MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Approximating Bayes in the 21st Century

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

December 2021

Working Paper 24/21

AACSB ACCREDITED EQUIS ACCREDITED AMBA ACCREDITED

Approximating Bayes in the 21st Century[*]

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

December 20, 2021

Abstract

The 21st century has seen an enormous growth in the development and use of approximate Bayesian methods. Such methods produce computational solutions to certain 'intractable' statistical problems that challenge exact methods like Markov chain Monte Carlo: for instance, models with unavailable likelihoods, high-dimensional models, and models featuring large data sets. These approximate methods are the subject of this review. The aim is to help new researchers in particular – and more generally those interested in adopting a Bayesian approach to empirical work – distinguish between different approximate techniques; understand the sense in which they are approximate; appreciate when and why particular methods are useful; and see the ways in which they can can be combined.

Keywords: Approximate Bayesian inference; intractable Bayesian problems; approximate Bayesian computation; Bayesian synthetic likelihood; variational Bayes; integrated nested Laplace approximation.

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

[*] The authors would 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. That earlier version, comprising a more extensive coverage of Bayesian computation *per se*, and an historical timeline for computational developments, appears under the title 'Computing Bayes: Bayesian Computation from 1763 to the 21st Century' (Martin *et al.*, 2020). Martin and Frazier have been supported by Australian Research Council Discovery Grants DP170100729 and DP200101414, 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).

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

1 Introduction

The advent of fast, accessible computers in the last two decades of the 20th century (Ceruzzi, 2003), allied with the exploitation of earlier insights into probabilistic simulation (Metropolis and Ulam, 1949; Metropolis *et al.*, 1953; Hammersley and Handscomb, 1964; Hastings, 1970), led to an explosion in the use of simulation-based computation to solve empirical Bayesian problems. Whilst significant advances were made in econometrics (Kloek and van Dijk, 1978; Bauwens and Richard, 1985; Geweke, 1989) and signal processing (Gordon *et al.*, 1993) using the principles of importance sampling, the 'computational revolution' – as it is often coined – was driven primarily by Markov chain Monte Carlo (MCMC) algorithms; see Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990) for seminal contributions, and Besag and Green (1993), Smith and Roberts (1993), Chib (2011), Geyer (2011) and Robert and Casella (2011) for selected reviews.

The impact of these computational advances was felt across a huge array of fields – genetics, biology, neuroscience, astrophysics, image analysis, ecology, epidemiology, engineering, education, economics, political science, marketing and finance, to name but some – and brought Bayesian analysis into the statistical mainstream. The three handbooks: *The Oxford Handbook of Applied Bayesian Analysis* (O'Hagan and West, 2010), *Handbook of Markov Chain Monte Carlo* (Brooks *et al.*, 2011) and *The Oxford Handbook of Bayesian Econometrics* (Geweke *et al.*, 2011), highlight the wide spectrum of fields, and broad scope of empirical problems to which MCMC and importance sampling 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 graphical processing units, and so forth). In short, the early algorithms were to stumble in the face of so-called 'intractable' statistical problems: data generating processes with likelihoods that are unavailable analytically; models with a very large number of unknowns; and models featuring 'big data'. 'Exact' solutions to such problems were simply not achievable via the early MCMC and importance sampling algorithms or, at least, were not available in a reasonable computing time; and 'approximate' solutions were, instead, often sought. It is those approximate solutions that are the subject of this review.[1]

Our overarching aim is to provide readers with some insight into questions such as: 'In what sense are approximate methods of computation *'approximate'*?', 'What are the connections between different approximate methods?', 'When does one use one approach, and when another?', and 'Can different methods be combined to tackle multiple, distinct instances of 'intractability'?'. In order to address such questions, we bring together in one place, and using a common notational framework, the four main

[1] We acknowledge, of course, that there have been *many* concurrent advances in MCMC and importance sampling designed, in particular, to deal with the problem of scale. We refer the reader to: Green *et al.* (2015), Robert *et al.* (2018) and Dunson and Johndrow (2019) for broad overviews of modern developments in MCMC; to Betancourt (2018) for a review of Hamiltonian Monte Carlo (HMC); to Naesseth *et al.* (2019) for a recent review of sequential Monte Carlo (exploiting importance sampling principles as it does); and to Hoogerheide *et al.* (2009), Tokdar and Kass (2010) and Elvira and Martino (2021) for other advances in importance sampling.

approximate techniques that have evolved during the 21st century: 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, with illustrative examples from the literature used to enhance this demonstration. This then provides some context for the *hybrid* computational methods that we then review. Whilst formally providing an exact solution and, hence, not a focus of this paper, we give a brief outline of pseudo-marginal methods, including particle MCMC (PMCMC), due to the role such methods play – in tandem with certain approximate techniques – under the 'hybrid' umbrella.

