Computing Bayes: Bayesian Computation from 1763 to the 21st Century*

Gael M. Martin†, David T. Frazier and Christian P. Robert

December 8, 2020

Abstract

The Bayesian statistical paradigm uses the language of probability to express uncertainty about the phenomena that generate observed data. Probability distributions thus characterize Bayesian analysis, with the rules of probability used to transform prior probability distributions for all unknowns — parameters, latent variables, models — into posterior distributions, subsequent to the observation of data. Conducting Bayesian analysis requires the evaluation of integrals in which these probability distributions appear. Bayesian computation is all about evaluating such integrals in the typical case where no analytical solution exists. This paper takes the reader on a chronological tour of Bayesian computation over the past two and a half centuries. Beginning with the one-dimensional integral first confronted by Bayes in 1763, through to recent problems in which the unknowns number in the millions, we place all computational problems into a common framework, and describe all computational methods using a common notation. The aim is to help new researchers in particular — and more generally those interested in adopting a Bayesian approach to empirical work — make sense of the plethora of computational techniques that are now on offer; understand when and why different methods are useful; and see the links that do exist, between them all.

Keywords: History of Bayesian computation; Laplace approximation; Markov chain Monte Carlo; importance sampling; approximate Bayesian computation; Bayesian synthetic likelihood; variational Bayes; integrated nested Laplace approximation.

MSC2010 Subject Classification: 62-03, 62F15, 65C60

*The first author presented a historical overview of Bayesian computation, entitled 'Computing Bayes: Bayesian Computation from 1763 to 2017!', at 'Bayes on the Beach' (Queensland, November, 2017) and Monash University (March, 2018). She would like to thank participants at both presentations for inspiring the eventual writing of this paper. Martin and Frazier have been supported by Australian Research Council Discovery Grant No. DP170100729 and the Australian Centre of Excellence in Mathematics and Statistics. Frazier has also been supported by Australian Research Council Discovery Early Career Researcher Award DE200101070. Robert has been partly supported by a senior chair (2016-2021) from l'Institut Universitaire de France and by a Prairie chair from the Agence Nationale de la Recherche (ANR-19-P3IA-0001). The authors would also like to thank the Editor, an associate editor and two anonymous reviewers for very constructive and insightful comments on an earlier draft of the paper.

†Corresponding author: gael.martin@monash.edu.

1 The Beginning

December 23 1763: London. Richard Price reads to the Royal Society a paper penned by a past Fellow, the late Thomas Bayes:

'*An Essay Towards Solving a Problem in the Doctrine of Chances.*'

With that reading, the concept of 'inverse probability' — *Bayesian inference* as we know it now — has its first public airing.

To our modern eyes, the problem tackled by Bayes in his essay is a simple one: If one performs n independent Bernoulli trials, with a probability, θ, of success on each trial, what is the probability — given n outcomes — of θ lying between two values, a and b? The answer Bayes offered is equally simple to re-cast in modern terminology. Define $Y_i|\theta \sim i.i.d.\ Bernoulli(\theta)$, $i = 1, 2, ..., n$; record the observed sequence of successes ($Y_i = 1$) and failures ($Y_i = 0$) as $\mathbf{y} = (y_1, y_2, ..., y_n)'$; denote by $p(\mathbf{y}|\theta)$ the likelihood function for θ; and invoke a Uniform prior, $p(\theta)$, on the interval $(0, 1)$. Bayes sought:

$$\mathbb{P}(a < \theta < b|\mathbf{y}) = \int_a^b p(\theta|\mathbf{y})d\theta, \tag{1}$$

where $p(\theta|\mathbf{y})$ denotes the posterior probability density function (pdf) for θ,

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}, \tag{2}$$

$p(\mathbf{y}) = \int_0^1 p(\mathbf{y}|\theta)p(\theta)d\theta$ defines the marginal likelihood, and the scale factor $[p(\mathbf{y})]^{-1}$ in (2) ensures that $p(\theta|\mathbf{y})$ integrates to one. Given the Bernoulli assumption for Y, the Uniform prior on θ, and defining $x = \Sigma_{i=1}^n y_i$, $p(\theta|\mathbf{y})$ has a closed-form representation as the Beta pdf,

$$p(\theta|\mathbf{y}) = [B(x + 1, n - x + 1)]^{-1} \theta^x (1 - \theta)^{n-x}, \tag{3}$$

where $B(x + 1, n - x + 1) = \Gamma(x + 1)\Gamma(n - x + 1)/\Gamma(n + 2) = \int_0^1 \theta^x (1 - \theta)^{n-x} d\theta$ is the Beta function, and $\Gamma(x)$ is the Gamma function. Bayesian inference — namely, quantification of uncertainty about an unknown θ, conditioned on known data, $\mathbf{y}$ — thus first emerges as the analytical solution to a particular inverse probability problem.[1]

Bayes, however, did not seek the pdf in (3) *per se*. Rather, he sought to evaluate the probability in (1) which, for either $a \neq 0$ or $b \neq 1$, involved evaluation of the incomplete Beta function. Except for the case when either x or $(n - x)$ were small, a closed-form solution to (1) eluded Bayes. Hence, along with the elegance of (2) — 'Bayes' theorem' as it is now commonly known — and the availability of the analytical expression in (3), came the need to approximate, or *compute*, the quantity of interest in (1). The quest for a computational solution to a Bayesian problem was thus born.

[1] Bayes expressed this problem in terms of its equivalent representation as one of deriving the probability of $a < \theta < b$ conditional on the value of x, where $X|\theta \sim \mathcal{B}(n, \theta)$. We have chosen to denote the conditioning values explicitly as a *sample* of n (Bernoulli) observations, $\mathbf{y}$, in order to establish the notation $p(\theta|\mathbf{y})$ from the outset. Due to the sufficiency of X, $p(\theta|x)$ is of course equivalent to $p(\theta|\mathbf{y})$. Bayes also cast this problem in physical terms: as one in which balls were rolled across a square table, or plane. Over time his pictorial representation of the problem has come to be viewed as a 'billiard table', despite Bayes making no overt reference to such an item in his essay. For this, and other historical anecdotes, see Stigler (1986a) and Fienberg (2006).

2 Bayesian Computation in a Nutshell

2.1 The Computational Problem

Bayes' probability of interest in (1) can, of course, be expressed as a posterior expectation, $\mathbb{E}(I_{[a,b]}|\mathbf{y}) = \int_\Theta \mathbb{I}_{[a,b]} p(\theta|\mathbf{y}) d\theta$, where the indicator function $\mathbb{I}_{[a,b]}$ equals 1 if $a < \theta < b$, and equals 0 otherwise. Generalizing at this point to any problem with unknown $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)' \in \Theta$ and joint posterior pdf $p(\boldsymbol{\theta}|\mathbf{y})$, most Bayesian quantities of interest are posterior expectations of some function $g(\boldsymbol{\theta})$ and, hence, can be expressed as,

$$\mathbb{E}(g(\boldsymbol{\theta})|\mathbf{y}) = \int_\Theta g(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{4}$$

In addition to posterior probabilities like that of Bayes, familiar examples include posterior moments, such as $\mathbb{E}(\boldsymbol{\theta}|\mathbf{y}) = \int_\Theta \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ and $\text{Var}(\boldsymbol{\theta}|\mathbf{y}) = \int_\Theta [\boldsymbol{\theta} - \mathbb{E}(\boldsymbol{\theta}|\mathbf{y})] [\boldsymbol{\theta} - \mathbb{E}(\boldsymbol{\theta}|\mathbf{y})]' p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$, plus marginal quantities like $p(\theta_1^*|\mathbf{y}) = \int_\Theta p(\theta_1^*|\theta_2, ..., \theta_p, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ (for θ_1^* a point in the support of $p(\theta_1|\mathbf{y})$). However, (4) also subsumes the case where $g(\boldsymbol{\theta}) = p(y_{n+1}^*|\boldsymbol{\theta}, \mathbf{y})$ (with y_{n+1}^* in the support of the 'out-of-sample' random variable, y_{n+1}), in which case (4) defines the *predictive* distribution for y_{n+1}, $p(y_{n+1}^*|\mathbf{y})$. Further, it encompasses $g(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, d)$, for $L(\boldsymbol{\theta}, d)$ the loss function associated with a decision d, in which case (4) is the quantity minimized in Bayesian decision theory (Berger, 1985; Robert, 2001). Finally, defining $g(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})$ as the likelihood function that explicitly conditions on the model, $\mathcal{M}$ say, the marginal likelihood of $\mathcal{M}$, $p(\mathbf{y}|\mathcal{M})$, is the expectation,

$$\mathbb{E}(g(\boldsymbol{\theta})|\mathcal{M}) = \int_\Theta g(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}, \tag{5}$$

with respect to the prior, $p(\boldsymbol{\theta}|\mathcal{M})$. The ratio of (5) to the comparable quantity for an alternative model defines the *Bayes factor* for use in choosing between the two models. In summary then, the key quantities that underpin all Bayesian analysis — inference, prediction, decision theory and model choice — can be expressed as expectations.

The need for numerical computation arises simply because analytical solutions to (4) and (5) are rare. Indeed, Bayes' original problem highlights that a solution to (4) can elude us *even* when the posterior pdf itself has a closed form. Typically, the posterior does not possess a closed form, as the move from the generative problem (the specification of $p(\mathbf{y}|\boldsymbol{\theta})$) to the inverse problem (the production of $p(\boldsymbol{\theta}|\mathbf{y})$), yields a posterior that is known only up to a constant of proportionality, as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}); \tag{6}$$

exceptions to this including when $p(\mathbf{y}|\boldsymbol{\theta})$ is from the exponential family, and either a natural conjugate, or convenient noninformative prior is adopted (as in Bayes' problem). The availability of $p(\boldsymbol{\theta}|\mathbf{y})$ only up to the integrating constant *immediately* precludes the analytical solution of (4), for any $g(\boldsymbol{\theta})$. By definition, a lack of knowledge of the integrating constant implies that the marginal likelihood for the model in (5) is unavailable. Situations where the likelihood function itself does not have a closed form render the analytical solution of (4) and (5) an even more distant dream. Hence the need for computational solutions.

2.2 The Computational Solutions

Despite their large number, it is useful to think about all Bayesian computational techniques falling into one or more of three broad categories:

1) Deterministic integration methods

2) Simulation methods

3) Approximation (including asymptotic) methods

Whilst all techniques are applicable to both the posterior expectation in (4) and the prior expectation in (5), we give emphasis throughout the paper to the computation of (4); reserving discussion of the computation of (5) until Section 6.

In brief, the methods in *1)* define L grid-points, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_L$, to span the support of $\boldsymbol{\theta}$, compute $g(\boldsymbol{\theta}_l)p(\boldsymbol{\theta}_l|\mathbf{y})$, for $l = 1, 2, ..., L$, and estimate (4) as a weighted sum of these L values of the integrand. Different deterministic numerical integration (or quadrature) rules are based on different choices for $\boldsymbol{\theta}_l$, $l = 1, 2, ..., L$, and different formulae for the weights. The methods in *2)* use simulation to produce M posterior draws of $g(\boldsymbol{\theta})$, $g(\boldsymbol{\theta}^{(i)})$, $i = 1, 2, ..., M$, a (weighted) mean of which is used to estimate (4). Different simulation methods are distinguished by the way in which the draws are produced and weighted. Finally, the methods in *3)* involve replacing the integrand in (4) with an approximation of some sort, and evaluating the resultant integral. Different approximation methods are defined by the choice of replacement for the integrand, with the nature of this replacement determining the way in which the final integral is computed. In particular, certain methods approximate $p(\boldsymbol{\theta}|\mathbf{y})$ itself and use simulation from this approximate posterior to evaluate the resultant integral. That is, the categories *1)* and *2)* are certainly not mutually exclusive. *Asymptotic* approximation methods replace the integrand with an expansion that is accurate for large n, and yield an estimate of (4) (via analytical means) that is accurate asymptotically.

2.3 The Aim and Scope of this Review

The aim of this review is to provide readers — in particular those new to the Bayesian paradigm — with some insights into questions like: 'Why are there so many different ways of performing Bayesian computation?', 'What are the connections between them'?, 'When does one use one approach, and when another?' and 'Are computational problems now different from computational problems of the past?'

To achieve this aim we have made two key decisions: *i)* to describe all methods using a common notation; and *ii)* to place the evolution of computational methods in a historical context. In so doing, we are able to present a coherent chronological narrative about Bayesian computation. Specifically, all methods can be seen to be, in essence, attempting to compute integrals like (4) and (5); the use of a common notation makes that clear. However, important details of those integrals have changed over time: the dimension of $\boldsymbol{\theta}$ (i.e. the number of 'unknowns'), the dimension of $\mathbf{y}$ (i.e. the 'size' of the data), and the nature of the integrand itself. Computation has evolved accordingly, and the chronological ordering helps make sense of that evolution. Hence, while we do make reference, where helpful, to the above

categories of computational methods, the over-arching structure that we adopt is one of chronology, as understanding *when* a computational method has appeared aids in the appreciation of *why*.

Excessive formalism, and extensive theoretical detail is avoided in order to make the paper as accessible as possible, in particular to researchers whose knowledge of Bayesian computation is not comprehensive. Whilst our referencing is reasonably thorough, we have still been selective; directing readers to key review papers, handbook chapters and other texts, for a more complete coverage of published work. We also defer to those other resources for descriptions of the dedicated software that is available for implementing specific computational techniques.

Importantly, to render the scope of the paper manageable, we have also had to be selective about the coverage of methods. As an overall principle, we give focus to computational developments that have ultimately been linked to the need to solve Bayesian problems. Hence, whilst deterministic numerical integration remains an important tool in the Bayesian arsenal, and despite the recent explosion of probabilistic numerics creating new connections between Bayesian concepts and numerical integration (Briol *et al.*, 2019), we make scant reference to developments in *1)* (See also Davis and Rabinowitz, 1975, Naylor and Smith, 1982, and Vanslette *et al.*, 2019, for relevant coverage). Instead, our focus is primarily on the techniques in categories *2) and 3)* that have either had their genesis within, or been transformational for, Bayesian analysis. We also consider computation only in the case of a parametric model, $p(\boldsymbol{y}|\boldsymbol{\theta})$, with a finite set of unknowns, and do not attempt to review Bayesian computation in nonparametric settings. (See Ghosal and Van der Vaart, 2017, for a thorough coverage of computation in that sphere.)

In addition, we do not cover Bayesian optimization (Gutmann and Corander, 2016; Frazier, 2018), the Bayesian bootstrap (Rubin, 1981), or the Bayesian empirical likelihood (Lazar, 2003; Chib *et al.*, 2018). Other than brief mentions made in Sections 5.2.1 and 6.1.1, we do not discuss sequential Monte Carlo (SMC) in any detail, referring the reader to Naesseth *et al.* (2019) for a recent review. The coverage of more recent developments in *Markov chain Monte Carlo* (MCMC) (in Section 5.5) is also very summary in nature. We categorize the primary goals of such developments, and defer to existing reviews for all details. Given our decision not to cover SMC, our citation of newer developments in *importance sampling* (IS) is less thorough, although IS methods that focus on estimation of the marginal likelihood are treated in Section 6.1.1. We refer to the reader to Hoogerheide *et al.* (2009) and Tokdar and Kass (2010) for discussion of more recent advances in IS.

2.4 The Structure of this Review

We begin, in Section 3, by returning to Bayes' integral in (1), briefly discussing the nature of the computational problem. We then use this as a springboard for pinpointing four particular points in time during the two centuries (or so) subsequent to 1763: 1774, 1953, 1964 and 1970. These time points correspond, in turn, to four publications — by Laplace, Metropolis *et al.,* Hammersley and Handscomb, and Hastings, respectively — in which computational methods that produce estimates of integrals like that of Bayes, were proposed. Whilst only the method of Laplace was explicitly set within the context of inverse probability (or Bayesian inference), all five methods of computing integrals can be viewed as harbingers of what was to come in Bayesian computation *per se*.

In Section 4, we look at Bayesian computation in the late 20th century, during which time the inexorable rise in the speed, and accessibility, of computers led to the pre-eminence of *simulation-based* computation. Whilst significant advances were made in econometrics (Kloek and van Dijk, 1978; Bauwens and Richard, 1985; Geweke, 1989a) and signal processing (Gordon *et al.*, 1993) using the principles of IS, the 'revolution' was driven primarily by MCMC algorithms. As many treatments of Bayesian computation have covered this period, we keep our coverage of this period very brief, deferring most details to more specialized reviews and seminal articles.

In contrast, the coverage in Section 5 — of what we term a 'second computational revolution' — is much more extensive, given that we bring together in one place, and using a common notational framework, the large variety of computational methods that have evolved during the 21st century. We begin with *pseudo-marginal* methods, including particle MCMC, before covering the main *approximate* methods: approximate Bayesian computation (ABC), Bayesian synthetic likelihood (BSL), variational Bayes (VB) and integrated nested Laplace approximation (INLA). One goal is to link the development of these new techniques to the increased complexity — and size — of the empirical problems being analyzed. A second goal is to draw out insightful links *and* differences between all and, in so doing, pinpoint when and why each technique has value. This provides some context for the *hybrid* computational methods that we then review. Section 5 is completed by a brief summary of important modifications and refinements of MCMC that have occurred since its initial appearance, including Hamiltonian up-dates, adaptive sampling, and coupling; developments that are, again, motivated by the challenges presented by modern problems, most notably, the need to process huge data sets and/or to infer high-dimensional unknowns.

We round off the review in Section 6 by switching focus from parameter inference to model choice and prediction, and to the role of computation therein. We then end the paper with Section 7, bravely entitled: 'The Future', in which we identify certain key computational challenges that remain, and the directions in which solutions to those challenges are being sought.

3 Some Early Chronological Signposts

3.1 1763: Bayes' Integral

Bayes' desire was to evaluate the probability in (1). As noted above, for either $a \neq 0$ or $b \neq 1$, this required evaluation of the incomplete Beta function. For either x or $(n-x)$ small, Bayes proposed a Binomial expansion and term-by-term integration to give an exact solution (his 'Rule 1'). However, for x and $(n-x)$ both large, this approach was infeasible: prompting Bayes (and, subsequently, Price himself; Price, 1764) to resort to producing upper and lower bounds for (1) using quadrature. Indeed, Stigler (1986a) speculates that the inability to produce an approximation to (1) that was sufficiently accurate may explain Bayes' reluctance to publish his work and, perhaps, the lack of attention it received subsequent to its (posthumous) presentation by Price in 1763 and publication the following year in Bayes (1764).[2]

[2]On November 10, 1763, Price sent an edited and annotated version of Bayes' essay to the Secretary of the Royal Society, with his own Appendix added. Price read the essay to the Society on December 23, as noted earlier. The essay and appendix were subsequently published in 1764, in the *Philosophical Transactions of the Royal Society of London*. The front matter of the issue appears here: https://royalsocietypublishing.org/cms/asset/f005dd95-c0f8-45b2-8347-0296a93c4272/front.pdf.

Whilst the integral that Bayes wished to compute was a very particular one, it was representative of the general hurdle that needed to be overcome if the principle of inverse probability were to be a useful practical tool. In brief, inference about θ was expressed in probabilistic terms and, hence, required either the direct computation of probability intervals, or the computation of distributional moments of some sort. Ironically, the choice of the Bernoulli model, possibly the simplest process for generating data 'forward' (conditional on θ) that Bayes could have assumed, exacerbated this problem, given that the 'inversion' problem does not possess the simplicity of the generative problem. What was required was a solution that was, in large measure, workable no matter what the nature of the generative model, and the first solution came via the 1774 *'Mémoire sur la probabilité des causes par les événemens'* by Pierre Laplace.

3.2 1774: Laplace and His Method of Asymptotic Approximation

Laplace envisaged an experiment in which n tickets were drawn with replacement from an urn containing a given proportion of white and black tickets. Recasting his analysis in our notation, θ is the probability of drawing a white ticket, $\mathbf{y} = (y_1, y_2, ..., y_n)'$ denotes the sequence of white tickets ($Y_i = 1$) and black tickets ($Y_i = 0$) in the n independent draws of $Y|\theta$, and $x = \Sigma_{i=1}^{n} y_i$ is the number of white tickets drawn. Laplace's aim was to show that, for arbitrary w: $\mathbb{P}(\left|\frac{x}{n} - \theta\right| < w|\mathbf{y}) = \mathbb{P}(\frac{x}{n} - w < \theta < \frac{x}{n} + w|\mathbf{y}) \to 1$ as $n \to \infty$. That is, Laplace wished to demonstrate *posterior consistency*: concentration of the posterior onto the true proportion of white tickets in the urn, $\theta_0 = \lim_{n\to\infty} \frac{x}{n}$. Along the way, however, he stumbled upon the same problem as had Bayes: computing the following probability of a Beta random variable,

$$\mathbb{P}(a < \theta < b|\mathbf{y}) = [B(x+1, n-x+1)]^{-1} \int_a^b \theta^x (1-\theta)^{n-x} d\theta, \tag{7}$$

with $a = \frac{x}{n} - w \neq 0$ and $b = \frac{x}{n} + w \neq 1$. Laplace's genius (allied with the power of asymptotics!) was to recognize that the exponential of the integrand in (7) has the bulk of its mass in the region of its mode, as n gets large, and that the integral can be computed in closed form in this case. This enabled him to prove (in modern notation) that $\mathbb{P}(|\theta_0 - \theta| > w|\mathbf{y}) = o_p(1)$, where p denotes the probability law for $\mathbf{y}$.

