53 (Springer, 1996).
This classic text highlights the connection between neural networks and Gaussian processes and the application of Bayesian approaches for fitting neural networks.
218. Williams, C. K. I. in *Advances in Neural Information Processing Systems* 295–301 (MIT Press, 1997).
219. MacKay David, J. C. A practical Bayesian framework for backprop networks. *Neural. Comput.* https://doi.org/10.1162/neco.1992.4.3.448 (1992).
220. Sun, S., Zhang, G., Shi, J. & Grosse, R. Functional variational Bayesian neural networks. Preprint at https://arxiv.org/abs/1903.05779 (2019).
221. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems* **30**, 6402–6413 (2017).
222. Wilson, A. G. The case for Bayesian deep learning. Preprint at https://arxiv.org/abs/2001.10995 (2020).
223. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
224. Gal, Y. & Ghahramani, Z. in *International Conference on Machine Learning* 1050–1059 (JMLR, 2016).
225. Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732 (1995).
226. 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).
227. Liang, F. & Wong, W. H. Evolutionary Monte Carlo: applications to Cp model sampling and change point problem. *Stat. Sinica* 317-342 (2000).
228. Liu, J. S. & Chen, R. Sequential Monte Carlo methods for dynamic systems. *J. Am. Stat. Assoc.* **93**, 1032–1044 (1998).
229. Sisson, S., Fan, Y. & Beaumont, M. *Handbook of Approximate Bayesian Computation* (Chapman and Hall/CRC 2018).
230. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Series B* **71**, 319–392 (2009).
231. Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* **10**, 325–337 (2000).
232. Ntzoufras, I. *Bayesian Modeling Using WinBUGS* Vol. 698 (Wiley, 2011).
233. Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* **10**, 325–337 (2000).
This paper provides an early user-friendly and freely available black-box MCMC sampler, opening up Bayesian inference to the wider scientific community.
234. Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. OpenBUGS User Manual version 3.2.3. *Openbugs* http://www.openbugs.net/w/Manuals?action=AttachFile&do=view&target=OpenBUGS_Manual.pdf (2014).
235. Plummer, M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proc. 3rd International Workshop on Distributed Statistical Computing* **124**, 1–10 (2003).
236. Plummer, M. rjags: Bayesian graphical models using MCMC. R package version, 4(6) (2016).
237. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2**, e55 (2016).
238. de Valpine, P. et al. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Stat.s* **26**, 403–413 (2017).
239. Dillon, J. V. et al. Tensorflow distributions. Preprint at https://arxiv.org/abs/1711.10604 (2017).
240. Keydana, S. tfprobability: R interface to TensorFlow probability. *github* https://rstudio.github.io/tfprobability/index.html (2020).
241. Bingham, E. et al. Pyro: deep universal probabilistic programming. *J. Mach. Learn. Res.* **20**, 973–978 (2019).
242. Bezanson, J., Karpinski, S., Shah, V. B. & Edelman, A. Julia: a fast dynamic language for technical computing. Preprint at https://arxiv.org/abs/1209.5145 (2012).
243. Ge, H., Xu, K. & Ghahramani, Z. Turing: a language for flexible probabilistic inference. *Proceedings of Machine Learning Research* **84**, 1682–1690 (2018).
244. Smith, B. J. et al. brian-j-smith/Mamba.jl: v0.12.4. *Zenodo* https://doi.org/10.5281/zenodo.3740216 (2020).
245. JASP Team. JASP (version 0.14) [computer software] (2020).
246. Lindgren, F. & Rue, H. Bayesian spatial modelling with R-INLA. *J. Stat. Soft.* **63**, 1–25 (2015).
247. Vanhatalo, J. et al. GPstuff: Bayesian modeling with Gaussian processes. *J. Mach. Learn. Res.* **14**, 1175–1179 (2013).
248. Blaxter, L. *How to Research* (McGraw-Hill Education, 2010).
249. Neuman, W. L. *Understanding Research* (Pearson, 2016).
250. Betancourt, M. Towards a principled Bayesian workflow. *github* https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html (2020).
251. Veen, D. & van de Schoot, R. Posterior predictive checks for the Premier League. *OSF* https://doi.org/10.17605/OSF.IO/7YRUD (2020).
252. Kramer, B. & Bosman, J. Summerschool open science and scholarship 2019 — Utrecht University. *ZENODO* https://doi.org/10.5281/ZENODO.3925004 (2020).
253. Rényi, A. On a new axiomatic theory of probability. *Acta Math. Hung.* **6**, 285–335 (1955).
254. Lesaffre, E. & Lawson, A. B. *Bayesian Biostatistics* (Wiley, 2012).
255. Hoijtink, H., Beland, S. & Vermeulen, J. A. Cognitive diagnostic assessment via Bayesian evaluation of informative diagnostic hypotheses. *Psychol Methods* **19**, 21–38 (2014).

Acknowledgements
R.v.d.S. was supported by grant NWO-VIDI-452-14-006 from the Netherlands Organization for Scientific Research. R.K. was supported by Leverhulme research fellowship grant reference RF-2019-299 and by The Alan Turing Institute under the EPSRC grant EP/N510129/1. K.M. was supported by a UK Engineering and Physical Sciences Research Council Doctoral Studentship. C.Y. is supported by a UK Medical Research Council Research Grant (Ref. MR/P02646X/1) and by The Alan Turing Institute under the EPSRC grant EP/N510129/1

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

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Methods Primers thanks D. Ashby, J. Doll, D. Dunson, F. Feinberg, J. Liu, B. Rosenbaum and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

RELATED LINKS
Dryad: https://datadryad.org/
Registry of Research Data Repositories: https://www.re3data.org/
***Scientific Data* list of repositories:** https://www.nature.com/sdata/policies/repositories
Zenodo: https://zenodo.org/