This paper is meant to serve both as a 'first port of call' for those who are new to modern Bayesian computation, and as a useful overview for practitioners with established, but selective, expertise. Hence, excessive formalism, and extensive algorithmic detail, is avoided in order to make the paper as accessible as possible, and to keep the focus on the key principles underpinning each computational method. We do not attempt to replicate the coverage of existing reviews of specific approximate methods. Rather, we direct readers to those review papers and handbook chapters where necessary, including for coverage of all published work. Whilst we make brief reference to various software packages, we also defer to those other resources for detailed descriptions of the dedicated software that is available for implementing particular computational techniques.

The remainder of the paper is as follows. In Section 2 we provide a brief outline of the general Bayesian computational problem, and explain when that problem may be viewed as intractable. Section 3 then outlines the approximate solutions to such intractable problems, ABC, BSL, VB and INLA, and hybridized versions thereof. For ease of exposition – and with acknowledgement that this categorization is imperfect – these main methods are grouped into: 'Simulation-based Approaches' (ABC and BSL) and 'Optimization Approaches' (VB and INLA). Several hybrid approximate methods are then described, in which distinct methods are amalgamated for the purpose of simultaneously solving multiple computational challenges (e.g., a high-dimensional model with a computationally expensive, or analytically unavailable likelihood). In order to illustrate the type of intractable problems for which approximate solutions have been sought, we display the results of selected empirical illustrations from the literature in which, respectively, simulation-based computation and optimization-based computation have been used. The paper concludes in Section 4 with some perspectives on the future. Particular attention is given to three directions that the authors believe to be worthy of attention: *1)* The performance of approximate methods under (model) misspecification; *2)* The use of approximate methods in generalized (non-likelihood) settings; and, finally, *3)* The role to be played by approximate inference in Bayesian prediction.

2 Bayesian Computation in a Nutshell

2.1 A short primer

We being by establishing notation. An n-dimensional vector of observed data $\mathbf{y} = (y_1, y_2, ..., y_n)'$ is assumed to be generated from some data generating process (DGP) $p(\mathbf{y}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ a p-dimension vector of unknown parameters, and where we possess prior beliefs on $\boldsymbol{\theta}$ specified by the prior probability density function (pdf) $p(\boldsymbol{\theta})$. By Bayes' rule, the joint posterior pdf $p(\boldsymbol{\theta}|\mathbf{y})$ is defined by

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1}$$

Most Bayesian quantities of interest are, in turn, posterior expectations of some function $g(\boldsymbol{\theta})$ and, hence, can be expressed as,

$$\mathbb{E}(g(\boldsymbol{\theta})|\mathbf{y}) = \int_\Theta g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \tag{2}$$

Familiar examples include posterior moments, such as

$$\mathbb{E}(\boldsymbol{\theta}|\mathbf{y}) = \int_\Theta \boldsymbol{\theta}p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \text{ and } \mathrm{Var}(\boldsymbol{\theta}|\mathbf{y}) = \int_\Theta \left[\boldsymbol{\theta} - \mathbb{E}(\boldsymbol{\theta}|\mathbf{y})\right]\left[\boldsymbol{\theta} - \mathbb{E}(\boldsymbol{\theta}|\mathbf{y})\right]' p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta},$$

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, (2) 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 (2) defines the *predictive* distribution for y_{n+1}:

$$p(y_{n+1}^*|\mathbf{y}) = \int_\Theta p(y_{n+1}^*|\boldsymbol{\theta}, \mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \tag{3}$$

It also encompasses $g(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, d)$, for $L(\boldsymbol{\theta}, d)$ a loss function associated with a decision d, in which case (2) is the quantity minimized in Bayesian decision theory (Berger, 1985; Robert, 2001). Further, defining $g(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})$ as the DGP that explicitly conditions on a model, $\mathcal{M}$ say, the marginal likelihood of $\mathcal{M}$, $p(\mathbf{y}|\mathcal{M})$, is the expectation,

$$\mathbb{E}(g(\boldsymbol{\theta})|\mathcal{M}) = \int_\Theta g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}, \tag{4}$$

with respect to the prior, $p(\boldsymbol{\theta}|\mathcal{M})$. The ratio of (4) to the comparable quantity for an alternative model $\mathcal{M}'$ defines the Bayes factor for use in choosing between the two models. In summary then, the key quantities that underpin the whole of 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 (2) and (4) are rare. 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 the constant of proportionality. The availability of $p(\boldsymbol{\theta}|\mathbf{y})$ only up to the integrating constant immediately precludes the analytical solution of (2), for any $g(\boldsymbol{\theta})$. By definition, a lack of knowledge of the integrating constant implies that the marginal likelihood for the model in (4) is unavailable. Hence the need for computational solutions.