The route he took to this proof, however, involved approximating the Beta posterior with a Normal distribution, which (under regularity) is an approach that can be used to provide a large sample approximation of virtually *any* posterior probability. Specifically, express an arbitrary posterior probability as

$$\mathbb{P}(a < \theta < b|\mathbf{y}) = \int_a^b p(\theta|\mathbf{y}) d\theta = \int_a^b \exp\{n f(\theta)\} d\theta, \tag{8}$$

where $f(\theta) = \log[p(\theta|\mathbf{y})]/n$, and assume appropriate regularity for $p(\mathbf{y}|\theta)$ and $p(\theta)$. What is now referred to as the *Laplace asymptotic approximation* involves first taking a second-order Taylor series approximation of $f(\theta)$ around its mode, $\widehat{\theta}$: $f(\theta) \approx f(\widehat{\theta}) + \frac{1}{2} f''(\widehat{\theta})(\theta - \widehat{\theta})^2$, where $f'(\widehat{\theta}) = 0$ by construction. Defining

The publication has been reprinted since, including in Barnard and Bayes (1958), with a biographical note by G.A. Barnard. Further historical detail on the important role played by Price in the dissemination of Bayes' ideas can be found in Hooper (2013) and Stigler (2018). As the submission of Bayes' essay by Price, and his presentation to the Royal Society occurred in 1763, and Volume 53 of the *Philosophical Transactions* in which the essay appears is 'For the Year 1763', Bayes' essay is often dated 1763. We follow Stigler (1986a) in using the actual publication date of 1764.

$\sigma^2 = -[nf''(\widehat{\theta})]^{-1}$, and substituting the expansion into (8) then yields

$$\begin{aligned}\mathbb{P}(a < \theta < b|\mathbf{y}) &\approx \exp\left\{nf(\widehat{\theta})\right\} \int_a^b \exp\left\{-\tfrac{1}{2\sigma^2}(\theta - \widehat{\theta})^2\right\} d\theta \\ &= \exp\left\{nf(\widehat{\theta})\right\} \sqrt{2\pi\sigma^2} \times \{\Phi[\tfrac{b-\widehat{\theta}}{\sigma}] - \Phi[\tfrac{a-\widehat{\theta}}{\sigma}]\},\end{aligned} \tag{9}$$

where $\Phi(.)$ denotes the standard Normal cumulative distribution function (cdf).[3]

With (9), Laplace had thus devised a general way of implementing inverse probability: probabilistic statements about an unknown parameter, θ, conditional on data generated from any (regular) model, could now be made, at least up to an error of approximation. Whilst his focus was solely on the computation of a specific posterior probability, and in a single parameter setting, his method was eventually used to approximate general posterior expectations of the form in (4) (Lindley, 1980; Tierney and Kadane, 1986; Tierney *et al.*, 1989) and, indeed, applied as an integral approximation method in its own right (De Bruijn, 1961). The approach also underpins the modern INLA technique to be discussed in Section 5.3.5 (Rue *et al.*, 2009).[4]

Meanwhile, it would take 170-odd years for the *next* major advance in the computation of probability integrals to occur; an advance that would eventually transform the way in which problems in inverse probability could be tackled. This development was based on a new form of thinking and, critically, required a platform on which such thinking could operate: namely, machines that could *simulate* repeated random draws of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{y})$, or from some representation thereof. Given a sufficient number of such draws, and the correct use of them, an estimate of (4) could be produced that — unlike the Laplace approximation — would be accurate for any sample size, n, and would require less analytical input. This potential to accurately estimate (4) for essentially any problem, and any given sample size, was the catalyst for a flourishing of Bayesian inference in the late 20th century and beyond. The 1953 publication in the *Journal of Chemical Physics* by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller: *'Equation of State Calculations by Fast Computing Machines'*, was a first major step in this journey.[5]

3.3 1953: Monte Carlo Simulation and the Metropolis Algorithm

The convergence of the idea of simulating random draws from a probability distribution, and the production of such draws by computing machines, occurred in the scientific hothouse of the Los Alamos

[3] Of course, buried within the symbol '$\approx$' in (9) is a rate of convergence that is a particular order of n, and is probabilistic if randomness in $\mathbf{y}$ is acknowledged. See Tierney and Kadane (1986) and Robert and Casella (2004) for futher elaboration; and Ghosal *et al.* (1995) and van der Vaart (1998) for more formal demonstrations of the conditions under which a posterior distribution converges in probability to a Normal distribution, and the so-called Bernstein-von Mises theorem — the modern day version of Laplace's approximation — holds.

[4] Stigler (1975, Section 2) states that he has found no documentary evidence that Laplace's ideas on inverse probability, as presented in the 1774 publication, including his own statement of 'Bayes' theorem' in (2), were informed by Bayes' earlier ideas. See Stigler (1986a, Chapter 3) for discussion of Laplace's later extensions of Bayes' theorem to the case of a non-Uniform prior, and See Stigler (1975), Stigler (1986a), Stigler (1986b) and Fienberg (2006) on matters of attribution. The first recorded reference to Bayes' prior claim to inverse probability is in the preface, written by Condorcet, to Laplace's later 1781 publication: '*Mémoire sur les probabilités*'.

[5] With reference to the mechanical simulation of a random variable, we acknowledge the earlier 1870s' invention of the *quincunx* by Francis Galton. This machine used the random dispersion of metal shot to illustrate (amongst other things) draws from a hierarchical Normal model and regression to the mean. Its use can thus be viewed as the first illustration of the conjugation of a Normal likelihood and a Normal prior. See Stigler (1986a) for more details, including Galton's graphical illustration of his machine in a letter to his cousin (and Charles Darwin's son), George Darwin.

Laboratory, New Mexico, in the 1940s and 1950s; the primary impetus being the need to simulate physical processes, including neutrons in the fissile material in atomic bombs. We refer the reader to Liu (2001), Hitchcock (2003), Gubernatis (2005) and Robert and Casella (2011) for reviews of this period, including details of the various personalities who played a role therein.[6] Our focus here is simply on the nature of the problem that was at the heart of Metropolis *et al.* (1953), the solution proposed, and the ultimate importance of that solution to Bayesian computation.

In short, the authors wished to compute an expectation of the form,

$$\mathbb{E}(g(\mathbf{x})) = \int_{\mathcal{X}} g(\mathbf{x})p(\mathbf{x})d\mathbf{x}, \tag{10}$$

where $p(\mathbf{x})$ denotes the so-called Boltzmann distribution of a set, $\mathbf{x}$, of N particles on $\mathbb{R}^2$. (See Robert and Casella, 2011, Section 2.1, for all details.) Two particular characteristics of (10) are relevant to us here: *i)* the integral is of very high dimension, $2N$, with N large; and *ii)* $p(\mathbf{x})$ is generally known only up to its integrating constant. The implication of *i)* is that a basic rectangular integration method, based on L grid-points in each of the $2N$ directions, is infeasible, having a computational burden of L^{2N} or, equivalently, an approximation error of $O(L^{-1/2N})$ (Kloek and van Dijk, 1978). The implication of *ii)* is that a Monte Carlo (MC) estimate of (10), based on M *i.i.d.* direct draws from $p(\mathbf{x})$, $\mathbf{x}^{(i)}$, $i = 1, 2, ..., M$: $\widehat{E}_{MC}(g(\mathbf{x})) = \frac{1}{M}\sum_{i=1}^{M} g(\mathbf{x}^{(i)})$, with approximation error of $O(M^{-1/2})$ independent of dimension, is not available.[7]

Features *i)* and *ii)* — either individually or in tandem — broadly characterize the posterior expectations in (4) that are the focus of this review. Hence the relevance to Bayesian computation of the solution offered by Metropolis *et al.* (1953) to the non-Bayesian problem in (10); and we describe their solution with direct reference to (4) and the notation used therein.

Specifically, the authors advocate computing an integral such as (4) via the simulation of a *Markov chain:* $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, ..., M$, with *invariant distribution* $p(\boldsymbol{\theta}|\mathbf{y})$. The draw at iteration $i + 1$ in the chain is created by taking the value at the ith iteration, $\boldsymbol{\theta}^{(i)}$, and perturbing it according to a random walk: $\boldsymbol{\theta}^c = \boldsymbol{\theta}^{(i)} + \delta\boldsymbol{\varepsilon}$, where each element of $\boldsymbol{\varepsilon}$ is drawn independently from $U(-1, 1)$, and δ 'tunes' the algorithm.[8] The 'candidate' draw $\boldsymbol{\theta}^c$ is accepted as draw $\boldsymbol{\theta}^{(i+1)}$ with probability:

$$\alpha = \min\{p^*(\boldsymbol{\theta}^c|\mathbf{y})/p^*(\boldsymbol{\theta}^{(i)}|\mathbf{y}), 1\}, \tag{11}$$

where p^* is a kernel of p. Using the theory of reversible Markov chains, it can be shown (see, for example, Tierney, 1994) that use of (11) to determine the $(i + 1)th$ value in the chain does indeed produce

[6] We make particular mention here of John and Klara von Neumann, and Stanislav Ulam, with the latter co-authoring the 1949 publication in the *Journal of the Americal Statistical Association: 'The Monte Carlo Method'* with Nicholas Metropolis. We also note the controversy concerning the respective contributions of the five authors of the 1953 paper (who included two married couples). On this particular point, we refer the reader to the informative 2005 article by Gubernatis, in which Marshall Rosenbluth gives a bird's eye account of who did what, and when. The article brings to light the important roles played by both Adriana Rosenbluth and Mici Teller.

[7] The authors actually make mention of a *naive* Monte Carlo method, based on *Uniform* sampling over the $2N$ dimensional space, followed by a reweighting of the Uniform draws by a kernel of $p(\mathbf{x})$. The idea is dismissed, however, as 'not practical'. In modern parlance, whilst this method would yield an $O(M^{-1/2})$ approximation error, the constant term within the order would be large, since the Uniform distribution used to produce draws of $\mathbf{x}$ differs substantially from the *actual* distribution of $\mathbf{x}$, $p(\mathbf{x})$.

[8] Metropolis *et al.* (1953) actually implemented their algorithm one element of $\boldsymbol{\theta}$ at a time, as a harbinger of the Gibbs sampler to come. See Robert and Casella (2011) for more details.

a dependent sequence of draws with invariant distribution $p(\boldsymbol{\theta}|\mathbf{y})$. Hence, subject to convergence to $p(\boldsymbol{\theta}|\mathbf{y})$ (conditions for which were verified by the authors for their particular problem) these draws can be be used to estimate (4) as the sample mean,

$$\overline{g(\boldsymbol{\theta})} = \frac{1}{M}\sum_{i=1}^{M} g(\boldsymbol{\theta}^{(i)}), \tag{12}$$

and an appropriate weak law of large numbers (WLLN) and central limit theorem (CLT) invoked to prove the $\sqrt{M}$-consistency and limiting normality of the estimator. (See Geyer, 2011b, for details.)

Due to the (positive) autocorrelation in the Markov chain, the variance of the *Metropolis estimator* (as it would become known) is larger than that of the (infeasible) MC estimate in (11), computed as in (12), but using $i.i.\dot{d}$ draws from $p(\boldsymbol{\theta}|\mathbf{y})$, namely:

$$\sigma^2_{MC} = \mathrm{Var}(g(\boldsymbol{\theta}))/M, \tag{13}$$

expressed here for the case of scalar $g(\boldsymbol{\theta})$. However, as is clear from (11), the Metropolis MCMC algorithm requires knowledge of $p(\boldsymbol{\theta}|\mathbf{y})$ only up to the normalizing constant, and does *not* require direct simulation from $p(\boldsymbol{\theta}|\mathbf{y})$ itself. It is this particular feature that would lend the technique its great power in the decades to come.[9]

3.4 1964: Hammersley and Handscomb: Importance Sampling

The obviation of the need to directly sample from $p(\boldsymbol{\theta}|\mathbf{y})$ also characterizes importance sampling, and underlies its eventual importance in solving difficult Bayesian computational problems. Nevertheless, Hammersley and Handscomb (1964) did not emphasize this feature but, rather, introduced the concept of IS for the express purpose of variance reduction in simulation-based estimation of integrals.[10] Again, the focus was not on Bayesian integrals, but we describe the method in that setting.

In brief, given an 'importance' (or 'proposal') density, $q(\boldsymbol{\theta}|\mathbf{y})$, that preferably mimics $p(\boldsymbol{\theta}|\mathbf{y})$ well, and M *i.i.d.* draws, $\boldsymbol{\theta}^{(i)}$, from $q(\boldsymbol{\theta}|\mathbf{y})$, an IS estimate of (4) is $\overline{g(\boldsymbol{\theta})}_{IS} = \frac{1}{M}\sum_{i=1}^{M} g(\boldsymbol{\theta}^{(i)})w(\boldsymbol{\theta}^{(i)})$, where $w(\boldsymbol{\theta}^{(i)}) = p(\boldsymbol{\theta}^{(i)}|\mathbf{y})/q(\boldsymbol{\theta}^{(i)}|\mathbf{y})$. In the typical case where $p(\boldsymbol{\theta}^{(i)}|\mathbf{y})$ is available only up to the integrating constant, and $w(\boldsymbol{\theta}^{(i)})$ cannot be evaluated as a consequence, the estimate is modified as

$$\overline{g(\boldsymbol{\theta})}_{IS} = \sum_{i=1}^{M} g(\boldsymbol{\theta}^{(i)})w(\boldsymbol{\theta}^{(i)}) \Big/ \sum_{j=1}^{M} w(\boldsymbol{\theta}^{(i)}), \tag{14}$$

with the weights re-defined as $w(\boldsymbol{\theta}^{(j)}) = p^*(\boldsymbol{\theta}^{(i)}|\mathbf{y})/q^*(\boldsymbol{\theta}^{(i)}|\mathbf{y})$, for kernels, $p^*(\boldsymbol{\theta}^{(i)}|\mathbf{y})$ and $q^*(\boldsymbol{\theta}^{(i)}|\mathbf{y})$, of $p(\boldsymbol{\theta}|\mathbf{y})$ and $q(\boldsymbol{\theta}|\mathbf{y})$ respectively. Once again, and under regularity conditions pertaining to the *importance density* $q(\boldsymbol{\theta}|\mathbf{y})$, asymptotic theory can be invoked to prove that (14) is a $\sqrt{M}$-consistent estimator of $\mathbb{E}(g(\boldsymbol{\theta})|\mathbf{y})$ (Geweke, 1989a). A judicious choice of $q(\boldsymbol{\theta}|\mathbf{y})$ is able to yield a sampling variance that is less than (13) in some cases, as befits the original motivation of IS as a variance reduction method. (See

[9] Dongarra and Sullivan (2000) rank the Metropolis algorithm as one of the 10 algorithms "with the greatest influence on the development and practice of science and engineering in the 20th century".

[10] One could in fact argue that a similar aim motivated Metropolis and co-authors, given that they drew a sharp contrast (in effect) between the efficiency of their method and that of the naive Monte Carlo technique based on Uniform sampling.

Geweke, 1989a, and Robert and Casella, 2004, for discussion.) Critically however, like the Metropolis method, (14) serves as a feasible estimate of $\mathbb{E}(g(\boldsymbol{\theta})|\mathbf{y})$ when $p(\boldsymbol{\theta}|\mathbf{y})$ cannot be easily simulated; hence the significance of IS in Bayesian computation. Moreover, its maintenance of independent draws, allied with its re-weighting of draws from an approximating density, has led to the emergence of IS as a vehicle for implementing SMC algorithms, like particle filtering, to be referenced in Section 5.2 in the context of particle MCMC.

3.5 1970: Hastings and his Generalization of the Metropolis Algorithm

The final publication that we pinpoint during the 200-odd year period subsequent to 1763, is the 1970 *Biometrika* paper, '*Monte Carlo Sampling Methods Using Markov Chains and Their Applications',* by Wilfred Keith Hastings. Whilst Metropolis *et al.* (1953) proposed the use of MCMC sampling to compute particular integrals in statistical mechanics, it was the Hastings paper that elevated the concept to a general one, and introduced it to the broader statistics community. Included in the paper is also the first mention of what would become known as the *Metropolis-within-Gibbs sampler* (Robert and Casella, 2011). Once again, the author's focus was not a Bayesian integral *per se*; however we describe the method in that context.

In contrast to Metropolis and co-authors, Hastings (1970) acknowledges up-front that the need to know $p(\boldsymbol{\theta}|\mathbf{y})$ only up to the integrating constant is a compelling feature of an MCMC-based estimate of (4). Hastings also generalizes the acceptance probability in (11) to one that accommodates a general 'candidate' distribution $q(\boldsymbol{\theta}|\mathbf{y})$ from which $\boldsymbol{\theta}^c$ is drawn, as:

$$\alpha = \min\left\{\left[p^*(\boldsymbol{\theta}^c|\mathbf{y})/q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^c,\mathbf{y})\right] \Big/ \left[p^*(\boldsymbol{\theta}^{(i)}|\mathbf{y})/q(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(i)},\mathbf{y})\right], 1\right\}, \tag{15}$$

which clearly collapses to (11) when $q(\boldsymbol{\theta}|\mathbf{y})$ is symmetric (in $\boldsymbol{\theta}^c$ and $\boldsymbol{\theta}^{(i)}$), as in the original random walk proposal of Metropolis *et al.* (1953). Importantly, the more general algorithm allows for a targeted choice of $q(\boldsymbol{\theta}|\mathbf{y})$ that reduces the need for tuning and which can, potentially, reduce the degree of dependence in the chain and, hence, the variance of the estimate of $\mathbb{E}(g(\boldsymbol{\theta})|\mathbf{y})$. Hastings formalizes the standard error of this estimate using time series theory, explicitly linking, for the first time, the autocorrelation in the Markov draws to the efficiency of the MCMC-based estimate of (4). Crucially, the author tackles the issue of dimension by advocating the treatment of one element (or several elements) of $\boldsymbol{\theta}$ at a time, conditional on all remaining elements.

In summary, all of the important ingredients from which the huge smorgasbord of future MCMC algorithms would eventually be constructed — for the *express* purpose of solving Bayesian problems — were now on the table, via this particular paper.

4 The Late 20th Century: Gibbs Sampling & the MCMC Revolution

Whilst the role that *could* be played by simulation in computation was thus known by the 1970s, the computing technology needed to exploit that knowledge lagged behind.[11] Over the next two decades,

[11] Many readers may be too young to remember the punchcards! But there was a time when RAND's 1955 *A Million Random Digits with 100,000 Normal Deviates* was more than an entry for sarcastic Amazon comments, as producing this million digits took more than two months at the time.

however, things changed. Indeed, *two* developments now went hand in hand to spawn a remarkable expansion in simulation-based Bayesian computation: *i)* the increased speed and availability of computers, including personal desktop computers (Ceruzzi, 2003), and *ii)* the collective recognition that MCMC draws from a joint posterior, $p(\boldsymbol{\theta}|\mathbf{y})$, could be produced via iterative sampling from lower dimensional, and often standard, *conditional* posteriors. When allied with both the concept of *augmentation,* and an understanding of the theoretical properties of combinations of MCMC algorithms, *ii)* would lead to *Gibbs sampling* (with or without *Metropolis-Hastings* (MH) subchains) becoming the work-horse of Bayesian computation in the 1990s.

An MH algorithm 'works', in the sense of producing a Markov chain that converges to the required distribution $p(\boldsymbol{\theta}|\mathbf{y})$, due to the form of the acceptance probability in (15) (or the nested version in (11)). More formally, the algorithm, as based on candidate density $q(\boldsymbol{\theta}|\mathbf{y})$, and acceptance probability as defined in (15), defines a *transition kernel* with invariant distribution, $p(\boldsymbol{\theta}|\mathbf{y})$. The 'Gibbs sampler' similarly yields a Markov chain with invariant distribution, $p(\boldsymbol{\theta}|\mathbf{y})$, but via a transition kernel that is defined as the product of full conditional posteriors associated with the joint. For the simplest case of a two-dimensional vector $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, the steps of the Gibbs algorithm are as follows: first, specify an initial value for θ_2, $\theta_2^{(0)}$; second, for $i = 1, 2, ..., M$, cycle iteratively through the two conditional distributions, drawing respectively: $\theta_1^{(i)}$ from $p_1(\theta_1^{(i)}|\theta_2^{(i-1)}, \mathbf{y})$, and $\theta_2^{(i)}$ from $p_2(\theta_2^{(i)}|\theta_1^{(i)}, \mathbf{y})$. Given the satisfaction of the required convergence conditions (which essentially place sufficient regularity on the conditionals), the draws $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)})'$, $i = 1, 2, ..., M$, converge in distribution to the joint posterior distribution as $M \to \infty$, and can be used to produce a $\sqrt{M}$-consistent estimator of (4) in the form of (12). Extension to higher-dimensional problems is obvious, although decisions about how to 'block' the parameters, and thereby define the conditionals, now play a role (Roberts and Sahu, 1997).[12]

Gibbs thus exploits the simplicity yielded by conditioning: whilst joint and marginal posterior distributions are usually complex in form, (full) conditional posteriors are often standard and, hence, able to be simulated from directly. While one may find hints in both Hastings (1970) and Besag (1974), this point was first made clearly by Geman and Geman (1984), who also coined the phrase 'Gibbs sampling' because their problem used Gibbs random fields in image restoration (named, in turn, after the physicist, Josiah Willard Gibbs). However, the later paper by Gelfand and Smith (1990) is generally credited with bringing this transformational idea to the attention of the broader statistical community, and illustrating its broad applicability.

