doi: 10.1038/s43586-020-00003-0; Article citation ID: 2021 1:3 For the Primer, visit doi: 10.1038/s43586-020-00001-2

PrimeView
Bayesian statistics and modelling

Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' theorem. All observed and unobserved parameters in a statistical model are given a joint probability distribution. The posterior distribution then reflects one's updated knowledge, balancing prior knowledge with observed data.

Experimentation

Bayes' theorem consists of three main components: the prior distribution, or 'prior' (**c**), which expresses knowledge about the parameters in question before seeing the data; the likelihood function, or 'likelihood' (**b**), which stochastically generates all the data; and the posterior distribution, or 'posterior' (**a**), to summarize one's updated knowledge, balancing prior knowledge with observed data. The marginalization (**d**) is a normalizing factor across all outcomes of the data.

$$\overset{a}{P(A|B)} = \frac{\overset{b}{P(B|A)}\,\overset{c}{P(A)}}{\underset{d}{P(B)}}$$

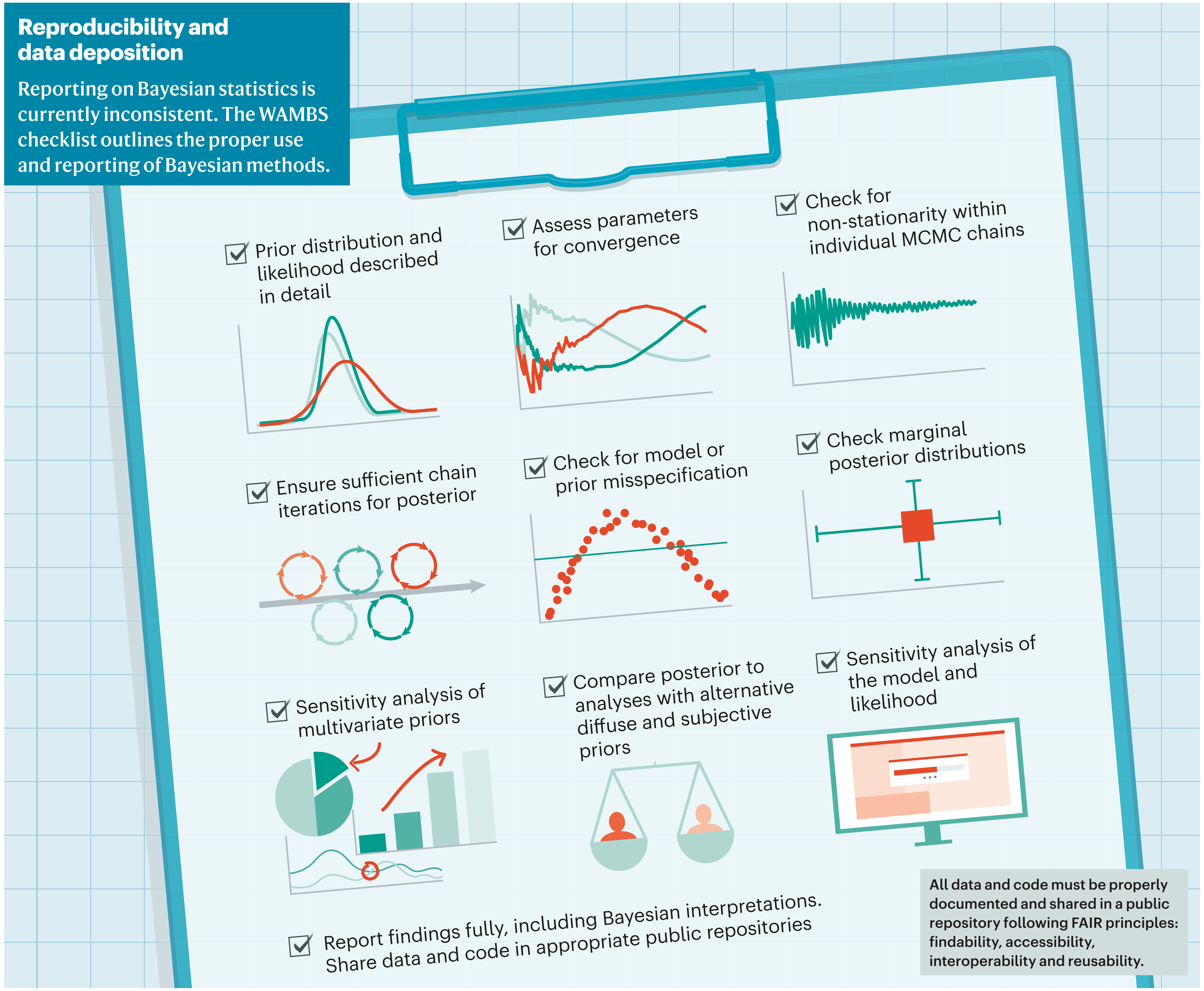
Priors can have a large impact on the final model results, and prior selection is considered one of the most important aspects of implementing Bayesian statistical analysis. Priors range in their level of certainty, from very high levels of certainty (informative priors) to low levels of certainty (diffuse priors), and their suitability must be verified using predictive checks.

Results

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. The priors and likelihood, along with the data, are combined to form the posterior distribution. The Markov chain Monte Carlo (MCMC) technique is often used used to approximate the posterior distribution by repeatedly drawing samples from it to form a distributional estimate of the posterior and its associated statistics. Posteriors can be used to extrapolate beyond the observed data in simulations and are useful for predicting future events.

Reproducibility and data deposition

Reporting on Bayesian statistics is currently inconsistent. The WAMBS checklist outlines the proper use and reporting of Bayesian methods.



☑ Prior distribution and likelihood described in detail

☑ Assess parameters for convergence

☑ Check for non-stationarity within individual MCMC chains

☑ Ensure sufficient chain iterations for posterior

☑ Check for model or prior misspecification

☑ Check marginal posterior distributions

☑ Sensitivity analysis of multivariate priors

☑ Compare posterior to analyses with alternative diffuse and subjective priors

☑ Sensitivity analysis of the model and likelihood

☑ Report findings fully, including Bayesian interpretations. Share data and code in appropriate public repositories

All data and code must be properly documented and shared in a public repository following FAIR principles: findability, accessibility, interoperability and reusability.

Limitations and optimizations

Prior and likelihood choice have a strong influence on the overall result, and in practice chosen models are never correct. Results that are indistinguishable from noise could be predicted to have a strong posterior distribution, leading to over-certainty. This limitation can be overcome by finding and fixing the problem with the model and showing that the inferences are robust to reasonable departures from the model.

Applications

Bayesian statistics and inference are versatile and have been used across scientific disciplines, from philosophy to quantum physics. Bayesian inference has been successfully used to predict behaviours based on various factors in the social sciences, from infant care-related stress and divorce rates, to limiting dietary sugar intake, to supreme court decisions and Presidential ideology. In ecology, Bayesian inference is used to predict survival, reproduction and population sizes and the influence of conservation efforts. Using Bayesian analysis in genetics and genomics is gaining popularity because of its ability to handle large datasets and make predictions on disease states using genetic data.

Outlook

A major challenge in Bayesian analysis is dealing with highly complex real-world situations, leading to larger datasets and model specification uncertainties. Advances in technology have changed how Bayesian approaches are being conducted, leading to the adoption of machine learning and artificial intelligence strategies for Bayesian inference. Deep neural networks are especially promising for model construction and in algorithms used to infer posterior distributions.

Written by Dominique Morneau; designed by Sam Whitham