It is useful to think about all Bayesian computational techniques falling into one of three broad categories: *1) Deterministic integration methods; 2) Exact simulation methods; 3) Approximate (including asymptotic) methods.* Whilst all techniques are applicable to both the posterior expectation in (2) and the prior expectation in (4), to keep the scope of the paper manageable we only consider the computation of (2).[2] 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 (2) 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 (see Davis and Rabinowitz, 1975, Naylor and Smith, 1982, Vanslette *et al.*, 2019, and Bilodeau *et al.*, 2021, for relevant coverage). Such methods remain an important tool in the Bayesian arsenal, and we note the recent explosion of probabilistic numerics creating new connections between Bayesian concepts and numerical integration (Briol *et al.*, 2019). However, deterministic integration – on its own – plays a relatively small role in Bayesian numerical work due primarily to the 'curse of dimensionality' from which it suffers.

The methods in *2)* use simulation to produce M draws of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, ..., M$, from $p(\boldsymbol{\theta}|\mathbf{y})$, with the mean of the M transformed draws, $g(\boldsymbol{\theta}^{(i)})$, often used to estimate (2). Different simulation methods are distinguished by the way in which the draws are produced, and whether those draws are independent (e.g. Monte Carlo simulation; importance sampling) or dependent (e.g., MCMC). However, under appropriate regularity, like finite variance, and subject to convergence in the case of MCMC, all such methods produce a $\sqrt{M}$-consistent estimate of (2), whatever the degree of dependence in the draws, with the dependence affecting the constant implicit in the $O(M^{-1/2})$ term, but not the rate itself (Geyer, 2011). Hence, in principle, any algorithm that simulates from $p(\boldsymbol{\theta}|\mathbf{y})$ can produce an estimate of (2) that is arbitrarily accurate for large enough M; justifying the use of the adjective 'exact'.

Finally, the methods in *3)* replace the integrand in (2) 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. *Asymptotic* approximation methods replace the integrand with an expansion that is accurate for large n, and yield an estimate of (2) that is accurate asymptotically.

It is the class of approximate methods in *3)* that is our focus, with our first task being to establish why *these* methods, rather than those in *2)*, are useful in intractable settings.

2.2 Intractable Bayesian problems

With reference to (1), two characteristics are worthy of note. First, as is common knowledge, in all but the most stylized problems (for example, when $p(\mathbf{y}|\boldsymbol{\theta})$ is from the exponential family, and either a natural conjugate, or convenient noninformative prior is adopted), $p(\boldsymbol{\theta}|\mathbf{y})$ 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 $p(\boldsymbol{\theta})$. 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 importance sampling methods obviate the first problem by drawing *indirectly* from $p(\boldsymbol{\theta}|\mathbf{y})$ via another ('candidate' or 'proposal') distribution from which simulation is feasible. However, these methods still require evaluation

[2] See Ardia *et al.* (2012) and Llorente *et al.* (2021) for extensive reviews of methods for computing marginal likelihoods.

of $p(\mathbf{y}|\boldsymbol{\theta})$: in the computation of the importance weights in importance sampling (Geweke, 1989; Tokdar and Kass, 2010), in the computation of the acceptance probability in any Metropolis-Hastings MCMC algorithm (Hastings, 1970; Chib and Greenberg, 1995), and in the implementation of any Gibbs-based MCMC algorithm, in which the conditional posteriors are required either in full form or at least up to a scale factor (Casella and George, 1992; Chib and Greenberg, 1996).[3]

The assumption that $p(\mathbf{y}|\boldsymbol{\theta})$ can be evaluated is a limitation for two reasons. First, some DGPs do not admit pdfs (or probability mass functions) 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 original MCMC and importance sampling methods are not readily *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). A prime example of this is when the vector of unknowns, $\boldsymbol{\theta}$, comprises both a set of fixed 'global' parameters that govern all the data (call this set $\boldsymbol{\phi}$), and a set of latent random variables that are 'local' to individual data points (call this set $\mathbf{x}$), and where $\mathbf{x}$ is very large – sometimes of dimension exceeding n (e.g. Tavaré *et al.*, 1997; Rue *et al.*, 2009; Beaumont, 2010; Braun and McAuliffe, 2010; Lintusaari *et al.*, 2017; Johndrow *et al.*, 2019). In such cases, standard MCMC methods – even if feasible in principle – may not enable an accurate estimate of (2) to be produced in finite computing time; i.e. such methods are *not necessarily scalable* in the dimension of the unknowns (Betancourt, 2018). We alert the reader to the fact that the term 'intractable likelihood' is sometimes used to refer to such cases, since the likelihood for the global parameters, $p(\mathbf{y}|\boldsymbol{\phi})$ – which requires integration over the latent parameters – is typically not available in closed form, even when the 'complete' likelihood, $p(\mathbf{y},\mathbf{x}|\boldsymbol{\phi}) = p(\mathbf{y}|\mathbf{x},\boldsymbol{\phi})p(\mathbf{x}|\boldsymbol{\phi})$, *is* available. We reserve the term 'intractable' or 'unavailable' likelihood for the case where the DGP cannot be expressed in closed form. The intractability that arises in problems with a large number of latent variables we view as simply an example of the computational difficulties that arise when $\boldsymbol{\theta}$ is of high dimension.