The idea of Gibbs sampling overlapped with a related proposal by Tanner and Wong (1987): that of 'augmenting' the set of unknowns ($\boldsymbol{\theta}$ in our notation) with latent data, $\mathbf{z} = (z_1, z_2, ..., z_n)'$, to yield conditionals — $p(\boldsymbol{\theta}|\mathbf{z}, \mathbf{y})$ and $p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$ — that facilitate the production of a simulation-based estimate of $p(\boldsymbol{\theta}|\mathbf{y})$; with $p(\boldsymbol{\theta}|\mathbf{z}, \mathbf{y})$, in particular, often being standard. The melding of these two ideas, i.e. sampling via conditionals *per se*, and yielding more tractable conditionals through the process of augmentation, enabled the analysis of complex models that had thus far eluded Bayesian treatment, due to their de-

[12]The Gibbs sampler can be viewed as a special case of a 'multiple-block' MH sampler, in which the candidate values for each block of parameters are drawn directly from their full conditional distributions and the acceptance probability in (each blocked version of) (15) is equal to one. (See, for example, Chib, 2011). See also Tran (2018) for further discussion of this point — conducted in the context of a generalized MH framework, in which many of the algorithms to be discussed in Sections 5.2 and 5.5 are also nested.

pendence on high-dimensional vectors of latent variables; selected examples being: Polson *et al.* (1992), Carter and Kohn (1994), Frühwirth-Schnatter (1994) and Jacquier *et al.* (1994). However, it also led to the realization that *artificial* latent variables could be judiciously introduced into a model for the sole purpose of producing tractable conditional posteriors over the augmented space, thereby opening up a whole range of additional models to a Gibbs-based solution (e.g. Albert and Chib, 1993; Diebolt and Robert, 1994; Higdon, 1998; Kim *et al.*, 1998; Damien *et al.*, 1999). The *slice sampler* (Roberts and Rosenthal, 1999; Neal, 2003) is one particularly notable, and generic, way of generating an MCMC algorithm via this principle of auxiliary variable augmentation.

Of course, in most high-dimensional problems — and in particular those in which latent variables feature — certain conditionals remain nonstandard, such that direct simulation from them is not possible. Critically though, the reduced dimension renders this a simpler problem than sampling from the joint itself: via either the inverse cumulative distribution function technique (Devroye, 1986) — approximated in the 'Griddy Gibbs' algorithm of Ritter and Tanner (1992) — or by embedding an MH algorithm within the outer Gibbs loop (a so-called 'Metropolis-within-Gibbs' algorithm).[13]

5 The 21st Century: A Second Computational Revolution

5.1 Why Did We Need a Second One?

The advent of accessible, fast computers in the last two decades of the 20th century, allied with the methodological and theoretical developments referenced above, led to an explosion in the use of simulation-based Bayesian computation, with variants of MCMC leading the charge. The impact of these developments was felt across a huge array of fields — genetics, biology, neuroscience, astrophysics, image analysis, ecology, epidemiology, education, economics, marketing and finance, to name but some — and brought Bayesian analysis into the statistical mainstream. The two 2011 Handbooks: *Handbook of Markov Chain Monte Carlo* (Brooks *et al.*, 2011) and *The Oxford Handbook of Bayesian Econometrics* (Geweke *et al.*, 2011), further highlight the wide spectrum of fields, and broad scope of empirical problems to which MCMC algorithms were (and continue to be) applied; as do certain contributions to the series of vignettes edited by Mengersen and Robert for *Statistical Science* (2014, Vol 29, No. 1), under the theme of 'Big Bayes Stories'.

Despite their unquestioned power and versatility however, these original simulation techniques did have certain limitations; with these limitations to become more marked as the empirical problems being tackled became more ambitious; and this despite a concurrent rise in computing power (parallel computing, access to GPUs etc.) over recent decades. With reference to the posterior pdf in (6), two characteristics are worthy of note. First, as already highlighted, in all but the most stylized problems $p(\boldsymbol{\theta}|\mathbf{y})$ is available only up to its integrating constant, and cannot be directly simulated. Second, representation of $p(\boldsymbol{\theta}|\mathbf{y})$ only as a kernel, $p^*(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, still requires closed forms for $p(\mathbf{y}|\boldsymbol{\theta})$ and

[13] We refer the reader to: Besag and Green (1993), Smith and Roberts (1993) and Chib and Greenberg (1996) for early reviews of MCMC sampling; Casella and George (1992) and Chib and Greenberg (1995) for descriptions of the Gibbs and MH algorithms (respectively) that are useful for practitioners; Robert (2015), Betancourt (2018) and Dunson and Johndrow (2019) for more recent reviews; and Andrieu *et al.* (2004) and Robert and Casella (2011) for historical accounts of MCMC sampling.

$p(\boldsymbol{\theta})$. With reference to $p(\mathbf{y}|\boldsymbol{\theta})$, this means that, for any $\boldsymbol{\theta}$, $p(\mathbf{y}|\boldsymbol{\theta})$ needs to be able to be evaluated at the observed $\mathbf{y}$. The MCMC and IS simulation methods obviate the first problem by drawing *indirectly* from $p(\boldsymbol{\theta}|\mathbf{y})$ via another distribution (or set of distributions) from which simulation is feasible. However, these methods still require evaluation of $p(\mathbf{y}|\boldsymbol{\theta})$: in the computation of the importance weights in (14), in the computation of the acceptance probability in (11) or (15), and in the implementation of any Gibbs-based algorithm, in which the conditional posteriors are required either in full form or at least up to a scale factor.

The assumption that $p(\mathbf{y}|\boldsymbol{\theta})$ can be evaluated is a limitation for two reasons. First, some DGPs do not admit pdfs in closed form; examples being: probability distributions defined by quantile or generating functions (Devroye, 1986; Peters *et al.*, 2012), continuous time models in finance with unknown transition densities (Gallant and Tauchen, 1996), dynamic equilibrium models in economics (Calvet and Czellar, 2015), certain deep learning models in machine learning (Goodfellow *et al.*, 2014); complex astrophysical models (Jennings and Madigan, 2017); and DGPs for which the normalizing constant is unavailable, such as Markov random fields in spatial modelling (Rue and Held, 2005; Stoehr, 2017). Second, pointwise evaluation of $p(\mathbf{y}|\boldsymbol{\theta})$ (at any $\boldsymbol{\theta}$) (in the case where $p(\cdot|\boldsymbol{\theta})$ has a closed form) entails an $O(n)$ computational burden; meaning that the MCMC and IS methods described above are *not scalable* to so-called 'big (or tall) data' problems (Bardenet *et al.*, 2017).

Just as important are the challenges that arise when the dimension of the unknowns themselves is very large (the so-called 'high-dimensional' problem); for instance, when a model contains a very large number of latent variables over which integration is required (e.g. Tavaré *et al.*, 1997; Braun and McAuliffe, 2010; Beaumont, 2010; Lintusaari *et al.*, 2017; Johndrow *et al.*, 2019). In such cases, standard MCMC methods — even if feasible in principle — may not (as highlighted further in Section 5.5) enable an accurate estimate of (4) to be produced in finite computing time; i.e. such methods are *not necessarily scalable* in the dimension of the unknowns.

Each of the techniques discussed in Sections 5.2, 5.3 and 5.4 relieves the investigator of one or more of these 'burdens'; although, as we shall see, the relief is not costless. In particular, the approximation methods covered in Section 5.3 and 5.4, whilst enabling some problems to be tackled that would be intractable for MCMC and IS, do not produce an estimate of (4) that is *exact* up to simulation error but, instead, produce an estimate that is only ever *approximate*.

Finally, we provide a very brief overview in Section 5.5 of further advances made in MCMC methods *per se*, primarily over the past decade or so; advances designed, in large measure, to improve the accuracy with which these dependent chains estimate posterior quantities of interest such as (4), most notably in large-scale problems.

5.2 Pseudo-Marginal Methods

5.2.1 The basic idea

Referencing the concept of data augmentation introduced in Section 4: given draws from the joint posterior of $\boldsymbol{\theta}$ and $\mathbf{z}$, $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$, the draws of $\boldsymbol{\theta}$ can be used to produce an exact simulation-based estimate of $p(\boldsymbol{\theta}|\mathbf{y})$ or any associated quantity, for large enough M. Again, the latent states, $\mathbf{z}$, may be either intrinsic to the model, or introduced 'artificially' as a computational device, as highlighted therein. Initially,

draws of $(\boldsymbol{\theta}, \mathbf{z})$ were produced via Gibbs-based MCMC schemes, with a variety of MH algorithms (based on alternative candidates) used to sample from $p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$, in the typical case where this conditional could not be simulated directly (see Fearnhead, 2011, and Giordani *et al.*, 2011, for reviews). As highlighted in Section 5.1 however, depending on the problem, such MH-within-Gibbs schemes can be slow to explore the joint space of $(\boldsymbol{\theta}, \mathbf{z})$ and, hence, to produce an accurate estimate of $p(\boldsymbol{\theta}|\mathbf{y})$.

The (combined) insight of Beaumont (2003) and Andrieu and Roberts (2009) was to recognize that draws of $\mathbf{z}$ could be used in a potentially more effective way to yield an estimate of the 'marginal' of interest: $p(\boldsymbol{\theta}|\mathbf{y})$ (and any integral of the form of (4)). Key to this insight is the following observation. Use $\mathbf{u} \in \mathcal{U}$ to denote all of the canonical (problem-specific) random variables that are used to generate $\mathbf{z}$. If draws of $\mathbf{u}$ can be used to produce an *unbiased* estimate of the likelihood function, $p(\mathbf{y}|\boldsymbol{\theta})$, then an MCMC scheme applied to the joint space $(\boldsymbol{\theta}, \mathbf{u})$, can target the required invariant distribution, $p(\boldsymbol{\theta}|\mathbf{y})$. An informal demonstration of this result is straightforward. Define $h(\mathbf{u})$ as the distribution of the random variables underpinning the generation of $\mathbf{z}$ (independently of the prior $p(\boldsymbol{\theta})$), and let $h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ denote an estimate of the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, that is unbiased in the sense that $E_{\mathbf{u}}[h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})] = p(\mathbf{y}|\boldsymbol{\theta})$. Then we have that $h(\boldsymbol{\theta}|\mathbf{y}) \propto \int_{\mathcal{U}} h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}) p(\boldsymbol{\theta}) h(\mathbf{u}) d\mathbf{u} = p(\boldsymbol{\theta}) E_{\mathbf{u}}[h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})] = p(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{y})$.

Use of $h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ within an MH algorithm amounts to replacing the acceptance probability in (15) with

$$\alpha = \min \left\{ \left[h(\mathbf{y}|\boldsymbol{\theta}^c, \mathbf{u}^c) p(\boldsymbol{\theta}^c) / q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^c, \mathbf{y}) \right] Big/ \left[h(\mathbf{y}|\boldsymbol{\theta}^{(i)}, \mathbf{u}^{(i)}) p(\boldsymbol{\theta}^{(i)}) / q(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(i)}, \mathbf{y}) \right], 1 \right\}, \qquad (16)$$

where $\boldsymbol{\theta}^c$ is proposed from $q(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(i)}, \mathbf{y})$ and $\mathbf{u}^{(i)}$ and $\mathbf{u}^c$ are independent draws from $h(\mathbf{u})$. The use of an estimate of the likelihood in (16) prompted use of the term 'pseudo'-marginal MH (PMMH) by Andrieu and Roberts (2009) although, as noted, this replacement still yields a chain with an invariant distribution equal to the correct marginal, $p(\boldsymbol{\theta}|\mathbf{y})$, when the estimate is unbiased.

When a likelihood estimate is produced specifically via the use of *particle filtering* in a state space model (SSM), the term *particle* MCMC (PMCMC) has also been coined (Andrieu *et al.*, 2011).[14] Whilst we omit details of the use of filtering to estimate a likelihood function (see reviews in Doucet *et al.*, 2001, and Giordani *et al.*, 2011), we do remark that particle filtering does involve the sequential application of IS, with independent, but differentially weighted draws of the latent states (from both 'filtered' and 'prediction' state distributions) being the outcome. As such, much of the early work on IS, including the impact of proposal choice on the efficiency of the sampler, has assumed a renewed importance in filtering-based settings, including PMCMC.

Whilst unbiasedness of the likelihood estimate is required for a general PMMH algorithm to 'work', the variance of the estimate also affects the performance of the sampler and, hence, the simulation efficiency of any estimate of (4) that is produced. However, improving the precision of the likelihood estimator by increasing the number of draws of $\mathbf{u}$ used in its production comes at a computational cost, and an 'optimal' number of draws that balances computational cost with an acceptable mixing of the chain needs to be sought. See Pitt *et al.* (2012), Doucet *et al.* (2015) and Deligiannidis *et al.* (2018) for discussion of the optimal structuring and tuning of pseudo-marginal algorithms.

[14] Whilst not a pseudo-marginal method, particle filtering has also been used to provide an estimate of $p(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{y})$ in a Gibbs scheme for an SSM — so-called 'particle Gibbs' (Andrieu *et al.*, 2011).

5.2.2 The benefits

The benefits of pseudo-marginal schemes are three-fold. First, in cases where the parameters and latent states are strongly correlated, use of a PMMH scheme rather than a Gibbs-based scheme (based on $p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$ and $p(\boldsymbol{\theta}|\mathbf{z}, \mathbf{y})$), may reap efficiency gains (conditional on appropriate 'tuning' choices, as flagged above). Linked to this, avoidance of the need to sample from $p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$ obviates the need to make choices regarding the blocking of $\mathbf{z}$ and the proposal densities for those blocks. Second, in cases where *only* forward simulation of the latent process is possible, and point-wise evaluation of $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$ is infeasible as a result, PMMH remains possible. For example, in an SSM an estimate of $p(\mathbf{y}|\boldsymbol{\theta})$ can be based on the bootstrap particle filter, for which only simulation from the transition density $p(z_t|z_{t-1}, \boldsymbol{\theta})$ (not evaluation thereof) is required. Third, in cases where the dimension of $\mathbf{y}$ is very large, an unbiased estimate of $p(\mathbf{y}|\boldsymbol{\theta})$ based on appropriately selected *subsamples* of data can be used to produce a valid PMMH scheme, at a much smaller computational cost than any scheme that requires full evaluation of $p(\mathbf{y}|\boldsymbol{\theta})$ (Bardenet *et al.*, 2017; Quiroz *et al.*, 2018b; Quiroz *et al.*, 2019).

5.3 Approximate Bayesian Inference

The goal of all simulation-based computational methods discussed thus far, including the pseudo-marginal techniques, has been to estimate the posterior expectation in (4) 'exactly', at least up to an order $O(M^{-1/2})$, where M is the number of draws that defines the simulation scheme. The alternative methods do, of course, differ one from the other in terms of the constant term that quantifies the precise error of approximation. Hence, it may be the case that even for a very large M, a nominally 'exact' method (despite being 'tuned' optimally) has an approximation error that is non-negligible; a point that we revisit in Section 5.5**.** Nevertheless, the convention in the literature is to refer to all simulation methods outlined to this point as *exact*, typically without qualification.[15]

In contrast, when applying an *approximation* method (using the taxonomy from Section 2.3), investigators make no claim to exactness, other than citing the asymptotic (in n) accuracy of the Laplace methods (in Sections 3.2 and 5.3.5), or the asymptotic validity of certain other approximations (Fearnhead, 2018). That is, for finite n at least, such methods are only ever acknowledged as providing an approximation to (4), with that approximation perhaps claimed to be as 'accurate' as possible, given the relevant choice variables that characterize the method; but not more.

So what benefits do such techniques offer, in return for sacrificing the goal of exact inference? With reference to the methods discussed below: ABC and BSL both completely obviate the need to evaluate $p(\mathbf{y}|\boldsymbol{\theta})$ and, in so doing, open up to Bayesian treatment a swathe of empirical problems — so-called *doubly-intractable* problems — that would otherwise not be amenable to Bayesian analysis. In computing (4), both methods replace the posterior in the integrand, $p(\boldsymbol{\theta}|\mathbf{y})$, with an approximation produced via simulation. A simulation-based estimate of the integral, $\overline{g(\boldsymbol{\theta})} = (1/M)\sum_{i=1}^{M} g(\boldsymbol{\theta}^{(i)})$, is then produced using draws, $\boldsymbol{\theta}^{(i)}$, from this approximate posterior.[16] In contrast, VB and INLA both require evaluation

[15] We note that we have omitted any mention herein of so-called 'quasi-Monte Carlo' integration schemes, which aim for exactness at a faster rate than $O(M^{-1/2})$. See Lemieux (2009) for a review of such methods, Chen *et al.* (2011) for the extension to quasi-MCMC algorithms, and Gerber and Chopin (2015) for an entry on sequential quasi-Monte Carlo.

[16] It can also be argued that these are simulation methods for an exact, albeit different posterior, using either a degraded version of the observations or a projection of them via a nonsufficient statistic (Wilkinson, 2013).

of $p(\mathbf{y}|\boldsymbol{\theta})$, but reap computational benefits in certain types of problems (in particular those of high-dimension) by replacing — at least in part — *simulation* with (in some cases closed-form) *optimization*. In the case of VB, the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ used to define (4) is replaced by an approximation produced via the calculus of variations. Depending on the nature of the problem, including the 'variational family' from which the 'optimal' approximation is produced, the integral is computed in either closed form or via a simulation step. With INLA, the approximation of $p(\boldsymbol{\theta}|\mathbf{y})$ is chosen in such a way that (4) can be computed with the aid of low-dimensional deterministic integration. Taking a wider perspective, these approximate methods may also be considered as novel approaches to inference and, hence, evaluated as such, rather than being viewed *solely* from the vantage point of computation.

5.3.1 Approximate Bayesian computation (ABC)

From its initial beginnings as a practical approach for inference in population genetics models with intractable likelihoods (Tavaré *et al.*, 1997; Pritchard *et al.*, 1999), ABC has grown in popularity and is now commonly applied in numerous fields; its broad applicability highlighted by the more than 11,000 citations garnered on Google Scholar since 2000. As such, not only do several reviews of the area exist (Marin *et al.*, 2011; Sisson and Fan, 2011), but the technique has recently reached 'handbook status', with the publication of Sisson *et al.* (2019); and it is to those resources that we refer the reader for extensive details on the method, application and theory of ABC. We provide only the essence of the approach here, including its connection to other computational methods.

The aim of ABC is to approximate $p(\boldsymbol{\theta}|\mathbf{y})$ in cases where — despite the complexity of the problem preventing the *evaluation* of $p(\mathbf{y}|\boldsymbol{\theta})$ — $p(\mathbf{y}|\boldsymbol{\theta})$ (and $p(\boldsymbol{\theta})$) can still be *simulated*. The simplest (accept/reject) form of the algorithm proceeds as follows: first, simulate $\boldsymbol{\theta}^i$, $i = 1, 2, ..., M$, from $p(\boldsymbol{\theta})$, and artificial data $\mathbf{x}^i$ from $p(\cdot|\boldsymbol{\theta}^i)$; second, use $\mathbf{x}^i$ to construct (a vector of) simulated summary statistics $\eta(\mathbf{x}^i)$, which is then compared against the (vector of) *observed* statistics $\eta(\mathbf{y})$ using a distance $d\{\cdot, \cdot\}$; third, retain all values of $\boldsymbol{\theta}^i$ that yield simulated statistics, $\eta(\mathbf{x}^i)$, for which $d\{\eta(\mathbf{x}^i), \eta(\mathbf{y})\} \leq \varepsilon$, for some small tolerance ε.

ABC thus produces draws of $\boldsymbol{\theta}$ from a posterior that conditions not on the full data set $\mathbf{y}$, but on statistics $\eta(\mathbf{y})$ (with dimension less than n) that summarize the key characteristics of $\mathbf{y}$. Only if $\eta(\mathbf{y})$ are sufficient for conducting inference on $\boldsymbol{\theta}$, and for $\varepsilon \to 0$, does ABC provide draws from the exact posterior $p(\boldsymbol{\theta}|\mathbf{y})$. In practice, the complexity of the models to which ABC is applied implies — almost by definition — that a low-dimensional set of sufficient statistics is unavailable, and the implementation of the method (in finite computing time) requires a non-zero value for ε, and a given number of draws, M. As such, draws from the ABC algorithm provide (via kernel density methods) a simulation-based approximation of $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$, which we denote by $\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$.

The difference between $\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ and the unattainable $p(\boldsymbol{\theta}|\mathbf{y})$ has two components: the difference between the 'partial' posterior, $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$, and $p(\boldsymbol{\theta}|\mathbf{y})$, and the difference between $\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ and $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$. The first difference is the critical one, and depends on the informativeness, or otherwise, of the chosen summaries; loosely speaking, the 'closer' is $\eta(\mathbf{y})$ to being sufficient for $\boldsymbol{\theta}$, the 'closer' is $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$ to $p(\boldsymbol{\theta}|\mathbf{y})$. Attention has been given to maximizing the information content of the summaries in some sense (e.g. Joyce and Marjoram, 2008; Blum, 2010; Fearnhead and Prangle, 2012). This includes the

idea of defining $\eta(\mathbf{y})$ as (some function of) the maximum likelihood estimator (MLE) of the parameter vector of an approximating 'auxiliary' model; thereby producing summaries that are — via the properties of the MLE — close to being *asymptotically* sufficient, depending on the accuracy of the approximating model (Drovandi *et al.*, 2011; Drovandi *et al.*, 2015; Martin *et al.*, 2019). This approach mimics, in the Bayesian setting, the frequentist methods of indirect inference (Gouriéroux *et al.*, 1993) and efficient method of moments (Gallant and Tauchen, 1996) using, as it does, an approximating model to produce feasible inference about an intractable true model. Whilst the price paid for the approximation in the frequentist case is reduced sampling efficiency, in the Bayesian case the cost is posterior inference that is conditioned on insufficient summaries, and is 'partial' inference as a consequence.