The techniques discussed in Section 3 relieve the investigator of one or more of these instances of intractability: unavailable likelihood; high-dimensional $\boldsymbol{\theta}$; 'big' $\mathbf{y}$. But they come at a cost. All such methods produce an estimate of (2) that is only ever intrinsically approximate.

3 Approximate Bayesian Methods

As noted above, the goal of all exact simulation-based computational methods (including the pseudo-marginal techniques that play a role in the hybrid approximation methods discussed in Section 3.3), is to

[3] Some versions of these methods only require a term of $p(\mathbf{y}|\boldsymbol{\theta})$ to be available or allow for its replacement by an unbiased estimate, as in pseudo-marginal MCMC; see Section 3.3.2.

estimate the posterior expectation in (2) 'exactly', at least up to some $O(M^{-1/2})$ term, 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. Nevertheless, the convention in the literature is to refer to all simulation methods outlined to this point as *exact*, typically without qualification.[4]

In contrast, when applying an *approximate* method (using the taxonomy in Section 2.1), investigators make no claim to exactness, other than citing the asymptotic (in n) accuracy of the Laplace approximation-based methods (Tierney and Kadane, 1986; Rue *et al.*, 2009), or the asymptotic validity of certain other approximations (Fearnhead, 2018; Frazier *et al.*, 2018; Frazier *et al.*, 2019b; Zhang and Gao, 2020). That is, for finite n at least, such methods are only ever acknowledged as providing an approximation to (2), with that approximation perhaps claimed to be as accurate as possible, given the relevant choice variables that characterize the method; but no more than that.

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 – in which neither $p(\boldsymbol{\theta}|\mathbf{y})$ *nor* $p(\mathbf{y}|\boldsymbol{\theta})$ is available analytically; problems that would otherwise not be amenable to Bayesian analysis. In computing (2), both methods replace the posterior in the integrand, $p(\boldsymbol{\theta}|\mathbf{y})$, with an approximate posterior based solely on *simulation* from $p(\mathbf{y}|\boldsymbol{\theta})$ (and $p(\boldsymbol{\theta})$). A simulation-based estimate of (2), $\overline{g(\boldsymbol{\theta})} = (1/M)\sum_{i=1}^{M} g(\boldsymbol{\theta}^{(i)})$, is then produced using draws, $\boldsymbol{\theta}^{(i)}$, from this approximate posterior. In contrast, VB and INLA both require evaluation of $p(\mathbf{y}|\boldsymbol{\theta})$, but reap computational benefits in certain types of problems (in particular those of high-dimension and/or based on huge data sets) 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 (2) is replaced by an approximation produced via 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 (2) can be computed with a combination of optimization and low-dimensional deterministic integration steps.

3.1 Simulation-based approaches

3.1.1 Approximate Bayesian computation (ABC)

From its initial beginnings as a practical approach for inference in population genetics models with computationally expensive 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 18,000 citations garnered on Google Scholar since 2000. As such, not only do several reviews of the area exist (e.g. Marin *et al.*, 2011; Sisson and Fan, 2011; Lintusaari *et al.*, 2017; Beaumont, 2019), but the technique has recently reached 'handbook status', with the publication of Sisson *et al.* (2019);

[4] We note that so-called 'quasi-Monte Carlo' integration schemes 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.

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 is given in Algorithm 1, where $d\{\cdot,\cdot,\}$ denotes a generic metric and $\varepsilon > 0$ a pre-specified or post-processing tolerance parameter:

Algorithm 1 Vanilla Accept/Reject ABC Algorithm

```
for $i = 1, \dots, M$ do
    Simulate $\boldsymbol{\theta}^i$, $i = 1, 2, ..., M$, from $p(\boldsymbol{\theta})$, and artificial data $\mathbf{z}^i$ from $p(\cdot|\boldsymbol{\theta}^i)$;
    Accept $\boldsymbol{\theta}^i$ if $d\{\mathbf{z}^i, \mathbf{y}\} \leq \varepsilon$
end for
```

An accepted $\boldsymbol{\theta}^i$ is a draw from the posterior:

$$p_\varepsilon(\boldsymbol{\theta}|\mathbf{y}) = \frac{\int_{\mathcal{X}} p_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) \mathrm{d}\mathbf{z}}{\int_\Theta \int_{\mathcal{X}} p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) \mathrm{d}\mathbf{z} \mathrm{d}\boldsymbol{\theta}}, \quad p_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \mathbb{I}\left[d\{\mathbf{y}, \mathbf{z}\} \leq \varepsilon\right] p(\mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

where $\mathbb{I}\left[\cdot\right]$ denotes the indicator function. Under regularity conditions, it can be shown that $\lim_{\varepsilon \to 0} p_\varepsilon(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{y})$. However, in practice the choice of $\varepsilon = 0$ is infeasible since if $\mathbf{y}$ is continuous, the event $\mathbf{z} = \mathbf{y}$ has zero probability. More generally, for a fixed computing budget, ensuring that $\alpha_n = \Pr\{d\{\mathbf{z}, \mathbf{y}\} \leq \varepsilon\}$ is non-negligible in practice for small ε is infeasible as n increases. Thus, unless we have very few observations, or are working with discrete data, Algorithm 1 cannot be implemented in anything but toy problems.

ABC using summary statistics

Since comparing high-dimensional Euclidean vectors $\mathbf{z}$ and $\mathbf{y}$ is computationally infeasible, the vast majority of ABC applications first degrade the datasets down to a vector of lower-dimensional statistics, customarily called *summary statistics.* Define $\eta : \mathcal{X} \to \mathcal{B} \subseteq \mathbb{R}^{k_\eta}$ as a summary statistic mapping. In general then, Algorithm 1 is replaced with:

Algorithm 2 Accept/Reject ABC Algorithm Based on Summary Statistics

```
for $i = 1, \dots, M$ do
    Simulate $\boldsymbol{\theta}^i$, $i = 1, 2, ..., M$, from $p(\boldsymbol{\theta})$, and artificial data $\mathbf{z}^i$ from $p(\cdot|\boldsymbol{\theta}^i)$;
    Accept $\boldsymbol{\theta}^i$ if $d\{\eta(\mathbf{z}^i), \eta(\mathbf{y})\} \leq \varepsilon$.
end for
```

In this more common formulation, 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. Consequently, since $\varepsilon > 0$, accepted draws from Algorithm 2 can only be seen as draws from the posterior $p_{\varepsilon}(\boldsymbol{\theta}|\eta(\boldsymbol{y}))$, which is an approximation to the 'partial' posterior $p(\boldsymbol{\theta}|\eta(\boldsymbol{y}))$ that results from using a non-zero tolerance.[5]

The accuracy of the posterior output by Algorithm 2 can be understood via the decomposition:

$$\begin{aligned}\int_{\Theta} |p_{\varepsilon}(\boldsymbol{\theta}|\eta(\mathbf{y})) - p(\boldsymbol{\theta}|\mathbf{y})|\mathrm{d}\boldsymbol{\theta} &= \int_{\Theta} |p_{\varepsilon}(\boldsymbol{\theta}|\eta(\mathbf{y})) - p(\boldsymbol{\theta}|\eta(\mathbf{y})) + p(\boldsymbol{\theta}|\eta(\mathbf{y})) - p(\boldsymbol{\theta}|\mathbf{y})|\mathrm{d}\boldsymbol{\theta} \\ &\leq \int_{\Theta} |p_{\varepsilon}(\boldsymbol{\theta}|\eta(\mathbf{y})) - p(\boldsymbol{\theta}|\eta(\mathbf{y}))|\mathrm{d}\boldsymbol{\theta} + \int_{\Theta} |p(\boldsymbol{\theta}|\eta(\mathbf{y})) - p(\boldsymbol{\theta}|\mathbf{y})|\mathrm{d}\boldsymbol{\theta}. \qquad (5)\end{aligned}$$

The first term in (5) captures the discrepancy between the partial posterior we *wish* to target, i.e., $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$, and the posterior that is targeted by Algorithm 2, i.e., $p_{\varepsilon}(\boldsymbol{\theta}|\eta(\mathbf{y}))$. The second term measures the discrepancy that results from the use of summary statistics $\eta(\mathbf{y})$ that are (most likely) insufficient for the data $\mathbf{y}$.