Regarding the second difference, at its simplest level: the smaller is ε and the larger is M, the more accurate will $\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ be as a kernel density estimator of $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$, for any given choice of $\eta(\mathbf{y})$, with the dimension of both $\eta(\mathbf{y})$ and $\boldsymbol{\theta}$ affecting accuracy (Blum *et al.*, 2013; Frazier *et al.*, 2018; Nott *et al.*, 2018). For given M (and, hence, a given computational burden), modifications of the basic accept/reject algorithm that improve the accuracy with which $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$ is estimated by $\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ have been proposed which, variously, involve post-sampling corrections of the draws (Beaumont *et al.*, 2002; Blum, 2010), the insertion of MCMC and/or SMC steps (Marjoram *et al.*, 2003; Sisson *et al.*, 2007; Beaumont *et al.*, 2009), or the use of randomized quasi-Monte Carlo, rather than (standard) Monte Carlo in the simulation of $\boldsymbol{\theta}^i$, $i = 1, 2, ..., M$ (Buchholz and Chopin, 2019).

Before concluding this section, we note that recent work has begun to explore the use of ABC methods that do not rely on summary statistics but, instead, match empirical measures calculated from the observed and simulated data using appropriate metrics. In particular, we highlight the work of Bernton *et al.* (2019) based on the Wasserstein distance, but note that similar approaches using alternative distances have also been proposed (e.g., Frazier, 2020, and Nguyen *et al.*, 2020). In the case of a scalar θ, and if the data is also scalar valued, *Wasserstein ABC* amounts to selecting draws of θ based on matching all n order statistics in the observed sample $\mathbf{y}$. While the use of the Wasserstein distance within ABC alleviates the need to choose summaries, it is unclear exactly when this approach can be expected to deliver superior inference to that obtained using summary statistics. Indeed, the general performance of distance-based ABC methods relative to methods based on very informative summaries is still a matter requiring further investigation.

5.3.2 Bayesian synthetic likelihood (BSL)

(Summary statistic-based) ABC thus targets $p(\boldsymbol{\theta}|\eta(\mathbf{y})) \propto p(\eta(\mathbf{y})|\boldsymbol{\theta})p(\boldsymbol{\theta})$, with $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$ itself, for non-sufficient $\eta(\mathbf{y})$, being an approximate representation of $p(\boldsymbol{\theta}|\mathbf{y})$. It is clear then that, embedded within the simplest accept/reject ABC algorithm, based on a tolerance ε, is a likelihood function of the form,

$$p_\varepsilon(\eta(\mathbf{y})|\boldsymbol{\theta}) = \int_{\mathcal{X}} p(\mathbf{x}|\boldsymbol{\theta}) \mathbb{I}\left(d\{\eta(\mathbf{y}), \eta(\mathbf{x})\} \leq \varepsilon\right) d\mathbf{x}. \tag{17}$$

For a given draw $\boldsymbol{\theta}^i$, and associated $\eta(\mathbf{x}^i)$, (17) is approximated by its simulation counterpart, $\widehat{p}_\varepsilon(\eta(\mathbf{y})|\boldsymbol{\theta}^i) = \mathbb{I}\left(d\{\eta(\mathbf{y}), \eta(\mathbf{x}^i)\} \leq \varepsilon\right)$, which can implicitly be viewed as a nonparametric estimator, based on a Uniform kernel, for the quantity of interest $p_\varepsilon(\eta(\mathbf{y})|\boldsymbol{\theta})$. Following Andrieu and Roberts (2009), and as illustrated in detail by Bornn *et al.* (2017), $\widehat{p}_\varepsilon(\eta(\mathbf{y})|\boldsymbol{\theta}^i)$ can serve as a likelihood estimate within

a form of pseudo-marginal MCMC scheme (referred to as ABC-MCMC by the authors) for sampling from $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y})) \propto p_\varepsilon(\eta(\mathbf{y})|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where in this context we take 'pseudo-marginal MCMC' to mean an MCMC scheme that replaces the intractable likelihood, $p_\varepsilon(\eta(\mathbf{y})|\boldsymbol{\theta})$, within the MH ratio by an unbiased estimator, $\widehat{p}_\varepsilon(\eta(\mathbf{y})|\boldsymbol{\theta}^i)$. (See also Marjoram *et al.*, 2003.) However, in contrast with other results in the pseudo-marginal literature, Bornn *et al.* (2017) demonstrate that the efficiency of the MCMC chain so produced is not necessarily improved by using more than one draw of $\eta(\mathbf{x}^i)$ for a given draw $\boldsymbol{\theta}^i$.

Bayesian synthetic likelihood (BSL) (Price *et al.*, 2018) also targets a posterior for $\boldsymbol{\theta}$ that conditions on $\eta(\mathbf{y})$, and requires only simulation from $p(\mathbf{y}|\boldsymbol{\theta})$ (not its evaluation) in so doing. However, in contrast to the nonparametric likelihood estimate that is implicit in ABC, BSL (building on Wood, 2010) adopts a Gaussian parametric approximation to $p(\eta(\mathbf{y})|\boldsymbol{\theta})$,

$$p_a(\eta(\mathbf{y})|\boldsymbol{\theta}) = \mathcal{N}\left[\eta(\mathbf{y}); \mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta})\right], \ \mu(\boldsymbol{\theta}) = \mathbb{E}[\eta(\mathbf{y})], \ \Sigma(\boldsymbol{\theta}) = \mathrm{Var}\left[\eta(\mathbf{y})\right]. \tag{18}$$

Use of this parametric kernel leads to the *ideal* BSL posterior,

$$p_a(\boldsymbol{\theta}|\eta(\mathbf{y})) \propto p_a(\eta(\mathbf{y})|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{19}$$

where the subscript 'a' highlights that (19) is still an approximation to $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$, due to the Gaussian approximation, $p_a(\eta(\mathbf{y})|\boldsymbol{\theta})$, of $p(\eta(\mathbf{y})|\boldsymbol{\theta})$.

In general, however, the mean and variance-covariance matrix of $\eta(\mathbf{y})$ are unknown and must be estimated via simulation. Given $\mathbf{x}_j \sim i.i.d.\ p(\cdot|\boldsymbol{\theta})$, $j = 1, \ldots, m$, we can estimate $\mu(\boldsymbol{\theta})$ and $\Sigma(\boldsymbol{\theta})$ in (18) via their empirical Monte Carlo averages, $\mu_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^m \eta(\mathbf{x}_j)$ and $\Sigma_m(\boldsymbol{\theta}) = \frac{1}{m-1}\sum_{j=1}^m \left(\eta(\mathbf{x}_j) - \mu_m(\boldsymbol{\theta})\right)\left(\eta(\mathbf{x}_j) - \mu_m(\boldsymbol{\theta})\right)'$, and thereby define

$$p_{a,m}(\eta(\mathbf{y})|\boldsymbol{\theta}) = \int_{\mathcal{X}} \mathcal{N}\left[\eta(\mathbf{y}); \mu_m(\boldsymbol{\theta}), \Sigma_m(\boldsymbol{\theta})\right] \prod_{j=1}^m p(\eta(\mathbf{x}_j)|\boldsymbol{\theta}) d\mathbf{x}_1 \ldots d\mathbf{x}_m, \tag{20}$$

and the associated *target* BSL posterior,

$$p_{a,m}(\boldsymbol{\theta}|\eta(\mathbf{y})) \propto p_{a,m}(\eta(\mathbf{y})|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{21}$$

Note that, even for a single draw $\eta(\mathbf{x}_j)$, $\mathbf{x}_j \sim p(\cdot|\boldsymbol{\theta})$, we have that $\mathcal{N}\left[\eta(\mathbf{y}); \mu_m(\boldsymbol{\theta}), \Sigma_m(\boldsymbol{\theta})\right]$ is an unbiased estimate of (20). Hence, with $p_{a,m}(\boldsymbol{\theta}|\eta(\mathbf{y}))$ then accessed via an MCMC algorithm, and with arguments in Drovandi *et al.* (2015) used to show that $p_{a,m} \to p_a$ as $m \to \infty$, BSL can yield a form of pseudo-marginal MCMC method.

5.3.3 ABC versus BSL

ABC and BSL both invoke two forms of approximation: one, replacement of the full dataset $\mathbf{y}$ by the summaries $\eta(\mathbf{y})$; two, approximation of the likelihood $p(\eta(\mathbf{y})|\boldsymbol{\theta})$. It is the difference in the second form of approximation that distinguishes the two approaches and leads to potential differences in performance. It is helpful to characterize these differences in terms of *i)* asymptotic (in n) behaviour and *ii)* computational efficiency. This then enables us to provide some guidelines as to when, and why, one might use one method over the other. We consider *i)* and *ii)* in turn.

i) As ABC has evolved into a common approach to inference, attention has turned to its asymptotic validation. This work demonstrates that, under certain conditions on $\eta(\mathbf{y})$, ε and M, the ABC posterior $\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ concentrates onto the true vector $\boldsymbol{\theta}_0$ (i.e. is Bayesian consistent), satisfies a Bernstein von Mises (BvM) theorem (i.e. is asymptotically Gaussian with credible sets that have the correct level of frequentist asymptotic coverage) and yields an ABC posterior mean with an asymptotically Gaussian sampling distribution. (See Frazier *et al.*, 2018, for this full suite of results, and Li and Fearnhead, 2018a, Li and Fearnhead, 2018b, and Frazier *et al.*, 2020, for related work.) Moreover, the conditions on $\eta(\mathbf{y})$ under which these results are valid are surprisingly weak, requiring only the existence of at least a polynomial moment (uniformly in the parameter space). In addition, we note that the ABC posterior mean is asymptotically as efficient as the maximum likelihood estimator based on $p(\eta(\mathbf{y})|\boldsymbol{\theta})$.

The required conditions on the tolerance, ε, for these results to be in evidence can be ordered in terms of the speed with which $\epsilon \to 0$ as $n \to \infty$: stronger results, such as a valid BvM, require faster rates of decay for ε than weaker results, such as posterior concentration. Such a taxonomy is important since the chosen tolerance ε largely determines the computational effort required for $\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ to be an accurate estimate of $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$. Broadly speaking, the smaller is ε, the smaller is $|\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y})) - p(\boldsymbol{\theta}|\eta(\mathbf{y}))|$. However, a smaller choice of ε requires a larger number of simulations (i.e., a larger value of M) and, hence, a greater computational burden. For instance, if we wish for credible sets obtained by $\widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ to be valid in the frequentest sense, M is required to diverge faster than $n^{\dim(\eta)/2}$ (Corollary 1 in Frazier *et al.*, 2018).

In contrast to ABC, as BSL is based on the Gaussian parametric approximation to the likelihood $p(\eta(\mathbf{y})|\boldsymbol{\theta})$, it does not require any choice of tolerance. However, in order for the BSL posterior $p_{a,m}(\boldsymbol{\theta}|\eta(\mathbf{y}))$ to be a reasonable approximation to $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$, the Gaussian approximation must be reasonable. More specifically, the summaries $\eta(\mathbf{y})$ and $\eta(\mathbf{z})$ themselves must satisfy a CLT (uniformly in the case of the latter) (see Frazier *et al.*, 2019b for details), and the variance of the summaries must be consistently estimated by $\Sigma_m(\boldsymbol{\theta})$ for some value of $\boldsymbol{\theta}$, as m (the number of data sets drawn for a given draw of $\boldsymbol{\theta}$) increases. If, moreover, we wish $p_{a,m}(\boldsymbol{\theta}|\eta(\mathbf{y}))$ to deliver asymptotically correct frequentest coverage, additional conditions on the summaries and m are required. In particular, Frazier *et al.* (2019a) demonstrate that if the summaries exhibit an exponential moment, then correct uncertainty quantification is achieved so long as $m/\log(n) \to \infty$.

Under the restrictions delineated above for $\eta(\mathbf{y})$, ε, M and m, the results of Frazier *et al.* (2018) and Frazier *et al.* (2019b) can then be used to deduce that the ABC and BSL posteriors are asymptotically equivalent, in the sense that $\int |p_{a,m}(\boldsymbol{\theta}|\eta(\mathbf{y})) - \widehat{p}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))| d\boldsymbol{\theta} \xrightarrow{p} 0$ as $n \to \infty$. That is, in large samples, and under regularity, we could expect the results obtained by both methods to be comparable. However, the above discussion makes plain that BSL requires much stronger conditions on the summaries than does ABC to produce equivalent asymptotic behaviour. Hence, in the case of summaries that have thick tails, non-Gaussian features, or non-standard rates of convergence, ABC would seem to be the better choice.

ii) Once computational efficiency is taken into account however, the comparison between the two methods becomes more nuanced. Frazier *et al.* (2019b) use theoretical arguments to compare the computational efficiency of BSL and accept/reject ABC, and demonstrate that BSL does not pay the same

penalty for summary statistic dimension as does ABC. In particular, the BSL acceptance probability is asymptotically non-vanishing, and does not depend on the dimension of the summaries, neither of which is true for accept/reject ABC, even under an optimal choice for M. Given this, when the summaries are approximately Gaussian, BSL is likely to be more computationally efficient than standard ABC.[17] That being said, if one considers post-sampling corrections to ABC (such as those cited in Section 5.3.1), or the newer ABC-Gibbs schemes (Section 5.4), whose specific goal is to mitigate the curse of dimensionality associated with the summaries, then these "corrected" ABC approaches are likely to be more efficient or comparable to standard BSL on computational grounds. What is still undocumented is the relative performance of corrected ABC methods and the more computationally efficient varieties of BSL, such as the shrinkage approaches of Ong *et al.* (2018b) or the whitening approach of Priddle *et al.* (2019).

5.3.4 Variational Bayes (VB)

The two approximate methods discussed thus far, ABC and BSL, target an approximation of the posterior that is (in a standard application of the methods) conditioned on a vector of low-dimensional summary statistics. As such, and most particularly when $\eta(\mathbf{y})$ is not sufficient for $\boldsymbol{\theta}$, these methods do not directly target the exact posterior $p(\boldsymbol{\theta}|\mathbf{y})$, nor any expectation, (4), defined with respect to it. In contrast, VB methods are a general class of algorithms that produce an approximation to the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ — and hence (4) — *directly*, by replacing simulation with optimization.

The idea of VB is to search for the best approximation to the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ over a class of densities $\mathcal{Q}$, referred to as the variational family, and where $q(\boldsymbol{\theta})$ indexes elements in $\mathcal{Q}$. The most common approach to VB is to find the best approximation to the exact posterior, in the class $\mathcal{Q}$, by minimizing the KL divergence between $q(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta}|\mathbf{y})$, which defines such a density as the solution to the following optimal optimization problem,

$$q^*(\boldsymbol{\theta}) := \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}\left[q(\boldsymbol{\theta})|p(\boldsymbol{\theta}|\mathbf{y})\right], \tag{22}$$

where

$$\mathrm{KL}\left[q(\boldsymbol{\theta})|p(\boldsymbol{\theta}|\mathbf{y})\right] = \int \log(q(\boldsymbol{\theta}))q(\boldsymbol{\theta})d\boldsymbol{\theta} - \int \log(p(\boldsymbol{\theta}|\mathbf{y}))q(\boldsymbol{\theta})d\boldsymbol{\theta} \equiv \mathbb{E}_q[\log(q(\boldsymbol{\theta}))] - \mathbb{E}_q[\log(p(\boldsymbol{\theta},\mathbf{y}))] + \log(p(\mathbf{y})). \tag{23}$$

Of course, the normalizing constant $p(\mathbf{y})$ is, in all but most simple problems (for which VB would not be required!), unknown; and the quantity in (23) inaccessible as a result. Rather, the approach adopted is to define the so-called evidence lower bound (ELBO),

$$\mathrm{ELBO}[q(\boldsymbol{\theta})] := \mathbb{E}_q[\log(p(\boldsymbol{\theta},\mathbf{y}))] - \mathbb{E}_q[\log(q(\boldsymbol{\theta}))], \tag{24}$$

where $\mathrm{KL}[q(\boldsymbol{\theta})|p(\boldsymbol{\theta}|\mathbf{y})]$ is equivalent to $-\mathrm{ELBO}[q(\boldsymbol{\theta})]$ up to the unknown constant, $\log(p(\mathbf{y}))$, with the latter not dependent on $q(\boldsymbol{\theta})$. Hence, we can obtain the variational density by solving an optimization problem that is equivalent to that in (22):

$$q^*(\boldsymbol{\theta}) := \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\max}\, \mathrm{ELBO}[q(\boldsymbol{\theta})]. \tag{25}$$

[17] BSL can often be implemented using the random walk MH algorithm, and often with minimal tuning required in practice (Price *et al.*, 2018). See also Frazier and Drovandi (2019) for a slice sampling approach to sampling the BSL posterior.

The beauty of VB is that, for *certain* problems, including certain choices of the class $\mathcal{Q}$, the optimization problem in (25) can either yield a closed-form solution, or be solved relatively quickly with various numerical algorithms; (see Ormerod and Wand, 2010, Blei *et al.*, 2017, and Zhang *et al.*, 2018, for reviews). Most importantly, given that — by design — the variational family is defined in terms of standard forms of distributions, replacement of $p(\boldsymbol{\theta}|\boldsymbol{y})$ by $q^*(\boldsymbol{\theta})$ in (4) yields an expectation that is either available in closed form, or amenable to a relatively simple simulation-based solution.

VB *truly* shines in cases where $\boldsymbol{\theta}$ is high-dimensional, and an efficient MCMC algorithm may well be simply out of reach. Indeed, in such cases, the family $\mathcal{Q}$ can be chosen in such a way that the resulting posterior approximations remain tractable even when the dimension of $\boldsymbol{\theta}$ is in the thousands, or the tens of thousands (Braun and McAuliffe, 2010; Kabisa *et al.*, 2016; Wand, 2017; Koop and Korobilis, 2018).

Finally, the link between (23) and (24) makes it clear that maximizing (24) to yield $q^*(\boldsymbol{\theta})$ produces, as a by-product, a lower bound on the logarithm of the 'evidence', or marginal likelihood, $p(\mathbf{y})$. Hence, ELBO$[q^*(\boldsymbol{\theta})]$ serves as an estimate of the quantity which, as noted in Section 2.1 and discussed further in Section 6.1, underpins model choice.

Recently, several authors have analyzed the asymptotic properties of VB methods; see, for example, Wang and Blei (2019a,b) and Zhang and Gao (2017). The most complete treatment can be found in Zhang and Gao, wherein the authors demonstrate that the rate at which the VB posterior concentrates is bounded above by the following two components: *i)* the concentration rate of the exact posterior, and *ii)* the approximation error incurred by the chosen variational family. This novel decomposition highlights the fundamental importance of the variational family that is used to approximate the posterior, something that is not present in other results on the asymptotic behavior of VB. Interestingly, while Zhang and Gao deliver a convenient upper bound in a general context, they also demonstrate that in specific examples, such as Gaussian sequence models and sparse linear regression models, the VB posterior can display concentration rates that are actually faster than those obtained by the exact posterior, owing to the fact that VB performs a type of 'internal regularization' as a consequence of the algorithm's optimization step.[18]

5.3.5 Integrated nested Laplace approximation (INLA)

We complete our review of 21st century approximate computational methods with a reminder of a computational innovation from the 18th! In 1774, in a quest to illustrate posterior consistency, Laplace produced an asymptotic (in n) approximation to a particular posterior probability. Not only did this result represent the first step in the development of Bayesian asymptotic *theory*, it also provided a simple practical solution to the *computation* of general posterior expectations like that in (4). In 1986, Tierney and Kadane revived and formalized the Laplace approximation: using it to yield an asymptotic approximation (of a given order) of any posterior expectation of the form of (4), including (in the multiple parameter case) marginal posterior densities. (See also Tierney *et al.*, 1989.) Two decades later, Rue *et al.* (2009) took the method further: adapting it to approximate marginal posteriors (and

[18] Huggins *et al.* (2019) propose a method for validating the accuracy of VB posterior approximations using alternative (nonasymptotic) principles. See also Yu *et al.* (2019) (and earlier references therein) for practical validation approaches that are relevant to approximate posteriors in general.

general expectations like those in (4)) in latent Gaussian models (LGMs). With the authors using a series of *nested* Laplace approximations, allied with low-dimensional numerical *integration*, they termed their method *integrated nested Laplace approximation,* or INLA for short. Since the LGM class encompasses a large range of empirically relevant models — including, generalized linear models, non-Gaussian state space (or hidden Markov) models, and spatial, or spatio-temporal models — a computational method tailored-made for such a setting is sufficiently broad in its applicability to warrant detailed consideration herein. In common with VB, and as follows from the use of Laplace approximations evaluated at modal values, INLA eschews simulation for optimization (in addition to using low-dimensional deterministic integration methods).

Deferring to Rue *et al.* (2009), Martino and Riebler (2019) and Wood (2019) for all details (including of the LGM structure), we provide here the *key* steps of INLA. To adhere to our goal of notational consistency, we reference the p-dimensional vector of static parameters in the LGM as $\boldsymbol{\theta}$, the n-dimensional vector of latent, random parameters as $\mathbf{z}$, and the n-dimensional vector of observed data as $\mathbf{y}$, and express the model as:

$$\mathbf{y}|\mathbf{z},\boldsymbol{\theta} \sim \prod_{i=1}^{n} p(y_i|z_i,\boldsymbol{\theta}) \qquad \mathbf{z}|\boldsymbol{\theta} \sim \mathcal{N}(0, Q^{-1}(\boldsymbol{\theta})) \qquad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}),$$

where $Q(\boldsymbol{\theta})$ is the precision matrix of the latent Gaussian field, assumed — for computational feasibility — to be sparse. Whilst the LGM allows for each element of $\mathbf{y}$ to be of dimension $d \geq 1$, we keep our description simple by assuming $d = 1$. The goal of the authors is to approximate the marginal posteriors; $p(\theta_j|\mathbf{y})$, $j = 1, 2, .., p$, and $p(z_i|\mathbf{y})$, $i = 1, 2, .., n$. The problems envisaged are those in which p is small and n is large (potentially in the order of millions), with MCMC algorithms deemed to be possibly unsuitable as a consequence, due to the scale of both $\mathbf{z}$ and $\boldsymbol{y}$.

Beginning with the expression of $p(\boldsymbol{\theta}|\mathbf{y})$ as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{z},\boldsymbol{\theta}|\mathbf{y})}{p(\mathbf{z}|\boldsymbol{\theta},\mathbf{y})} \propto \frac{p(\mathbf{z},\boldsymbol{\theta},\mathbf{y})}{p(\mathbf{z}|\boldsymbol{\theta},\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{z}|\boldsymbol{\theta},\mathbf{y})}, \tag{26}$$

and recognizing that the proportionality sign arises due to the usual lack of integrating constant (over $\mathbf{z}$ and $\boldsymbol{\theta}$), the steps of the algorithm (in its simplest form) are as follows. First, on the assumption that all components of the model can be evaluated and, hence, that the numerator is available, $p(\boldsymbol{\theta}|\mathbf{y})$ in (26) is approximated as

$$\widetilde{p}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{p(\mathbf{y}|\widehat{\mathbf{z}}(\boldsymbol{\theta}),\boldsymbol{\theta})p(\widehat{\mathbf{z}}(\boldsymbol{\theta})|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p_G(\widehat{\mathbf{z}}(\boldsymbol{\theta})|\boldsymbol{\theta},\mathbf{y})}. \tag{27}$$

The denominator in (27) represents a Gaussian approximation of $p(\mathbf{z}|\boldsymbol{\theta},\mathbf{y})$, $p_G(\mathbf{z}|\boldsymbol{\theta},\mathbf{y}) = \mathcal{N}(\widehat{\mathbf{z}}(\boldsymbol{\theta}), \widehat{\Sigma}(\boldsymbol{\theta}))$, evaluated at the mode, $\widehat{\mathbf{z}}(\boldsymbol{\theta})$, of $p(\mathbf{z},\boldsymbol{\theta},\mathbf{y})$ (at a given value of $\boldsymbol{\theta}$), where $\widehat{\Sigma}(\boldsymbol{\theta})$ is the inverse of the Hessian of $-\log p(\mathbf{z},\boldsymbol{\theta},\mathbf{y})$ with respect to $\mathbf{z}$, also evaluated at $\widehat{\mathbf{z}}(\boldsymbol{\theta})$. The expression in (27) can obviously be further simplified to

$$\widetilde{p}(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\widehat{\mathbf{z}}(\boldsymbol{\theta}),\boldsymbol{\theta})p(\widehat{\mathbf{z}}(\boldsymbol{\theta})|\boldsymbol{\theta})p(\boldsymbol{\theta}) \left|\widehat{\Sigma}(\boldsymbol{\theta})\right|^{1/2}. \tag{28}$$

With appropriate adjustments made for notation, and noting that the expression is given up to the integrating constant only, (28) can be seen to be identical to the Laplace approximation of a marginal density in Tierney and Kadane (1986, equation (4.1)). (Rue *et al.*, 2009 discuss the circumstances in which the order of approximation proven in Tierney and Kadane, 1986 applies to the LGM setting.)

With marginal posterior for the ith element of $\mathbf{z}$ defined as

$$\widetilde{p}(z_i|\mathbf{y}) = \int_{\Theta} \widetilde{p}(z_i|\boldsymbol{\theta},\mathbf{y})\widetilde{p}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \tag{29}$$

a second application of a Laplace approximation would yield

$$\widetilde{p}(z_i|\boldsymbol{\theta},\mathbf{y}) \propto p(\mathbf{y}|\widehat{\mathbf{z}}_{-i}(\boldsymbol{\theta},z_i),\boldsymbol{\theta})p(\widehat{\mathbf{z}}_{-i}(\boldsymbol{\theta},z_i)|\boldsymbol{\theta})p(\boldsymbol{\theta})\left|\widehat{\Sigma}_{-i}(\boldsymbol{\theta},z_i)\right|^{1/2}, \tag{30}$$

where $\widehat{\mathbf{z}}_{-i}(\boldsymbol{\theta},z_i)$ is the mode of $p(\mathbf{z}_{-i},z_i,\boldsymbol{\theta},\mathbf{y})$ (at given values of $\boldsymbol{\theta}$ and z_i, with $\mathbf{z}_{-i}$ denoting all elements of $\mathbf{z}$ other than the ith); and where $\widehat{\Sigma}_{-i}(\boldsymbol{\theta},z_i)$ is the inverse of the Hessian of $-\log p(\mathbf{z}_{-i},z_i,\boldsymbol{\theta},\mathbf{y})$ with respect to $\mathbf{z}_{-i}$, also evaluated at $\widehat{\mathbf{z}}_{-i}(\boldsymbol{\theta},z_i)$. Computation of (30) for each z_i would, however, involve n optimizations (over $\mathbf{z}_{-i}$) plus n specifications of the high-dimensional matrix $\widehat{\Sigma}_{-i}(\boldsymbol{\theta},z_i)$. Rue *et al.* (2009) avoid this computational burden by modifying the approximation in (30) in a number of alternative ways, all details of which are provided in the references cited above. Once a representation of $\widetilde{p}(z_i|\boldsymbol{\theta},\mathbf{y})$ is produced, (29) is computed using a deterministic numerical integration scheme defined over a grid of values for the low-dimensional $\boldsymbol{\theta}$.

Defining the marginal posterior for the jth element of $\boldsymbol{\theta}$ as $\widetilde{p}(\theta_j|\mathbf{y}) = \int_{\Theta_{-j}} \widetilde{p}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}$, where $\boldsymbol{\theta}_{-j}$ denotes all elements of $\boldsymbol{\theta}$ excluding θ_j, this integral is computed using deterministic integration over $\boldsymbol{\theta}_{-j}$. Finally, if required, the marginal likelihood, $p(\mathbf{y})$ can be approximated by computing the normalizing constant in (28), $\int_{\Theta} p(\mathbf{y}|\widehat{\mathbf{z}}(\boldsymbol{\theta}),\boldsymbol{\theta})p(\widehat{\mathbf{z}}(\boldsymbol{\theta})|\boldsymbol{\theta})p(\boldsymbol{\theta})\left|\widehat{\Sigma}(\boldsymbol{\theta})\right|^{1/2} d\boldsymbol{\theta}$, using deterministic integration over $\boldsymbol{\theta}$.[19]

5.4 Hybrid Methods

We remind the reader at this point of the following: *i)* whilst ABC and BSL are advantageous when $p(\mathbf{y}|\boldsymbol{\theta})$ cannot be evaluated, a large dimension for $\boldsymbol{\theta}$ (and, hence, for $\eta(\mathbf{y})$) causes challenges (albeit to differing degrees) for both; *ii)* VB is much better equipped to deal with high-dimensional $\boldsymbol{\theta}$, but requires the evaluation of $p(\boldsymbol{\theta},\mathbf{y})$ and, thus, $p(\mathbf{y}|\boldsymbol{\theta})$; *iii)* pseudo-marginal MCMC circumvents a challenging evaluation of $p(\mathbf{y}|\boldsymbol{\theta})$ by using an unbiased estimate, $h(\mathbf{y}|\boldsymbol{\theta},\mathbf{u})$, as defined in Section 5.2.1. Recently, hybrid algorithms that meld aspects of all four methods — ABC, BSL, VB and pseudo-marginal MCMC — have been used to deal with settings in which the likelihood is intractable *and* $\boldsymbol{\theta}$ is high-dimensional.

Tran *et al.* (2017) devise a VB method for use when the likelihood is intractable, coining the technique 'VBIL'. To appreciate the principles of the method, consider that the variational approximation is indexed by a finite dimensional parameter λ, so that $\mathcal{Q} := \{\lambda \in \Lambda : q_\lambda\}$. The variational approximation is then obtained by maximizing the ELBO, $\mathcal{L}(\lambda) := \text{ELBO}[q_\lambda(\boldsymbol{\theta})]$, over Λ. VBIL replaces the intractable likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ with an estimator $h(\mathbf{y}|\boldsymbol{\theta},\mathbf{u})$, such that $E_{\mathbf{u}}[h(\mathbf{y}|\boldsymbol{\theta},\mathbf{u})] = p(\mathbf{y}|\boldsymbol{\theta})$, and considers as target distribution the joint posterior

$$h(\boldsymbol{\theta},z|\mathbf{y}) \propto \pi(\boldsymbol{\theta})h(\mathbf{y}|\boldsymbol{\theta},\mathbf{u})\exp(z)g(z|\boldsymbol{\theta}), \text{ where } z := \log h(\mathbf{y}|\boldsymbol{\theta},\mathbf{u}) - \log p(\mathbf{y}|\boldsymbol{\theta}),$$

and where $g(z|\boldsymbol{\theta})$ denotes the distribution of $z|\boldsymbol{\theta}$. Given that $h(\mathbf{y}|\boldsymbol{\theta},\mathbf{u})$ is, by construction, an unbiased estimator of $p(\mathbf{y}|\boldsymbol{\theta})$, it follows that marginalizing over z in $h(\boldsymbol{\theta},z|\mathbf{y})$, yields the posterior distribution

[19] All steps of the algorithm are provided in a dedicated package, R-INLA, for the general LGM framework, with particular packages also available for implementing INLA in more specific models nested within the LGM class; see Martino and Riebler (2019) for a listing of all such packages.

of interest, namely $p(\boldsymbol{\theta}|\mathbf{y})$. Tran *et al.* then minimize $\text{KL}[q_\lambda(\boldsymbol{\theta}, z)|p(\boldsymbol{\theta}, z|\mathbf{y})]$ over the augmented space of $(\boldsymbol{\theta}, z)$, using as the variational family $\mathcal{Q}$ distributions of the form $q_\lambda(\boldsymbol{\theta}, z) = q_\lambda(\boldsymbol{\theta})g(z|\boldsymbol{\theta})$. Whilst, in general, minimization of $\text{KL}[q_\lambda(\boldsymbol{\theta}, z)|p(\boldsymbol{\theta}, z|\mathbf{y})]$ is not the same as minimization of $\mathcal{L}(\lambda)$, the authors demonstrate the two solutions *do* correspond under particular tuning regimes for $h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$.

Following Tran *et al.* (2017), Ong *et al.* (2018a) propose an alternative VB method for intractable likelihood problems. The authors begin with the recognition that establishing the conditions under which the minimizers of $\text{KL}[q_\lambda(\boldsymbol{\theta}, z)|p(\boldsymbol{\theta}, z|\mathbf{y})]$ and $\mathcal{L}(\lambda)$ coincide is non-trivial, and that in certain types of problems it may be difficult to appropriately tune $h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ so that they coincide. This acknowledgement then prompts them to construct a variational approximation of a simpler target, namely the BSL posterior in (18). By focusing on the (simpler) approximate posterior, rather than the exact posterior $p(\boldsymbol{\theta}|\mathbf{y})$, the approach of Tran *et al.* can be recycled using any unbiased estimator of the synthetic likelihood, $p_a(\eta(\mathbf{y})|\boldsymbol{\theta})$ — which we recall is nothing but a Normal likelihood with unknown mean and variance-covariance matrix — of which several closed-form examples exist. Moreover, since the approach of Ong *et al.* does not rely on the random variables $\mathbf{u}$ in order for its likelihood estimate to be unbiased, no tuning of $h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ is required, and the minimizers of $\text{KL}[q_\lambda(\boldsymbol{\theta}, z)|p(\boldsymbol{\theta}, z|\mathbf{y})]$ and $\mathcal{L}(\lambda)$ will always coincide.

While useful, it must be remembered that the approach of Ong *et al.* (2018a) targets only the *partial* posterior $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$. Furthermore, given the discussion in Section 5.3.2, the approach is likely to perform poorly when the summaries used to construct the unbiased estimator of the synthetic likelihood $p_a(\eta(\mathbf{y})|\boldsymbol{\theta})$ are non-Gaussian. Given that, by definition, the problem is a high-dimensional one, thereby requiring a large collection of summaries, the Gaussian approximation for $\eta(\mathbf{y})$ may not be accurate.

Similar to the above, Barthelmé and Chopin (2014) and Barthelmé *et al.* (2018) propose the use of variational methods to approximate the ABC posterior. The approach of Barthelmé and Chopin is based on 'local' collections of summary statistics that are computed by first partitioning the data into $b \leq n$ distinct 'chunks', $\mathbf{y}_1, \ldots, \mathbf{y}_b$, with possibly differing lengths and support, and then computing the summaries $\eta(\mathbf{y}_i)$ for each of the b chunks. Using this collection of local summaries, the authors then seek to compute an approximation to the following ABC posterior:

$$p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y})) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{b} \left\{ \int p(\mathbf{z}_i|\mathbf{y}_{1:i-1}, \boldsymbol{\theta}) \mathbb{I}\left\{ \|\eta(\mathbf{z}_i) - \eta(\mathbf{y}_i)\| \leq \varepsilon \right\} d\mathbf{z}_i \right\} = p(\boldsymbol{\theta}) \prod_{i=1}^{b} \ell_i(\boldsymbol{\theta}), \tag{31}$$

which implicitly maintains that the 'likelihood chunks', $\ell_i(\boldsymbol{\theta})$, $i = 1, 2, ..., b$, are conditionally independent.

The posterior in (31) is then approximated using expectation propagation (EP) (see Bishop, 2006, Ch 10 for details). The EP approximation seeks to find a tractable density $q_\lambda(\boldsymbol{\theta}) \in \mathcal{Q}$ that is close to $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ by minimizing $\text{KL}\left[p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))|q_\lambda(\boldsymbol{\theta})\right]$. The reader will note that this minimization problem is actually the reverse of the standard variational problem in (25), and is a feasible variational problem because $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ is accessible. Using a factorizable Gaussian variational family with chunk-specific location and covariance parameters, μ_i and variance Σ_i, respectively, $i = 1, \ldots, b$, i.e., $\mathcal{Q} := \{\lambda = (\lambda_1, \ldots, \lambda_b) \in \Lambda : q_\lambda(\boldsymbol{\theta}) := \prod_{i=1}^{b} q_{i,\lambda_i}(\boldsymbol{\theta})\}$, this minimization problem is solved iteratively by minimizing the KL divergence between $\ell_i(\boldsymbol{\theta})$ and $q_{i,\lambda_i}(\boldsymbol{\theta})$ for $i = 1, 2, ..., b$. A coordinate ascent optimization approach allows the i-th variational component to be updated by calculating (using Monte Carlo integration) the

mean and variance of $q_{i,\lambda_i}(\boldsymbol{\theta})$, based on data simulated from $\ell_i(\boldsymbol{\theta})$, conditional on $\boldsymbol{\theta}$ drawn from the variational approximation based on the remaining $j \neq i$ chunks.

By chunking data to create conditionally independent likelihood increments, and by employing (conditionally independent) Gaussian approximations over these chunks, EP-ABC creates a (sequentially updated) Gaussian pseudo-posterior that serves as an approximation to the original ABC posterior. Given that EP-ABC requires the posterior approximation to be Gaussian (or more generally within the linear exponential family), the resulting EP-ABC posterior may not be a reliable approximation to the ABC posterior if the data has strong, or nonlinear, dependence, or (similar to the problem identified for BSL) if (31) has non-Gaussian features, such as thick tails, multimodality or boundary issues. Moreover, the need to generate synthetic data sequentially according to different chunks of the likelihood is unlikely to be feasible in models where there is strong or even moderate serial dependence, and generation of new data requires simulating the entire path history up to that point.

We complete this section on 'hybrid methods' by noting that in state space settings ABC principles have also been used to implement particle filtering, in particular in cases where the measurement density has no closed form and, hence, cannot be used to define the particle weights in the usual way. ABC filtering can be used to estimate the likelihood function, as either a basis for producing frequentist point estimates of $\boldsymbol{\theta}$ (Jasra *et al.*, 2012; Calvet and Czellar, 2015) or as an input into a PMCMC scheme (Dean *et al.*, 2014; Jasra, 2015). More generally, recent algorithms have combined the principles of ABC and Gibbs sampling (Clarté *et al.*, 2020; Rodrigues *et al.*, 2019), with the primary aim being to reduce the impact of dimension on the accuracy of ABC. We reference also Dehideniya *et al.* (2019) for the production of an approximate posterior via the melding of BSL steps with a Laplace approximation.

5.5 MCMC Algorithms Revisited

Despite the rich pickings now on offer with all of the new (including approximate) computational methods, it is *far* from the case that the stalwart of the late 20th century — MCMC — has run its race! Hence, we complete this section with a brief overview of the key advances in MCMC that have been made subsequent to the appearance of the first algorithms, many of which have been motivated by the by the need to deal effectively with large data sets and/or high-dimensional unknowns. Brevity is adopted, not because this segment of the literature is not ripe with developments; in fact, attempts to improve the performance of the original (Gibbs and MH) schemes began early, have been continual, and engage a substantial number of researchers. Rather, we choose to be brief simply because the fundamental principles of the newer advances remain essentially faithful to the original principles of MCMC, and those principles have already been covered herein.[20]

Indeed, we begin with three reminders about MCMC algorithms:

1. *First*, an *MC*-MC algorithm is just that — a *Markov chain* Monte Carlo algorithm. As such, an MCMC scheme — by design — produces a *local* exploration of the target posterior, with the

[20] We acknowledge here a slight inconsistency in our approach, by having allocated (above) a full section to pseudo-marginal MCMC methods, methods that also remain faithful to the fundamental principles of MCMC. However, the goals of the pseudo-marginal literature are arguably broader than just improving algorithmic performance, as we have touched on in Section 5.2.2.

location in the parameter space of any simulated draw being dependent on the location of the previous draw, in a manner that reflects the specific structure of the algorithm. Most notably, an MCMC algorithm with a high degree of dependence will potentially be slow in exploring the high mass region of the target posterior (or the 'target set', in the language of Betancourt, 2018), with this problem usually being more severe the larger is the dimension of the parameter space. Looked at through another lens: for M MCMC draws, the greater the degree of (typically positive) dependence in those draws, the less efficient is the MCMC-based estimate of (4), relative to an estimate based on M *i.i.d.* draws from the target. This loss of efficiency is measured by the so-called *inefficiency factor* (IF), defined (in the case of scalar $g(\boldsymbol{\theta})$) as the ratio of the *MCMC standard error*, σ_{MCMC} with $\sigma^2_{MCMC} = \sqrt{\text{Var}(g(\boldsymbol{\theta}))[1 + 2\sum_{l=1}^{\infty}\rho_l]/M}$ to the standard error associated with M *i.i.d.* draws, σ^2_{MC}, with σ^2_{MC} as given in (13), where ρ_l is the lag-l autocorrelation of the draws of $g(\boldsymbol{\theta})$ over the history of the chain. This ratio, in turn, defines the *effective sample size* of the MCMC algorithm, $ESS = M/[1 + 2\sum_{l=1}^{\infty}\rho_l]$. Improving the efficiency of an MCMC algorithm, for any given value of M, thus equates to increasing ESS to its maximum possible value of M by reducing the dependence in the draws.

2. *Second*: an MC-*MC* algorithm is also a Markov chain *Monte Carlo* algorithm. That is, under appropriate regularity it produces a $\sqrt{M}$-consistent estimate of (4), whatever the degree of dependence in the chain, with the dependence affecting the constant term implicit in the $O(M^{-1/2})$ rate of convergence, but not the rate itself. Hence, in principle, any MCMC algorithm, no matter how inherently inefficient, can produce an estimate of (4) that is arbitrarily accurate, simply through an increase in M. However, an increase in M entails an increase in *computational cost*, measured, say, by *computing clock-time*. The extent of this increase depends, in turn, on the (per-iteration) cost of generating a (proposal/candidate) draw and, with an MH step, the cost of calculating the acceptance probability. Both component costs will (for any algorithm) clearly increase with the number of unknowns that need to be simulated, and assessed, at each iteration. Either cost, or both, will also increase with the sample size, given the need for pointwise evaluation of the likelihood function across the elements of $\mathbf{y}$.