With regard to the second term in (5), the difference is characterized by the informativeness, or otherwise, of the chosen summaries. Under regularity conditions that ensure both posteriors $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$ and $p(\boldsymbol{\theta}|\mathbf{y})$ are asymptotically Gaussian, as $n \to \infty$, and concentrate onto the same value in Θ, this difference is determined by the difference in the posterior variances: in particular, for $n \to \infty$,

$$\int_{\Theta} |p(\boldsymbol{\theta}|\eta(\mathbf{y})) - p(\boldsymbol{\theta}|\mathbf{y})|^2 \mathrm{d}\boldsymbol{\theta} \leq \mathrm{KL}\left[p(\boldsymbol{\theta}|\mathbf{y}), p(\boldsymbol{\theta}|\eta(\mathbf{y}))\right] \approx \frac{1}{2}\left[\ln\frac{|\mathcal{I}_{\eta}|}{|\mathcal{I}|} - \dim(\boldsymbol{\theta}) + \mathrm{Tr}\left[\mathcal{I}_{\eta}^{-1}\mathcal{I}\right]\right],$$

where $\mathcal{I}$ (respectively, $\mathcal{I}_{\eta}$) denotes the Fisher information matrix of the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ (respectively, $p(\eta(\mathbf{y})|\boldsymbol{\theta})$), $|\cdot|$ denotes the determinant, and $\mathrm{Tr}(\cdot)$ the trace operator. Clearly, the above is zero if and only if $\mathcal{I}_{\eta} = \mathcal{I}$, i.e., if and only if the summaries are sufficient. More generally, the above relationship demonstrates that the more informative are the summaries, i.e., the closer $\eta(\mathbf{y})$ is to being sufficient, the closer the partial posterior $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$ will be to the exact posterior $p(\boldsymbol{\theta}|\mathbf{y})$. To this end, some 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.

Analyzing the first term in (5), we note that if $\varepsilon \to 0$, then under reasonable assumptions, (such as, e.g., the dominance condition $p_{\varepsilon}(\boldsymbol{\theta}|\eta(\mathbf{y})) \leq C < \infty$, for all $\boldsymbol{\theta} \in \Theta$ and all $\varepsilon \to 0$), this first term will converge to zero. However, in practice, since Algorithm 2 generates *i.i.d.* draws of $\boldsymbol{\theta}$ under the prior,

[5] Wilkinson (2013) argues that an equally valid interpretation of Algorithm 2 is that it produces exact draws from a controlled approximation to the target posterior, $p(\boldsymbol{\theta}|\eta(\boldsymbol{y}))$. This controlled approximation $p_{\varepsilon}(\boldsymbol{\theta}|\eta(\boldsymbol{y}))$ is actually expressible as a convolution of the exact partial posterior with a kernel function that is used to represent error in the summary statistics. This convolution can itself be interpreted as an exact posterior associated with a randomised version of $\eta(\boldsymbol{y})$.

many of the subsequent $\eta(\mathbf{z}^i)$ values will be far away from $\eta(\mathbf{y})$; hence a large value of ε may be required to obtain a reasonable acceptance rate for the algorithm. Consequently, obtaining draws from $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ can be difficult as ε becomes small. In the regime when ε is large, the first term in equation (5) can be large even if the second term in equation (5) is small. To address this issue, several extensions of the basic ABC algorithm have been proposed that seek to increase the mass of simulated summaries $\eta(\mathbf{z})$ in the region of $\eta(\mathbf{y})$, with the hope being that these methods yield a more accurate approximation to $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$. These proposals broadly fall into two classes, and are often used in conjunction: the first is the use of post-processing corrections, the second is the use of proposals that 'learn' regions of $(\boldsymbol{\theta}, \mathbf{z})$ where $\eta(\mathbf{z})$ is closer to $\eta(\mathbf{y})$.[6]

Broadly speaking, post-processing corrections adjust the accepted draws obtained from an initial ABC algorithm according to a given model, the most common being some form of regression model, in an attempt to increase the accuracy of the posterior approximation at a fixed value of ε; see, Beaumont *et al.* (2002), Blum (2010), Blum and François (2010) for examples, and Blum (2017) for a review. Alternatively, methods that 'learn' proposal distributions can deliver more simulations in regions of $(\boldsymbol{\theta}, \mathbf{z})$ where $\eta(\mathbf{z})$ is close to $\eta(\mathbf{y})$. One such approach is to insert an MCMC step, and associated proposal distribution, within Algorithm 2 in order to more effectively explore the space, which yields an ABC-MCMC algorithm (Marjoram *et al.*, 2003). A more efficient exploration of the posterior space for $(\boldsymbol{\theta}, \mathbf{z})$ increases the likelihood that we obtain draws of $\eta(\mathbf{z})$ closer to $\eta(\mathbf{y})$, and subsequently ensures that, all else equal, ABC-MCMC can use a smaller tolerance than that used in Algorithm 2, and thus obtain a more accurate approximation to $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$.