3. *Third*: the very concept of efficiency is relevant only if the Markov chain is (asymptotically in M) *unbiased*, which depends critically on draws being produced from the correct invariant distribution. That is, the production of an accurate MCMC-based estimate of (4) depends, not just on reducing the degree of dependence in the chain, or on increasing the number of draws, but on ensuring that the chain *actually* explores the target set, and thereby avoids bias in the estimation of (4).[21]

Hence, all advances in MCMC - at their core — aim to increase the effectiveness with which an

[21] It is acknowledged in the literature that MCMC algorithms produce potentially strong biases in their initial phase of 'convergence' to the typical set from an initial point in the parameter space. However, under appropriate regularity, such biases are transient, and their impact on the estimation of (4) able to be eliminated by discarding a sufficiently large number of 'burn-in' or 'warm-up' draws from the computation. (See Robert and Casella, 2004, and Gelman and Shirley, 2011, for textbook discussions of convergence, including diagnostic methods.) Some of the more recent literature is concerned with removing this transitory bias after a finite number of iterations; e.g. Jacob *et al.* (2020). Other literature is concerned with ensuring that an MCMC algorithm does not yield a bias that is *non-transitory* due to the inability of the algorithm to effectively explore the target set at all (within a meaningful time frame); see e.g. Betancourt (2018).

algorithm explores the high mass region of the target posterior and, hence, the accuracy with which (4) is estimated, by doing one (or more) of three things: reducing dependence in the chain, reducing the computational cost per iteration of the chain (thus enabling more draws to be produced), or eliminating bias. Focus is increasingly directed towards developing algorithms that *scale well*, in terms of the dimension of the data and/or the number of unknowns.

With our goal of brevity in mind, we simply list the main contenders here, including certain key references or reviews, deflecting both to those papers, and to the broad overviews of modern developments in MCMC in Green *et al.* (2015), Robert *et al.* (2018) and Dunson and Johndrow (2019) for all details. Of particular note is the recent survey on Bayesian methods for 'Big Data' in Jahan *et al.* (2020) (Section 5.1 being most pertinent), which describes the precise manner in which certain of the methods cited below (and others) tackle the problem of scale. We categorize the methods according to whether improved performance is achieved (primarily): *i)* via the exploitation of more geometric information about the target posterior; *ii)* by better choice of proposals; *iii)* by the use of parallel, batched, subsample, coupled or ensemble sampling methods; or *iv)* by the explicit use of variance reduction methods.

i) Hamiltonian Monte Carlo (HMC) (Neal, 2011b; Carpenter *et al.*, 2017; Betancourt, 2018); no U-turn sampling (NUTS) (Hoffman and Gelman, 2014);[22] Metropolis-Adjusted Langevin algorithm (MALA) (Roberts *et al.*, 1996; Roberts and Rosenthal, 1998); stochastic gradient MCMC (Nemeth and Fearnhead, 2019); piecewise deterministic Markov processes (PDMP) (Bierkens *et al.*, 2018; Fearnhead *et al.*, 2018; Bierkens *et al.*, 2019).

ii) Optimal scaling of random-walk MH (Roberts *et al.*, 1997); adaptive sampling (Nott and Kohn, 2005; Roberts and Rosenthal, 2009; Rosenthal, 2011); MCMC with ordered overrelaxation (Neal, 1998); simulated tempering, parallel tempering and tempered transition methods (Geyer, 1991; Marinari and Parisi, 1992; Neal, 1996; Gramacy *et al.*, 2010; Geyer, 2011a; Tawn *et al.*, 2020); delayed rejection sampling (Tierney and Mira, 1998); delayed acceptance sampling (Christen and Fox, 2005; Golightly *et al.*, 2015; Wiqvist *et al.*, 2018; Banterle *et al.*, 2019); multiple try MCMC (Liu *et al.*, 2000; Bédard *et al.*, 2012; Martino, 2018; Luo and Tjelmeland, 2019); taylored randomized block MH (TaRB-MH) (Chib and Ramamurthy, 2010); tempered Gibbs sampling (TGS) (Zanella and Roberts, 2019); quasi-stationary Monte Carlo and subsampling (Pollock *et al.*, 2020).

iii) Parallelized MCMC (Jacob *et al.*, 2011; Wang and Dunson, 2013); subposterior (batched) methods (Neiswanger *et al.*, 2013; Scott *et al.*, 2016); subsampling methods based on pseudo-marginal MCMC (Bardenet *et al.*, 2017; Quiroz *et al.*, 2018b; Quiroz *et al.*, 2019); perfect sampling (Propp and Wilson, 1996; Casella *et al.*, 2001, Craiu and Meng, 2011; Huber, 2016); unbiased MCMC via coupling (Glynn and Rhee, 2014; Glynn, 2016; Middleton *et al.*, 2018; Jacob *et al.*, 2020); unbiased MCMC for doubly-intractable problems using pseudo-marginal principles (Lyne *et al.*, 2015); ensemble MCMC (Iba, 2000; Cappé *et al.*, 2004; Neal, 2011a).

[22] As described in Neal (2011b), simulation methods based on Hamiltonian dynamics can actually be viewed as having as long a history as MCMC itself. The more modern manifestations of HMC, however, including NUTS, can be viewed as Markov chain algorithms that simply explore the parameter space more effectively than (say) a default random walk scheme. The probabilistic programming platform Stan (Carpenter *et al.*, 2017) enables implementation of NUTS, in addition to certain variants of VB.

iv) Rao-Blackwellization (Casella and Robert, 1996; Robert and Casella, 2004; Douc and Robert, 2011); antithetic variables (Frigessi *et al.*, 2000; Craiu and Meng, 2005); control variates (Dellaportas and Kontoyiannis, 2012; Baker *et al.*, 2019); thinning (Owen, 2017).

6 The Role of Computation in Model Choice and Prediction

Thus far, our focus has been on computing the posterior expectation in (4) defined in the context of an assumed model. Other than the brief reference made to the expectation that defines the marginal likelihood in (5), to the lower bound on the marginal likelihood yielded by the VB procedure, and to the deterministic approximation of it produced by INLA, the issue of model uncertainty itself has not been addressed; nor have the specific issues that arise when the expectation in (4) defines the predictive distribution. We touch on these topics below. In Section 6.1 each model in the assumed model space is treated separately, and the set of *posterior model probabilities* so produced used to make decisions. In Section 6.2 model uncertainty is directly incorporated via the principle of augmentation, with inference about the model being a direct outcome. In Section 6.3 we look at Bayesian prediction.

6.1 Model Uncertainty and Marginal Likelihood Computation

We begin by adopting the simplest possible characterization of model uncertainty, where the model space is spanned by two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, with prior probabilities, $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$ respectively. A simple application of the Bayesian calculus leads to the following expression for the ratio of posterior model probabilities (or *posterior odds ratio):*

$$\frac{p(\mathcal{M}_1|\mathbf{y})}{p(\mathcal{M}_2|\mathbf{y})} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \times \frac{p(\mathbf{y}|\mathcal{M}_1)}{p(\mathbf{y}|\mathcal{M}_2)}, \tag{32}$$

where

$$p(\mathbf{y}|\mathcal{M}_k) = \int_{\Theta_k} p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{M}_k) p(\boldsymbol{\theta}_k|\mathcal{M}_k) d\boldsymbol{\theta}_k, \tag{33}$$

$\boldsymbol{\theta}_k$ is the unknown parameter (vector) for model $\mathcal{M}_k$, $k = 1, 2$, and $p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)$ and $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ define respectively the likelihood and prior conditioned explicitly on $\mathcal{M}_k$. The density in (33) defines, equivalently, the *marginal likelihood*, the *marginal data density*, or the *evidence* of model $\mathcal{M}_k$, and the ratio of the two such densities on the right-hand-side of (32) defines the *Bayes factor*.

Computation of $p(\mathcal{M}_1|\mathbf{y})$ and $p(\mathcal{M}_2|\mathbf{y})$ proceeds via (32) allied with the restriction that $p(\mathcal{M}_1|\mathbf{y}) + p(\mathcal{M}_2|\mathbf{y}) = 1$ (with the extension to $K > 2$ models being obvious). *Model choice* can be performed by invoking decision-theoretic arguments, and minimizing expected posterior loss. This leads to $\mathcal{M}_1$ being chosen if $p(\mathcal{M}_1|\mathbf{y})/p(\mathcal{M}_2|\mathbf{y})$ exceeds the ratio of losses of 'Type 2' and 'Type 1' errors. *Model averaging* can also be used, whereby the posterior expectation of a quantity of interest is computed for both models, then averaged, using $p(\mathcal{M}_1|\mathbf{y})$ and $p(\mathcal{M}_2|\mathbf{y})$ as the weights (an example of which is given in Section 6.3). Textbook illustrations of all such steps can be found in Zellner (1971), Koop (2003) and Robert (2007).

Key to all of this is the evaluation of the two integrals in (33). As noted in Section 2.1, analytical solutions to (33) are available only for certain special cases. Whilst the VB approach to computing

$\mathbb{E}(g(\boldsymbol{\theta})|\boldsymbol{y})$ within the context of a given model yields, as a bi-product, a lower bound on the evidence for that model, we focus in this section on methods that target the marginal likelihood *directly*. We do not reproduce here the INLA-based method for computing $p(\mathbf{y})$ that has been described in Section 5.3.5.

The integral in (33) is the prior expectation in (5) defined for $\mathcal{M}_k$, $k = 1, 2$. Being a prior, rather than a posterior expectation has two consequences. First, it is well-defined only if $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ is a *proper* density function. Hence, Bayes factors, and the posterior odds ratios that they imply, cannot be computed with impunity under (typically improper) non-informative, or objective, priors. Attempts to incorporate objective prior information into Bayes factors have been made (see Strachan and van Dijk, 2014, for a recent treatment and relevant referencing); however the convention in the literature remains one of adopting informative, proper priors in the computation of posterior model probabilities. The *second* consequence relates to computation: the most direct approach to computing (33) via simulation, namely drawing $\boldsymbol{\theta}_k$ from $p(\boldsymbol{\theta}_k)$, and averaging $p(\mathbf{y}|\boldsymbol{\theta}_k)$ over the draws, is typically inaccurate, as $p(\boldsymbol{\theta}_k)$ will not necessarily have high mass in the high-mass region of the likelihood function. All simulation-based estimates of (33) thus use draws of $\boldsymbol{\theta}_k$ that are informed by the data in some way, to improve accuracy.

Alternative methods for simulation-based estimation of (33) have been proposed, in addition to — and sometimes combined with — either partial analytical solutions or asymptotic (Laplace) approximations. We refer the reader to Kass and Raftery (1995), Geweke (1999), Marin and Robert (2011), Chib (2011) and Fourment *et al.* (2018) for reviews. We simply emphasize below three distinct uses of simulation, all of which nest, or can be linked to, a range of specific methods, not all of which we cover here. (We refer the reader to Aridia *et al.*, 2012, for a useful comparative study of some of the simulation methods we review below, plus additional referencing.) Section headings that indicate the over-arching *principle* underlying each approach are adopted.

6.1.1 Importance sampling

As already noted, the original motivation of Hammersley and Handscomb (1964) in using IS to compute integrals was that of variance reduction. The authors highlight that a judicious choice of importance density can yield a more accurate estimator of any integral than a 'crude' Monte Carlo estimator, as they term it, by targeting the parts of the support where the integrand is 'important'.

The problem of computing (33) can be approached in the same way. Defining $w^{(i)} = p(\mathbf{y}|\boldsymbol{\theta}_k^{(i)})p(\boldsymbol{\theta}_k^{(i)})/q(\boldsymbol{\theta}_k^{(i)}|\mathbf{y})$, $i = 1, 2, ..., M$, for draws, $\boldsymbol{\theta}_k^{(i)}$, from some suitable importance density $q(\cdot|\mathbf{y})$, an IS approach produces $\widehat{p}(\mathbf{y}|\mathcal{M}_k) = \sum_{i=1}^{M} w^{(i)}/M$. Subject to regularity on $q(\cdot|\mathbf{y})$, the usual asymptotic (in M) arguments can be invoked to prove the consistency and asymptotic normality of $\widehat{p}(\mathbf{y}|\mathcal{M}_k)$ as an estimator of (33) (see Geweke, 1999). Note that, on the assumption that $p(\mathbf{y}|\boldsymbol{\theta}_k^{(i)}, \mathcal{M}_k)$ and $p(\boldsymbol{\theta}_k^{(i)})$ are available in closed form, and that $q(\cdot|\mathbf{y})$ is known in in its entirety (i.e. including its integrating constant), no additional normalization step (like that in (14)) is required. See Geweke (1989b), Gelfand and Dey (1994) and Raftery (1996) for early examples of this approach, and Geweke (1999) for illustration of a non-*i.i.d.* version, based on draws from an MH candidate distribution, $q(\cdot|\mathbf{y})$.

The general principle of IS has, of course, two aspects to it: *i)* the use of draws simulated from the importance density to compute a weighted mean; and *ii)* evaluation of the importance density in the

weight. The 'reciprocal IS' (RIS) method (Gelfand and Dey, 1994; Frühwirth-Schnatter, 2004) uses the second aspect, whilst taking draws from the posterior itself. Simple calculations can be used to show that, for some $q(\cdot|\mathbf{y})$ that is contained in the support of $p(\boldsymbol{\theta}_k|\mathbf{y})$, and defining $g(\boldsymbol{\theta}_k) = q(\boldsymbol{\theta}_k|\mathbf{y}) / [p(\mathbf{y}|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k)]$,

$$\mathbb{E}(g(\boldsymbol{\theta}_k)|\mathbf{y}) = [p(\mathbf{y}|\mathcal{M}_k)]^{-1}. \tag{34}$$

Again, under stringent regularity conditions, including some on the support of $q(\boldsymbol{\theta}_k|\mathbf{y})$, draws from $p(\boldsymbol{\theta}_k|\mathbf{y})$, can be used to estimate (34), and its reciprocal used as an estimate of (33) itself. The 'harmonic mean estimator' of Newton and Raftery (1994) is a special, if infamous, case of (34) in which $q(\boldsymbol{\theta}_k|\mathbf{y}) = p(\boldsymbol{\theta}_k)$. The complete untrustworthiness of the output when the prior has fatter tails than the posterior (Neal, 1994, 1999) have led to this method being largely eschewed in the literature. The 'bridge sampler', on the other hand, provides a more robust version of (34) by exploiting draws from *both* the posterior and the weight density, $q(\boldsymbol{\theta}_k|\mathbf{y})$ (Meng and Wong, 1996; Meng and Schilling, 2002; Frühwirth-Schnatter, 2004; Gronau *et al.*, 2017). More recent versions of the RIS method retain draws from the posterior, but use VB approximations of $p(\boldsymbol{\theta}_k|\mathbf{y})$ to define $q(\boldsymbol{\theta}_k|\mathbf{y})$ (Fourment *et al.*, 2018; Hajargasht and Woźniak, 2018).

Finally, building on the thermodynamic integration method from theoretical physics, Gelman and Meng (1998) develop the 'path sampler'. They illustrate the sense in which the earlier IS schemes for estimating (33), followed by bridge sampling, and then path sampling, represent a natural methodological lineage. See also Rischard *et al.* (2018) for a recent amalgam of path sampling with unbiased MCMC, and Neal (2001) for related work using 'annealed importance sampling' to estimate (33).[23]

6.1.2 Multiple runs of MCMC

Given the definition of $p(\boldsymbol{\theta}_k|\mathbf{y})$, we can produce a representation of the marginal likelihood as:

$$p(\mathbf{y}|\mathcal{M}_k) = \frac{p(\mathbf{y}|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k)}{p(\boldsymbol{\theta}_k|\mathbf{y})}. \tag{35}$$

The insight of Chib (1995) was to recognize that (35) holds for all $\boldsymbol{\theta}$. Hence, an estimate of $p(\mathbf{y}|\mathcal{M}_k)$ is simply produced as: $\widehat{p}(\mathbf{y}|\mathcal{M}_k) = p(\mathbf{y}|\boldsymbol{\theta}_k^*)p(\boldsymbol{\theta}_k^*)/p(\boldsymbol{\theta}_k^*|\mathbf{y})$, where the convention is to take $\boldsymbol{\theta}_k^*$ as some high posterior value. Whilst the (common) availability of the likelihood and prior in closed form renders the ordinates of the factors in the numerator readily accessible in most cases, the denominator and, indeed, the value of $\boldsymbol{\theta}_k^*$ itself are, by the very nature of the problem, inaccessible without further work. However, defining $\boldsymbol{\theta}_k^* = (\theta_{k,1}^*, \theta_{k,2}^*, ..., \theta_{k,p_k}^*)'$, the joint posterior (evaluated at $\boldsymbol{\theta}_k^*$) can be decomposed as:

$$p(\boldsymbol{\theta}_k^*|\mathbf{y}) = p(\theta_{k,1}^*|\theta_{k,2}^*, ..., \theta_{k,p_k}^*, \mathbf{y})p(\theta_{k,2}^*|\theta_{k,3}^*, ..., \theta_{k,p_k}^*, \mathbf{y})...p(\theta_{k,p_k}^*|\mathbf{y}). \tag{36}$$

The last term on the right-hand-side of (36) can simply be estimated in the usual way using a full run of an MCMC sampler. The remaining conditionals can be estimated from additional applications of

[23] On a related thread, note that SMC has often been advocated for estimating the evidence (Doucet *et al.*, 2000; Fiel and Wyse, 2012; Everitt *et al.*, 2020). Similarly, techniques incorporating the intractable marginal likelihood as a supplementary parameter can be traced back to Geyer (1993), with more recent occurrences like noise-contrastive estimation (Gutmann and Hyvärinen, 2012), being based on a so-called 'logistic trick' that turns the approximation of the evidence into the estimation of the intercept of a logistic classification program. See also Barthelmé and Chopin (2015) and Lyne *et al.* (2015) for related work.

the same simulation scheme, but with the appropriate sets of parameters held fixed. Modification of the original (pure Gibbs) approach proposed in Chib (1995) to cater for full conditionals that are not available in closed form, by using MH steps, is detailed in Chib and Jeliazkov (2001). Recent work in Chib *et al.* (2020) highlights the effective use of parallelization to reduce the computational cost of the additional 'reduced' MCMC runs.

6.1.3 Nested sampling

Nested sampling (Skilling, 2007) is yet another method for producing a simulation-based estimate of $p(\mathbf{y}|\mathcal{M}_k)$. Whilst it gained an immediate foothold in astronomy (Mukherjee *et al.*, 2006) — possibly due to the availability of dedicated software like MultiNest and Dynesty — and remains popular in that field (Feroz *et al.*, 2019), it has not gained wide acceptance beyond that field. A cartoon description of the method is as the simulation version of Lebesgue integration, in that M points are simulated on slices of the likelihood function (delineated by points t, $t-1$ in the support of $\boldsymbol{\theta}$), $\{\boldsymbol{\theta};\ p(\mathbf{y}|\boldsymbol{\theta}_{t-1}) \leq p(\mathbf{y}|\boldsymbol{\theta}) \leq p(\mathbf{y}|\boldsymbol{\theta}_t)\}$, with each slice having approximately a prior probability, $\exp\{-(t-1)/M\}$, of occurring. As first shown in Chopin and Robert (2010), nested sampling is a Monte Carlo method with a $\sqrt{M}$ rate of convergence, whose performance relies on the ability to efficiently simulate parameters within the above slices, which is challenging when these are not connected.

6.2 Reversible Jump MCMC

All methods described above for computing $p(\mathbf{y}|\mathcal{M}_k)$ have one thing in common: the marginal likelihood for *each* model is tackled as a separate computational exercise. Once each $p(\mathbf{y}|\mathcal{M}_k)$ is computed, the posterior model probabilities follow, and model choice, or model averaging, can proceed.

Alternatively, uncertainty about the model can be used to *augment* the unknowns, and a posterior sampler designed to target this augmented space. This is the principle adopted in a range of papers, including those that focus on variable selection in regression models, and the number of components in finite mixture models; and we refer the reader to George (2000), Marin *et al.* (2005) and Chib (2011) for reviews and references. We focus here on one particular approach only: that of Green (1995).

Green (1995) characterizes the problem of an unknown model as one in which the dimension of the (model-specific) unknowns *varies*, depending on which model is in play. He thus designs an MCMC sampler that is allowed to *jump* between parameter spaces of differing dimensions; coining the term: 'reversible jump MCMC' (RJMCMC). At its core though, Green's approach is one of augmentation, and can be viewed as a particular application of the original idea of Tanner and Wong (1987), with the extra complexity of dimension variation as the sampler traverses the augmented space.

In brief, Green (1995) assumes a countable collection of candidate models $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, ...\}$, indexed by $k = 1, 2, ...$ Each model has a p_k-dimensional set of unknown parameters $\boldsymbol{\theta}_k$. Using obvious notation for the likelihood and prior for the kth model, and the prior $p(k)$ for the model index itself, the target of the RJMCMC algorithm is then: $p(k, \boldsymbol{\theta}_k|\mathbf{y}) = p(\mathbf{y}|k, \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|k)p(k)/p(\mathbf{y})$. The RJMCMC sampler moves between any two parameters spaces by creating temporary auxiliary variables that bring the dimensions of the augmented spaces to be equal, with a reversibility constraint on the proposed moves between these models. Such draws from the joint space of $\{k, \boldsymbol{\theta}_k\}$ can be used to compute

any particular $\mathbb{E}(g(\boldsymbol{\theta}_k)|\mathbf{y})$ of interest. Indeed, the draws can also be used to compute an expectation of the form: $\mathbb{E}(g(k)|\mathbf{y})$, which nests the marginal posterior probability attached to the kth model: $p(k|\mathbf{y})$. That is, posterior model probabilities are an automatic outcome of the simulation scheme. Moreover, the computation of *any* expectation of interest incorporates all uncertainty associated with both the parameters of each model and the model itself; hence model averaging is automatic. See Green (2003), Fan and Sisson (2011) and Geyer (2011b) for reviews of RJMCMC, including: its theoretical and implementation properties, its links to other 'multi-model' samplers, and the scope of its application.

6.3 Computation in Bayesian Prediction

Conditional on the assumed model that underpins the likelihood function being 'correctly specified' — i.e. $p(\mathbf{y}|\boldsymbol{\theta})$ coinciding with the true data generating process (DGP) — the 'gold standard' for Bayesian prediction is the specification of (4) with $g(\boldsymbol{\theta}) = p(y^*_{n+1}|\boldsymbol{\theta}, \mathbf{y})$, which yields:

$$p(y^*_{n+1}|\mathbf{y}) = \int_\Theta p(y^*_{n+1}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{37}$$

The distribution $p(y^*_{n+1}|\mathbf{y})$ summarizes all uncertainty about y_{n+1}, conditional on both the assumed model — which underpins the structure of both the conditional predictive, $p(y^*_{n+1}|\boldsymbol{\theta}, \mathbf{y})$, and the posterior itself — and the prior beliefs that inform $p(\boldsymbol{\theta}|\mathbf{y})$. Point and interval predictions of y_{n+1}, and indeed any other distributional summary, can be extracted from (37). In the case where the model itself is uncertain, and a finite set of models, $\mathcal{M}_1$, $\mathcal{M}_2$,...,$\mathcal{M}_K$, is assumed to span the model space, the principle of model averaging can be used to produce a 'model-averaged' predictive, $p_{MA}(y^*_{n+1}|\mathbf{y})$ as

$$p_{MA}(y^*_{n+1}|\mathbf{y}) = \sum_{k=1}^{K} p(y^*_{n+1}|\mathbf{y}, \mathcal{M}_k) p(\mathcal{M}_k|\mathbf{y}), \tag{38}$$

where $p(y^*_{n+1}|\mathbf{y}, \mathcal{M}_k)$ denotes the density in (37), but conditioned explicitly on the kth model in the set. In the typical case where (37) and (38) are unavailable analytically, *any* of the computational methods that have been discussed thus far could be used to compute either $p(y^*_{n+1}|\mathbf{y})$ or $p_{MA}(y^*_{n+1}|\mathbf{y})$, and any summaries thereof. In some ways then, this could be viewed as completing the Bayesian prediction story.

However, Bayesian prediction is more than just a special case of the general Bayesian computational problem defined by (4). Predicting outcomes that have not yet been observed is arguably the most stringent test to which any statistical methodology can be put. Moreover, out-of-sample accuracy is the ultimate arbiter in this setting, and any computational method used to produce predictions needs to be assessed in this light.

Such is the motivation of recent works that pose the following question: In cases (such as those highlighted in Section 5.3) where the exact posterior is inaccessible and the exact predictive in (37) is thus also unavailable, what are the implications for predictive accuracy of adopting an approximation to $p(\boldsymbol{\theta}|\mathbf{y})$ and, hence, $p(y^*_{n+1}|\mathbf{y})$? Frazier *et al.* (2019a), for instance, document that an 'approximate predictive' produced by replacing $p(\boldsymbol{\theta}|\mathbf{y})$ with an ABC-based posterior, is numerically indistinguishable (in the cases investigated) from the exact predictive and, thus, yields equivalent predictive accuracy. Further, under certain regularity conditions, the exact and approximate predictives are shown to be asymptotically

(in n) equivalent. The tenor of related work exploring prediction in approximate inference (including VB) settings — Park and Nassar (2014), Canale and Ruggiero (2016), Koop and Korobilis (2018), Quiroz *et al.* (2018a) and Kon Kam King *et al.* (2019) — is somewhat similar to that of Frazier *et al.*; that is, computing $p(\boldsymbol{\theta}|\mathbf{y})$ via an approximate method does not *necessarily* reduce predictive accuracy.

However, an arguably more challenging question is: How does one even think about Bayesian prediction — and the use of computation therein — once one acknowledges that, in reality: *i)* any given model is misspecified; and *ii)* any finite set of models does not span the truth? The 'probably approximately correct (PAC)-Bayes' approach in machine learning (reviewed in Guedj, 2019) replaces the likelihood function in the Bayesian update with the exponential of a general loss function, with a view to producing predictions that are more targeted to the problem at hand, rather than being tied to a particular model specification. This approach yields so-called 'Gibbs posteriors' (Zhang, 2006a; Zhang, 2006b; Jiang and Tanner, 2008), and mimics — in a prediction setting — the generalized *inferential* methods proposed in (*inter alia*) Bissiri *et al.* (2016), Giummolè *et al.* (2017), Holmes and Walker (2017), Lyddon *et al.* (2019) and Syring and Martin (2019). In addition, PAC-Bayes uses updates based on 'tempered', or 'power' likelihoods, in which robustness to model misspecification is sought by raising the likelihood associated with an assumed model to a particular power.

Other work tackles this issue by estimating weighted combinations of predictives via either forecast accuracy criteria or the criterion of predictive calibration (Dawid, 1982; Gneiting *et al.*, 2007), without assuming that the true model lies within the set of constituent predictives (Billio *et al.*, 2013; Casarin *et al.*, 2015; Pettenuzzo and Ravazzolo, 2016; Bassetti *et al.*, 2018; Baştürk *et al.*, 2019). In a similar spirit, Loaiza-Maya *et al.* (2020) propose the use of *focused Bayesian prediction*, in which the Bayesian update is driven by a criterion that captures a user-specified measure of predictive accuracy. The authors do indeed find that focusing on the loss that matters produces superior predictive performance relative to using a misspecified likelihood update.

In summary, once attention is on the particular version of (4) that yields a predictive distribution, it is the updating criterion function itself that is increasingly being viewed as key, with uncertainty about the unknown parameters — and the computational method used to quantify that uncertainty — assuming a somewhat secondary role.

7 The Future

Our journey with Bayesian computation began in *1763*: with a posterior probability defined in terms of a scalar θ, whose solution challenged Bayes. We now end our journey in *2020*: having referenced posterior distributions defined over thousands, possibly millions of unknowns, and computational problems with a degree of complexity — and scale — to match. Along the way, we have seen the huge variety of imaginative computational solutions that have been brought to bear on all such problems, over the span of 250 years. A natural question to ask then is: 'what is there left to do?'

Judging from the wealth of contributions to '*Bayes Comp 2020*', in January 2020 (http://users.stat.ufl.edu/~jhobert/BayesComp2020/Conf_Website/), the answer to that question is: 'a great deal!'; and *most certainly* as pertains to matters of scale. Indeed, of the 129 abstracts recorded on the conference

website (and acknowledging some double counting), 16 make explicit use of the term *scalability* (or something comparable); 10 refer to the ability of a proposed computational method to deal effectively with *large data sets*, and 22 refer to *high-dimensional problems* of one form or another. There are attempts to scale most categories of computational methods reviewed in this paper, and the scope of the empirical problems to which these advances are applied — from probabilistic topic models, health studies on hypertension and sepsis, problems in neuroimaging, genomics, biology, epidemiology, ecology, psychology and econometrics, through to probabilistic record linkage and geographic profiling — is extremely broad, highlighting again the critical importance of Bayesian computation to diverse fields.

But if *scale* may be viewed as a key focus of this selection of research, there is another theme that can also be discerned. A second glance at these conference proceedings, and at recent journal publications, pinpoints a growing interest in the impact of *model misspecification* on computational methods; specifically: *1)* what are the implications for computation if an assumed parametric model is misspecified?; and, *2)* what are the implications for computation if the conventional likelihood-based paradigm is eschewed altogether (as is already happening in some prediction settings)? Remembering that our interest here is on the implications for *computation per se* of both misspecified and non-likelihood settings, we note that specific attempts to address *1)* and *2)* directly are still quite small in number, if growing.[24] We complete our review by briefly summarizing *five* recent (sets of) papers in this vein:

i) *First*, Lyddon *et al.* (2019) and Syring and Martin (2019) use bootstrap principles to compute so-called *general Bayesian posteriors*, in which the likelihood function associated with an assumed (and potentially misspecified) model is replaced by a more general loss function that is not tied to a particular model specification. Huggins and Miller (2019) also use the bootstrap to construct 'bagged' posteriors (*BayesBag*), and thereby conduct Bayesian inference that is robust to model misspecification.

ii) *Second*, Frazier *et al.* (2020) analyze the theoretical properties of ABC under model misspecification; outlining when ABC concentrates posterior mass on an appropriately defined pseudo-true value, and when it does not. The nonstandard asymptotic behaviour of the ABC posterior, including its failure to yield credible sets with valid frequentist coverage, is highlighted. The authors also devise techniques for diagnosing model misspecification in the context of ABC. Frazier and Drovandi (2019) devise a version of BSL that is robust to model misspecification, and demonstrate that this version can be much more computationally efficient than standard BSL when the model is misspecified.

iii) *Third,* Wang and Blei (2019a) investigate VB under model misspecification. They demonstrate that the VB posterior both concentrates on the value that minimizes the Kullback-Leibler (KL) divergence from the true DGP, and is asymptotically normal; as is the VB posterior mean. These

[24] We refer to Kleijn and van der Vaart (2012) and Muller (2013) for general treatments of Bayesian likelihood-based inference in misspecified models; to some of the literature cited in Section 6.3, plus Chernozhukov and Hong (2003),Gallant (2016), Chib *et al.* (2018), and Miller and Dunson (2019), for various generalizations of the standard Bayesian paradigm, including posterior up-dates driven by problem-specific loss (or moment) functions. It can be argued, however, that in none of these papers are the implications of model misspecification and/or non-likelihood up-dates for computation *per se* the primary focus.

results generalize the asymptotic results for VB of Wang and Blei (2019b), derived under correct specification, to the misspecification case.

iv) *Fourth*, Knoblauch *et al.* (2019) propose what they term *generalized variational inference,* by extending the specification of the Bayesian paradigm to accommodate general loss functions (thereby avoiding the reliance on potentially misspecified likelihoods) and building an optimization-based computational tool within that setting.

v) *Fifth*, building on earlier work in the context of HMC (and which is cited therein), Bornn *et al.* (2019) derive an MCMC scheme for sampling on the lower-dimensional manifold implied by the moment conditions that they adopt within a Bayesian framework. Whilst this work embeds the moments within a nonparametric Bayesian set-up — and we have not discussed computation in nonparametric settings in this review — we make note of this work as an example of a fundamental shift in computational design that is required when moving to a particular *non-likelihood* Bayesian up-date. Whilst not motivated by this same problem, the work on extending VB to manifolds by Tran *et al.* (2019) is also relevant here.

Bayesian computation in parametric models is thus beginning to confront — and adapt to — the reality of misspecified DGPs, and the generalizations beyond the standard likelihood-based up-date that are evolving. Allied with the growing ability of computational methods to also deal with the scale of modern problems, the future of the paradigm in the 21st century thus seems assured. And with this, the 18th century Bayes (and his loyal champion, Price) would no doubt be duly impressed!

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. American Statist. Assoc.*, 88:669–679.

Andrieu, C., Doucet, A., and Holenstein, R. (2011). Particle Markov chain Monte Carlo. *J. Royal Statist. Society Series B*, 72(2):269–342. With discussion.

Andrieu, C., Doucet, A., and Robert, C. (2004). Computational advances for and from Bayesian analysis. *Statist. Science*, 19(1):118–127.

Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725.

Baker, J., Fearnhead, P., Fox, E., and Nemeth, C. (2019). Control variates for stochastic gradient mcmc. *Statist. Comp.*, 29:599–615.

Banterle, M., Grazian, C., Lee, A., and Robert, C. P. (2019). Accelerating Metropolis-Hastings algorithms by delayed acceptance. *Foundations of Data Science*, 1(2):103–128.

Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *J. Machine Learning Res.*, 18(1):1515–1557.

Barnard, G. and Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3/4):293–315.

Barthelmé, S., Chopin, N., and Cottet, V. (2018). Divide and conquer in ABC: Expectation-propagation algorithms for likelihood-free inference. *Handbook of Approximate Bayesian Computation*, pages 415–34. Chapman & Hall/CRC. Eds. Sisson, S., Fan, Y., Beaumont, M.

Barthelmé, S. and Chopin, N. (2014). Expectation propagation for likelihood-free inference. *J. American Statist. Assoc.*, 109(505):315–333.

Barthelmé, S. and Chopin, N. (2015). The Poisson transform for unnormalised statistical models. *Stat. Comput.*, 25:767–780.

Bassetti, F., Casarin, R., and Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *J. American Statist. Assoc.*, 113(522):675–685.

Bauwens, L. and Richard, J. (1985). A 1-1 Poly-t random variable generator with application to Monte Carlo integration. *J. Econometrics*, 29(1):19–46.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London for 1973*, 53:370–418.

Baştürk, N., Borowska, A., Grassi, S., Hoogerheide, L., and van Dijk, H. (2019). Forecast density combinations of dynamic models and data driven portfolio strategies. *Journal of Econometrics*, 210(1):170–186.

Beaumont, M. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160.

Beaumont, M. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406.

Beaumont, M., Cornuet, J.-M., Marin, J.-M., and Robert, C. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990.

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bédard, M., Douc, R., and Moulines, E. (2012). Scaling analysis of multiple-try MCMC methods. *Stochastic Processes and their Applications*, 122(3):758–786.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the wasserstein distance. *J. Royal Statist. Society Series B*, 81(2):235–269.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Society Series B*, 36(2):192–326. With discussion.

Besag, J. and Green, P. (1993). Spatial statistics and Bayesian computation. *J. Royal Statist. Society Series B*, 55(1):25–37. With discussion.

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *https://arxiv.org/abs/1701.02434v2*.

Bierkens, J., Bouchard-Côté, A., Doucet, A., Duncan, A. B., Fearnhead, P., Lienart, T., Roberts, G., and Vollmer, S. J. (2018). Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains. *Statist. Prob. Letters*, 136:148–154.

Bierkens, J., Fearnhead, P., and Roberts, G. (2019). The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Statist.*, 47(3):1288–1320.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213–232.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *J. Royal Statist. Society Series B*, 78(5):1103–1130.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *J. American Statist. Assoc.*, 112(518):859–877.

Blum, M. (2010). Approximate Bayesian computation: a non-parametric perspective. *J. American Statist. Assoc.*, 105(491):1178–1187.

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statist. Science*, 28(2):189–208.

Bornn, L., Pillai, N. S., Smith, A., and Woodard, D. (2017). The use of a single pseudo-sample in approximate Bayesian computation. *Statist. Comp.*, 27(3):583–590.

Bornn, L., Shephard, N., and Solgi, R. (2019). Moment conditions and Bayesian non-parametrics. *J. Royal Statist. Society Series B*, 81(1):5–43.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *J. American Statist. Assoc.*, 105(489):324–335.

Brooks, S., Gelman, A., Jones, G., and Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Taylor & Francis.

Buchholz, A. and Chopin, N. (2019). Improving approximate Bayesian computation via quasi-Monte Carlo. *J. Comput.*

Graph. Statist., 28(1):205–219.

Calvet, L. E. and Czellar, V. (2015). Accurate methods for approximate Bayesian computation filtering. *J. Finan. Econometrics*, 13(4):798–838.

Canale, A. and Ruggiero, M. (2016). Bayesian nonparametric forecasting of monotonic functional time series. *Electronic Journal of Statistics*, 10(2):3265–3286.

Cappé, O., Guillin, A., Marin, J., and Robert, C. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.*, 13(4):907–929.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1).

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Casarin, R., Leisen, F., Molina, G., and ter Horst, E. (2015). A Bayesian beta Markov random field calibration of the term structure of implied risk neutral densities. *Bayesian Analysis*, 10(4):791–819.

Casella, G. and George, E. (1992). An introduction to Gibbs sampling. *American Statist.*, 46:167–174.

Casella, G., Lavine, M., and Robert, C. P. (2001). Explaining the perfect sampler. *American Statist.*, 55(4):299–305.

Casella, G. and Robert, C. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94.

Ceruzzi, P. (2003). *A History of Modern Computing*. MIT Press, second edition.

Chen, S., Dick, J., and Owen, A. B. (2011). Consistency of Markov chain quasi-Monte Carlo on continuous state spaces. *Ann. Statist.*, 39(2):673–701.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *J. Econometrics*, 115(2):293–346.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. American Statist. Assoc.*, 90(432):1313–1321.

Chib, S. (2011). Introduction to simulation and MCMC methods. *The Oxford Handbook of Bayesian Econometrics*, pages 183–217. OUP. Eds. Geweke, J., Koop, G. and van Dijk, H.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *American Statist.*, 49:327–335.

Chib, S. and Greenberg, E. (1996). Markov chain Monte Carlo simulation methods in econometrics. *Econometric Theory*, 12(3):409–431.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *J. American Statist. Assoc.*, 96(453):270–281.

Chib, S. and Ramamurthy, S. (2010). Tailored randomized block MCMC methods with application to DSGE models. *Journal of Econometrics*, 155(1):19 – 38.

Chib, S., Shin, M., and Simoni, A. (2018). Bayesian estimation and comparison of moment condition models. *J. American Statist. Assoc.*, 113(524):1656–1668.

Chib, S., Sin, M., and Tan, F. (2020). Estimating high-dimensional DSGE models with the TaRB-MH algorithm. Technical report, Olin Business School, Washington University.

Chopin, N. and Robert, C. (2010). Properties of nested sampling. *Biometrika*, 97(3):741–755.

Christen, J. and Fox, C. (2005). Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Statist.*, 14(4):795–810.

Clarté, G., Robert, C. P., Ryder, R., and Stoehr, J. (2020). Component-wise approximate Bayesian computation via Gibbs-like steps. *Biometriak*. To appear.

Craiu, R. V. and Meng, X.-L. (2005). Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. *Ann. Statist.*, 33(2):661–697.

Craiu, R. V. and Meng, X.-L. (2011). Perfection within reach: exact MCMC sampling. *Handbook of Markov Chain Monte Carlo*, pages 199–226. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Royal Statist. Society Series B*, 61(2):331–344.

Davis, P. and Rabinowitz, P. (1975). *Numerical Methods of Integration*. Academic Press, New York.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

De Bruijn, N. (1961). *Asymptotic methods in analysis. 2nd Edn., Sect. 4.2*. Amsterdam: North-Holland.

Dean, T. A., Singh, S. S., Jasra, A., and Peters, G. W. (2014). Parameter estimation for hidden Markov models with intractable likelihoods. *Scandinavian Journal of Statistics*, 41(4):970–987.

Dehideniya, M., Overstall, A. M., Drovandi, C. C., and McGree, J. M. (2019). A synthetic likelihood-based Laplace approximation for efficient design of biological processes. *https://arXiv:1903.04168*.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). The correlated pseudomarginal method. *J. Royal Statist. Society Series B*, 80(5):839–870.

Dellaportas, P. and Kontoyiannis, I. (2012). Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *J. Royal Statist. Society Series B*, 74(1):133–161.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Society Series B*, 56(2):363–375.

Dongarra, J. and Sullivan, F. (2000). Guest editors' introduction: The top 10 algorithms. *Computing in Science & Engineering*, 2(1):22–23.

Douc, R. and Robert, C. P. (2011). A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms. *Ann. Statist.*, 39(1):261–277.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte-Carlo sampling methods for Bayesian filtering. *Statist. Comp.*, 10:197–208.

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313.

Drovandi, C., Pettitt, A., and Faddy, M. (2011). Approximate Bayesian computation using indirect inference. *J. Royal Statist. Society Series A*, 60(3):503–524.

Drovandi, C. C., Pettitt, A. N., and Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statist. Science*, 30(1):72–95.

Dunson, D. and Johndrow, J. (2019). The Hastings algorithm at fifty. *Biometrika*, 107(1):1–23.

Everitt, R., Culliford, R., Medina-Aguayo, F., and Wilson, D. (2020). Sequential Monte Carlo with transformations. *Stat. Comput.*, 30:663–676.

Fan, Y. and Sisson, S. A. (2011). Reversible jump MCMC. *Handbook of Markov Chain Monte Carlo*, pages 67–92. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Fearnhead, P. (2011). MCMC for state-space models. *Handbook of Markov Chain Monte Carlo*, pages 513–529. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Fearnhead, P. (2018). Asymptotics of ABC. *Handbook of Approximate Bayesian Computation*, pages 269–288. Chapman & Hall/CRC. Eds. Sisson, S., Fan, Y., Beaumont, M.

Fearnhead, P., Bierkens, J., Pollock, M., and Roberts, G. O. (2018). Piecewise deterministic Markov processes for continuous-time Monte Carlo. *Statist. Sci.*, 33(3):386–412.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. Royal Statist. Society Series B*, 74(3):419–474. With discussion.

Feroz, F., Hobson, M. P., Cameron, E., and Pettitt, A. N. (2019). Importance nested sampling and the multinest algorithm. *The Open Journal of Astrophysics*, 2(1).

Fiel, N. and Wyse, J. (2012). Estimating the evidence: a review. *Statistica Neerlandica*, 66(3):288–308.

Fienberg, S. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, 1(1):1–40.

Fourment, M., Magee, A. F., Whidden, C., Bilge, A., Matsen, I., Frederick, A., and Minin, V. N. (2018). 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *https://arXiv:1811.11804*.

Frazier, D. T. (2020). Robust and efficient Approximate Bayesian Computation: A minimum distance approach. *arXiv preprint arXiv:2006.14126*.

Frazier, D. T. and Drovandi, C. (2019). Robust approximate Bayesian inference with synthetic likelihood. *https://arXiv:1904.04551*.

Frazier, D. T., Maneesoonthorn, W., Martin, G. M., and McCabe, B. P. (2019a). Approximate Bayesian forecasting. *Intern.*

J. Forecasting, 35(2):521–539.

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607.