The downside of ABC-MCMC is that unless the tolerance ε is carefully tuned, the resulting Markov chain can mix poorly, thus leading to unreliable inference. Consequently, the use of approaches other than MCMC within Algorithm 2, for instance based on a decreasing sequence of tolerances, is commonplace. Indeed, arguably the most popular current approach to conducting ABC inference is to insert sequential, or 'adaptive', proposals within Algorithm 2, which leads to ABC-SMC/ABC-population(P)MC algorithms; see Sisson *et al.* (2007) and Beaumont *et al.* (2009) for examples, and Sisson and Fan (2019) for a review. ABC-SMC learns effective proposal distributions sequentially as part of the algorithm, which, all else equal, can deliver better approximations to $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$ than Algorithm 2. Importantly, since ABC-SMC is based on sequential importance sampling, the resulting independent posterior draws are free from the stickiness that can arise in ABC-MCMC. Furthermore, most common ABC-SMC algorithms sequentially learn the tolerance ε also so that explicit tuning of the tolerance is not required.[7]

As a final point, we make note of the well-known curse of dimensionality to which ABC is subject. At its simplest level, the estimation of $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ (for any given ε) using the M draws of Algorithm 2, is equivalent to nonparametric conditional density estimation. As such, the accuracy of the estimate degrades as the dimension of $\eta(\mathbf{y})$ increases. Equivalently, a given level of accuracy requires a larger

[6] We make mention here of a further method that shares some features in common with these two categories of method, namely Bayesian optimization for likelihoood-free inference, or BOLFI (Gutmann and Corander, 2016). BOLFI uses Bayesian optimization to iteratively build a probabilistic model for the relationship between $\boldsymbol{\theta}$ and the distance function $d\{\eta(\mathbf{y}), \eta(\mathbf{z})\}$. The effect of this is to produce draws that yield small values for $d\{\cdot, \cdot\}$ and, hence, to reduce the number of required model simulations. The principle is equally applicable to the BSL technique to be discussed below.

[7] We refer to Kousathanas *et al.* (2019) for a review of software that enables many of the ABC algorithms discussed in this section to be easily implemented.

value of M and, hence, entails a higher computational burden, the larger is the dimension of $\eta(\mathbf{y})$. Whilst the modifications of ABC noted above potentially reduce the computational burden associated with any given $\eta(\mathbf{y})$ – by either correcting draws post-simulation, or producing more effective draws in the first place – the issue of dimension still obtains, and is simply intrinsic to the selection method that underpins ABC. See Blum *et al.* (2013) and Nott *et al.* (2018) for in-depth discussions, and also Frazier *et al.* (2018) for additional insights into the impact of the dimension of $\boldsymbol{\theta}$ on the asymptotic behaviour of ABC.

ABC using full data distances

Recently, several researchers have 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 such cases, the accept/reject step in Algorithm 2 is simply replaced with a discrepancy over the space of probability measures. More formally, let $\hat{\mu}$ denote the empirical measure of the observed sample and $\hat{\mu}_{\boldsymbol{\theta}}$ the empirical measure calculated from the simulated sample $\mathbf{z}$. Then, for $\mathcal{D}(\hat{\mu}_n, \hat{\mu}_{\boldsymbol{\theta}})$ denoting a generic discrepancy that measures the difference between $\hat{\mu}_n$ and $\hat{\mu}_{\boldsymbol{\theta}}$, the distance between the summaries, $d\{\eta(\mathbf{y}), \eta(\mathbf{z})\}$, is replaced by $\mathcal{D}(\hat{\mu}_n, \hat{\mu}_{\boldsymbol{\theta}})$.

Several choices of $\mathcal{D}(\cdot,\cdot)$ have been proposed, including the Wasserstein distance (Bernton *et al.*, 2019), KL divergence (Jiang, 2018), minimum mean discrepancy (Park *et al.*, 2016), the energy distance (Nguyen *et al.*, 2020) and the Cramer-von Mises distance (Frazier, 2020). Recently, Drovandi and Frazier (2021) have undertaken an in-depth comparison of these different methods for conducting inference, and compared the results with a generic summary statistic-based ABC approach across several examples. The authors' main findings are three-fold. First, the distance-based approaches are found to be promising, and to deliver reasonable inferences in many cases, whilst obviating the need to seek a vector of informative summary statistics. Secondly, and as a slight qualification to the first finding, the authors find that distance-based approaches must be combined with summary statistics to ensure identification of $\boldsymbol{\theta}$ in certain classes of models. Lastly, at least in their experiments, the best performing summary statistic-based approach always performs at least as well as the best distance-based approach, which suggests that if one can find informative summary statistics they may outperform distance-based approaches in general.