Frazier, D. T., Nott, D. J., Drovandi, C., and Kohn, R. (2019b). Bayesian inference using synthetic likelihood: Asymptotics and adjustments. *https://arXiv:1902.04827*.

Frazier, D. T., Robert, C. P., and Rousseau, J. (2020). Model misspecification in approximate Bayesian computation: consequences and diagnostics. *J. Royal Statist. Society Series B*.

Frazier, P. (2018). A tutorial on Bayesian optimization. *https://arxiv.org/abs/1807.02811*.

Frigessi, A., Gasemyr, J., and Rue, H. (2000). Antithetic coupling of two Gibbs sampler chains. *The Annals of Statistics*, 28(4):1128–1149.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Ser. Anal.*, 15(2):183–202.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7(1):143–167.

Gallant, A. R. (2016). Reflections on the probability space induced by moment conditions with implications for Bayesian inference. *J. Finan. Econometrics*, 14(2):229–247.

Gallant, A. R. and Tauchen, G. (1996). Which moments to match? *Econometric theory*, 12(4):657–681.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Royal Statist. Society Series B*, 56(3):501–514.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398–409.

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Science*, 13:163–185.

Gelman, A. and Shirley, K. (2011). Inference from simulations and monitoring convergence. *Handbook of Markov chain Monte Carlo*, pages 163–174. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741.

George, E. (2000). The variable selection problem. *J. American Statist. Assoc.*, 95(452):1304–1308.

Gerber, M. and Chopin, N. (2015). Sequential quasi Monte Carlo. *J. Royal Statist. Society Series B*, 77(3):509–579.

Geweke, J. (1989a). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1340.

Geweke, J. (1989b). Exact predictive densities for linear models with ARCH disturbances. *J. Econometrics*, 40(1):63–86.

Geweke, J. (1999). Using simulation methods for Bayesian econometric models: Inference, development, and communication. *Econometric Reviews*, 18(1):1–73. With discussion.

Geweke, J., Koop, G., and van Dijk, H. (2011). *The Oxford Handbook of Bayesian Econometrics*. OUP.

Geyer, C. (1991). Markov chain Monte Carlo maximum likelihood. In Keramigas, E., editor, *Computing Science and Statistics: Proceedings of the 32rd Symposium on the Interface*, pages 156–163, Fairfax. Interface Foundation.

Geyer, C. (1993). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report 568, School of Statistics, Univ. of Minnesota.

Geyer, C. J. (2011a). Importance sampling, simulated tempering, and umbrella sampling. *Handbook of Markov chain Monte Carlo*, pages 295–311. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Geyer, C. J. (2011b). Introduction to Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo*, pages 3–48. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). On convergence of posterior distributions. *The Annals of Statistics*, 23(6):2145–2152.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press.

Giordani, P., Pitt, M., and Kohn, R. (2011). Bayesian inference for time series state space models. *The Oxford Handbook of Bayesian Econometrics*, pages 61–124. OUP. Eds. Geweke, J., Koop, G. and van Dijk, H.

Giummolè, F., Mameli, V., Ruli, E., and Ventura, L. (2017). Objective Bayesian inference with proper scoring rules. *TEST*, 28(3):1–28.

Glynn, P. W. (2016). Exact simulation vs. exact estimation. *Proceedings of the 2016 Winter Simulation Conference*, WSC16:193–205.

Glynn, P. W. and Rhee, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.*, 51(A):377–389.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Golightly, A., Henderson, D. A., and Sherlock, C. (2015). Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statist. Comp.*, 25(5):1039–1055.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gordon, N., Salmond, J., and Smith, A. (1993). A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing*, 140(2):107–113.

Gouriéroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *J. Applied Econometrics*, 8:85–118.

Gramacy, R., Samworth, R., and King, R. (2010). Importance tempering. *Statist. Comp.*, 20(1):1–7.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P., Latuszynski, K., Pereyra, M., and Robert, C. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statist. Comp.*, 25:835–862.

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. *Oxford Statistical Science Series*, 27:179–198.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97.

Gubernatis, J. E. (2005). Marshall Rosenbluth and the Metropolis algorithm. *Physics of Plasmas*, 12(5):057303.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*.

Gutmann, M. U. and Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *The Journal of Machine Learning Research*, 17(1):4256–4302.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(Feb):307–361.

Hajargasht, G. and Woźniak, T. (2018). Accurate computation of marginal data densities using variational Bayes. *https://arXiv:1805.10036*.

Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods*. John Wiley, New York.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57(1):97–109.

Higdon, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association*, 93(442):585–595.

Hitchcock, D. H. (2003). A history of the Metropolis–Hastings algorithm. *American Statist.*, 57(4):254–257.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Hoogerheide, L. F., van Dijk, H. K., and van Oest, R. D. (2009). Simulation based Bayesian econometric inference: principles and some recent computational advances. *Handbook of Computational Econometrics*, pages 215–280. John Wiley & Sons. Eds. van Dijk, H. and van Oest, R.

Hooper, M. (2013). Richard Price, Bayes' theorem, and God. *Significance*, 10(1):36–39.

Huber, M. L. (2016). *Perfect simulation*. Chapman & Hall/CRC.

Huggins, J. H., Kasprzak, M., Campbell, T., and Broderick, T. (2019). Validated variational inference via practical posterior error bounds. *https://arXiv:1910.04102*.

Huggins, J. H. and Miller, J. W. (2019). Using bagged posteriors for robust inference and model criticism. *https://arXiv:1912.07104*.

Iba, Y. (2000). Population-based Monte Carlo algorithms. *Trans. Japanese Soc. Artificial Intell.*, 16(2):279–286.

Jacob, P., Robert, C., and Smith, M. (2011). Using parallel computation to improve independent Metropolis–Hastings based estimation. *J. Comput. Graph. Statist.*, 20(3):616–635.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *J. Royal Statist. Society Series B*, 82:1–32. With discussion.

Jacquier, R., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models. *J. Business and Economic Statistics*, 12(4):371–389. With discussion.

Jahan, F., Ullah, I., and Mengersen, K. (2020). A review of Bayesian statistical approaches for Big Data. In Mengersen, K., Pudlo, P., and Robert, C., editors, *Case Studies in Applied Bayesian Science*, pages 17–44. Springer.

Jasra, A. (2015). Approximate Bayesian computation for a class of time series models. *International Statistical Review*, 83(3):405–435.

Jasra, A., Singh, S., Martin, J., and McCoy, E. (2012). Filtering via approximate Bayesian computation. *Statist. Comp.*, 22:1223–1237.

Jennings, E. and Madigan, M. (2017). AstroABC: an approximate Bayesian computation sequential Monte Carlo sampler for cosmological parameter estimation. *Astronomy and Computing*, 19:16–22.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data-mining. *Annals of Statistics*, 36(5):2207–2231.

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2019). MCMC for imbalanced categorical data. *J. American Statist. Assoc.*, 114(527):1394–1403.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1):article 26.

Kabisa, S., Dunson, D. B., and Morris, J. S. (2016). Online variational Bayes inference for high-dimensional correlated data. *J. Comput. Graph. Statist.*, 25(2):426–444.

Kass, R. and Raftery, A. (1995). Bayes factors. *J. American Statist. Assoc.*, 90:773–795.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393.

Kleijn, B. and van der Vaart, A. (2012). The Bernstein-von-Mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381.

Kloek, T. and van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica*, 46(1):1–19.

Knoblauch, J., Jewson, J., and Damoulas, T. (2019). Generalized variational inference. *https://arXiv:1904.02063*.

Kon Kam King, G., Canale, A., and Ruggiero, M. (2019). Bayesian functional forecasting with locally-autoregressive dependent processes. *Bayesian Anal.*, 14(4):1121–1141.

Koop, G. and Korobilis, D. (2018). Variational Bayes inference in high-dimensional time-varying parameter models. *SSRN 3246472*.

Koop, G. M. (2003). *Bayesian Econometrics*. John Wiley & Sons Inc.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika*, 90:319–326.

Lemieux, C. (2009). *Monte Carlo and quasi-Monte Carlo sampling*. Springer Science & Business Media.

Li, W. and Fearnhead, P. (2018a). Convergence of regression-adjusted approximate Bayesian computation. *Biometrika*, 105(2):301–318.

Li, W. and Fearnhead, P. (2018b). On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika*, 105(2):285–299.

Lindley, D. (1980). Approximate Bayesian methods. In *Bayesian statistics: Proceedings of the first international meeting held in Valencia (Spain), May 28 to June 2, 1979*.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic biology*, 66(1):e66–e82.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Verlag, New-York.

Liu, J. S., Liang, F., and Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. American Statist. Assoc.*, 95(449):121–134.

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2020). Focused Bayesian prediction. *Forthcoming. Journal of Applied Econometrics*.

Luo, X. and Tjelmeland, H. (2019). A multiple-try Metropolis-Hastings algorithm with tailored proposals. *Computational Statistics*, 34(3):1109–1133.

Lyddon, S., Holmes, C., and Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478.

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., and Simpson, D. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statist. Sci.*, 30(4):443–467.

Marin, J., Pudlo, P., Robert, C., and Ryder, R. (2011). Approximate Bayesian computational methods. *Statist. Comp.*, 21(2):279–291.

Marin, J. and Robert, C. (2011). Importance sampling methods for Bayesian discrimination between embedded models. In Chen, M.-H., Dey, D., Müller, P., Sun, D., and Ye, K., editors, *Frontiers of Statistical Decision Making and Bayesian Analysis*. Springer-Verlag, New York.

Marin, J.-M., Mengersen, K., and Robert, C. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, pages 459–507. Elsevier. Eds. Rao, C. and Dey, D.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19(6):451–458.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100(26):15324–15328.

Martin, G. M., McCabe, B. P., Frazier, D. T., Maneesoonthorn, W., and Robert, C. P. (2019). Auxiliary likelihood-based approximate Bayesian computation in state space models. *J. Comput. Graph. Statist.*, 28(3):508–522.

Martino, L. (2018). A review of multiple try MCMC algorithms for signal processing. *Digital Signal Processing*, 75:134–152.

Martino, S. and Riebler, A. (2019). Integrated nested Laplace approximations (INLA). *https://arXiv:1907.01248*.

Meng, X. and Schilling, S. (2002). Warp bridge sampling. *J. Comput. Graph. Statist.*, 11(3):552–586.

Meng, X. and Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica*, 6:831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2018). Unbiased Markov chain Monte Carlo for intractable target distributions. *https://arXiv:1807.08691*.

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *J. American Statist. Assoc.*, 114(527):1113–1125.

Mukherjee, P., Parkinson, D., and Liddle, A. (2006). A nested sampling algorithm for cosmological model selection. *The Astrophysical Journal*, 638(2):L51–L54.

Muller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849.

Naesseth, C. A., Lindsten, F., Schön, T. B.,*et al.* (2019). Elements of sequential Monte Carlo. *Foundations and Trends in Machine Learning*, 12(3):307–392.

Naylor, J. and Smith, A. (1982). Application of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31(3):214–225.

Neal, R. (1994). Contribution to the discussion of "Approximate Bayesian inference with the weighted likelihood bootstrap" by Michael A. Newton and Adrian E. Raftery. *J. Royal Statist. Society Series B*, 56(1):41–42.

Neal, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statist. Comp.*, 6:353–366.

Neal, R. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. *Learning in Graphical Models*, pages 205–228. Chapman & Hall/CRC. Ed. Jordan, M. I.

Neal, R. (1999). Erroneous results in 'Marginal likelihood from the Gibbs output'. Technical report, University of Toronto.

Neal, R. (2001). Annealed importance sampling. *Stat. Comput.*, 11(2):125–139.

Neal, R. (2003). Slice sampling. *Ann. Statist.*, 31(3):705–767. With discussion.

Neal, R. (2011a). MCMC using ensembles of states for problems with fast and slow variables such as Gaussian process

regression. *https://arXiv:1101.0387*.

Neal, R. (2011b). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Neiswanger, W., Wang, C., and Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. *https://arXiv:1311.4780*.

Nemeth, C. and Fearnhead, P. (2019). Stochastic gradient Markov chain Monte Carlo. *https://arXiv:1907.06986*.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Royal Statist. Society Series B*, 56(1):3–48.

Nguyen, H. D., Arbel, J., Lü, H., and Forbes, F. (2020). Approximate Bayesian computation via the energy statistic. *IEEE Access*, 8:131683–131698.

Nott, D., Ong, V. M.-H., Fan, Y., and Sisson, S. (2018). High-dimensional ABC. *Handbook of Approximate Bayesian Computation*, pages 211–242. Chapman & Hall/CRC. Eds. Sisson, S., Fan, Y., Beaumont, M.

Nott, D. J. and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika*, 92(4):747–763.

Ong, V. M., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2018a). Variational Bayes with synthetic likelihood. *Statist. Comp.*, 28(4):971–988.

Ong, V. M.-H., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2018b). Likelihood-free inference in high dimensions with synthetic likelihood. *Computational Statistics & Data Analysis*, 128:271–291.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *American Statist.*, 64(2):140–153.

Owen, A. B. (2017). Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Statist.*, 26(3):738–744.

Park, M. and Nassar, M. (2014). Variational Bayesian inference for forecasting hierarchical time series. In *Divergence Methods in Probabilistic Inference (DMPI) workshop at International Conference on Machine Learning (ICML)*. Citeseer.

Peters, G. W., Sisson, S. A., and Fan, Y. (2012). Likelihood-free Bayesian inference for α-stable models. *Comput. Statist. Data Anal.*, 56(11):3743–3756.

Pettenuzzo, D. and Ravazzolo, F. (2016). Optimal portfolio choice under decision-based model combinations. *Journal of Applied Econometrics*, 31(7):1312–1332.

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics*, 171(2):134–151.

Pollock, M., Fearnhead, P., Johansen, A. M., and Roberts, G. O. (2020). Quasi-stationary Monte Carlo and the ScaLE algorithm. *J. Royal Statist. Society Series B*, 82(5):1167–1221. (With discussion.).

Polson, N. G., Carlin, B. P., and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. American Statist. Assoc.*, 87(418):493–500.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *J. Comput. Graph. Statist.*, 27(1):1–11.

Price, R. (1764). A demonstration of the second rule in the essay towards the solution of a problem in the doctine of chances. *Philosophical Transactions of the Royal Society of London*, 54:296–325.

Priddle, J. W., Sisson, S. A., Frazier, D. T., and Drovandi, C. (2019). Efficient Bayesian synthetic likelihood with whitening transformations. *arXiv preprint arXiv:1909.04857*.

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.*, 16:1791–1798.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, volume 9, pages 223–252.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *J. American Statist. Assoc.*, 114(526):831–843.

Quiroz, M., Nott, D. J., and Kohn, R. (2018a). Gaussian variational approximation for high-dimensional state space models. *https://arXiv:1801.07873*.

Quiroz, M., Tran, M.-N., Villani, M., and Kohn, R. (2018b). Speeding up MCMC by delayed acceptance and data subsampling. *J. Comput. Graph. Statist.*, 27(1):12–22.

Raftery, A. (1996). Hypothesis testing and model selection. In Gilks, W., Spiegelhalter, D., and Richardson, S., editors,

Markov Chain Monte Carlo in Practice, pages 163–188. Chapman and Hall, New York, London.

Rischard, M., Jacob, P. E., and Pillai, N. (2018). Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation.

Ritter, C. and Tanner, M. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *J. American Statist. Assoc.*, 87(419):861–868.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition.

Robert, C. (2007). *The Bayesian Choice*. Springer-Verlag, New York.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition.

Robert, C. and Casella, G. (2011). A history of Markov chain Monte Carlo—subjective recollections from incomplete data. *Statist. Science*, 26(1):102–115.

Robert, C. P. (2015). *The Metropolis–Hastings Algorithm*, pages 1–15. Wiley StatsRef: Statistics Reference Online. American Cancer Society.

Robert, C. P., Elvira, V., Tawn, N., and Wu, C. (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1435.

Roberts, G., Gelman, A., and Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Applied Prob.*, 7(1):110–120.

Roberts, G. and Rosenthal, J. (1999). Convergence of slice sampler Markov chains. *J. Royal Statist. Society Series B*, 61(3):643–660.

Roberts, G. and Sahu, S. (1997). Updating schemes, covariance structure, blocking and parametrisation for the Gibbs sampler. *J. Royal Statist. Society Series B*, 59(2):291–317.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. Royal Statist. Society Series B*, 60(1):255–268.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.*, 18(2):349–367.

Roberts, G. O., Tweedie, R. L.,*et al.* (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Rodrigues, G., Nott, D. J., and Sisson, S. (2019). Likelihood-free approximate Gibbs sampling. *https://arXiv:1906.04347*.

Rosenthal, J. S. (2011). Optimal proposal distributions and adaptive MCMC. *Handbook of Markov Chain Monte Carlo*, pages 93–111. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.*, 9(1):130–134.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *J. Royal Statist. Society Series B*, 71(2):319–392.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88.

Sisson, S. and Fan, Y. (2011). Likelihood-free Markov chain Monte Carlo. *Handbook of Markov Chain Monte Carlo*, pages 313–333. Chapman & Hall/CRC. Eds. Brooks, S., Gelman, A., Jones, G., Meng, X-L.

Sisson, S. A., Fan, Y., and Beaumont, M. (2019). *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC.

Sisson, S. A., Fan, Y., and Tanaka, M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104(6):1760–1765.

Skilling, J. (2007). Nested sampling for Bayesian computations. *Bayesian Analysis*, 1(4):833–859.

Smith, A. and Roberts, G. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Royal Statist. Society Series B*, 55(1):3–24. With discussion.

Stigler, S. (1986a). *The History of Statistics*. Belknap, Cambridge.

Stigler, S. (1986b). Memoir on inverse probability. *Statistical Science*, 1(3):359–363.

Stigler, S. M. (1975). Studies in the history of probability and statistics. XXXIV Napoleonic statistics: The work of Laplace.

Biometrika, 62(2):503–517.

Stigler, S. M. (2018). Richard Price, the first Bayesian. *Statist. Sci.*, 33(1):117–125.

Stoehr, J. (2017). A review on statistical inference methods for discrete Markov random fields. *https://arXiv:1704.03331*.

Strachan, R. and van Dijk, H. (2014). Divergent priors and well behaved Bayes factors. *Central European Journal of Economic Modelling and Econometrics, CEJEME*, 6(1):1–31.

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486.

Tanner, M. A. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *J. American Statist. Assoc.*, 82(398):528–550. With discussion.

Tavaré, S., Balding, D., Griffith, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518.

Tawn, N. G., Roberts, G. O., and Rosenthal, J. S. (2020). Weight-preserving simulated tempering. *Statist. Comp.*, 30:27–41.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1762. With discussion and a rejoinder by the author.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *J. American Statist. Assoc.*, 81(393):82–86.

Tierney, L., Kass, R., and Kadane, J. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions. *J. American Statist. Assoc.*, 84(407):710–716.

Tierney, L. and Mira, A. (1998). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, 18:2507–2515.

Tokdar, S. and Kass, R. (2010). Importance sampling: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:54 – 60.

Tran, K. T. (2018). A common derivation for Markov chain Monte Carlo algorithms with tractable and intractable targets. *arXiv:*, 1607.01985.

Tran, M.-N., Nguyen, D. H., and Nguyen, D. (2019). Variational Bayes on manifolds. *https://arXiv:1908.03097*.

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). Variational Bayes with intractable likelihood. *J. Comput. Graph. Statist.*, 26(4):873–882.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vanslette, K., Alsheikh, A. A., and Youcef-Toumi, K. (2019). Why simple quadrature is just as good as Monte Carlo. *https://arXiv:1908.00947*.

Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *J. American Statist. Assoc.*, 112(517):137–168.

Wang, X. and Dunson, D. (2013). Parallel MCMC via Weierstrass sampler. *https://arXiv:1312.4605*.

Wang, Y. and Blei, D. (2019a). Variational Bayes under model misspecification. In *Advances in Neural Information Processing Systems*, pages 13357–13367.

Wang, Y. and Blei, D. M. (2019b). Frequentist consistency of variational Bayes. *J. American Statist. Assoc.*, 114(527):1147–1161.

Wilkinson, R. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2):129–141.

Wiqvist, S., Picchini, U., Forman, J. L., Lindorff-Larsen, K., and Boomsma, W. (2018). Accelerating delayed-acceptance Markov chain Monte Carlo algorithms. *https://arXiv:1806.05982*.

Wood, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102—1104.

Wood, S. (2019). Simplified integrated nested Laplace approximation. *Biometrika*, 107(1):223–230.

Yu, X., Nott, D. J., Tran, M.-N., and Klein, N. (2019). Assessment and adjustment of approximate inference algorithms using the law of total variance. *https://arXiv:1911.08725*.

Zanella, G. and Roberts, G. (2019). Scalable importance tempering and Bayesian variable selection. *J. Royal Statist. Society Series B*, 81(3):489–517.

Zellner, A. (1971). *An Introduction to Bayesian Econometrics*. John Wiley, New York.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2018). Advances in variational inference. *IEEE transactions on*

pattern analysis and machine intelligence, 41(8):2008–2026.

Zhang, F. and Gao, C. (2017). Convergence rates of variational posterior distributions. *https://arXiv:1712.02519*.

Zhang, T. (2006a). From eps-entropy to KL entropy: analysis of minimum information complexity density estimation. *Annals of Statistics*, 34:2180–2210.

Zhang, T. (2006b). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Info. Theory*, 52(4):1307–1321.