3.1.2 Bayesian synthetic likelihood (BSL)

Summary statistic-based ABC 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 insufficient $\eta(\mathbf{y})$, 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,

$$p_{\varepsilon}(\eta(\mathbf{y})|\boldsymbol{\theta}) = \int_{\mathcal{X}} p(\mathbf{z}|\boldsymbol{\theta})\mathbb{I}\left(d\{\eta(\mathbf{y}), \eta(\mathbf{z})\} \leq \varepsilon\right) d\mathbf{z}. \tag{6}$$

For a given draw $\boldsymbol{\theta}^i$, and associated $\eta(\mathbf{z}^i)$, (6) is approximated by its unbiased simulation counterpart, $\widehat{p}_{\varepsilon}(\eta(\mathbf{y})|\boldsymbol{\theta}^i) = \mathbb{I}\left(d\{\eta(\mathbf{y}), \eta(\mathbf{z}^i)\} \leq \varepsilon\right)$, 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 a Metropolis-Hastings 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 $\eta(\mathbf{z}^i)$ 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) overwhelmingly adopts a Gaussian parametric approximation to $p(\eta(\mathbf{y})|\boldsymbol{\theta})$,

$$p_a(\eta(\mathbf{y})|\boldsymbol{\theta}) = \mathcal{N}\left[\eta(\mathbf{y}); \mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta})\right], \; \mu(\boldsymbol{\theta}) = \mathbb{E}[\eta(\mathbf{y})], \; \Sigma(\boldsymbol{\theta}) = \text{Var}\left[\eta(\mathbf{y})\right]. \tag{7}$$

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

$$p_a(\boldsymbol{\theta}|\eta(\mathbf{y})) \propto p_a(\eta(\mathbf{y})|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{8}$$

where the subscript 'a' highlights that (8) 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 $\mathbf{z}_j \sim i.i.d.\ p(\cdot|\boldsymbol{\theta})$, $j = 1, \dots, m$, we can estimate $\mu(\boldsymbol{\theta})$ and $\Sigma(\boldsymbol{\theta})$ in (7) via their empirical Monte Carlo averages, $\mu_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^m \eta(\mathbf{z}_j)$ and $\Sigma_m(\boldsymbol{\theta}) = \frac{1}{m-1}\sum_{j=1}^m (\eta(\mathbf{z}_j) - \mu_m(\boldsymbol{\theta}))(\eta(\mathbf{z}_j) - \mu_m(\boldsymbol{\theta}))'$, and thereby define

$$p_{a,m}(\eta(\mathbf{y})|\boldsymbol{\theta}) = \int_{\mathcal{X}} \mathcal{N}\left[\eta(\mathbf{y}); \mu_m(\boldsymbol{\theta}), \Sigma_m(\boldsymbol{\theta})\right] \prod_{j=1}^m p(\eta(\mathbf{z}_j)|\boldsymbol{\theta}) d\mathbf{z}_1 \dots d\mathbf{z}_m, \tag{9}$$

and the associated *target* BSL posterior,

$$p_{a,m}(\boldsymbol{\theta}|\eta(\mathbf{y})) \propto p_{a,m}(\eta(\mathbf{y})|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{10}$$

Note that, even for a single draw $\eta(\mathbf{z}_j)$, $\mathbf{z}_j \sim p(\cdot|\boldsymbol{\theta})$, we have that $\mathcal{N}\left[\eta(\mathbf{y}); \mu_m(\boldsymbol{\theta}), \Sigma_m(\boldsymbol{\theta})\right]$ is an unbiased estimate of (9). 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. Pseudo-code for generic MCMC sampling of the BSL posterior in (10) is given in Algorithm 3. We refer the interested reader to the R language (R Core Team, 2020) package BSL (An *et al.*, 2019), which can be used to implement BSL and its common variants.

3.1.3 ABC and BSL

Whilst (summary statistic-based) ABC and BSL target the same posterior, $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$, both methods produce posteriors that differ from this target, and from one another. Therefore, it is helpful to characterize the difference between these posteriors in in terms of their *i)* large sample (in n) behaviour and

Algorithm 3 Vanilla BSL MCMC Algorithm

```
for $i = 1, \dots, M$ do
    Draw $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{i-1})$
    Produce $\mu_m(\boldsymbol{\theta})$ and $\Sigma_m(\boldsymbol{\theta})$ using $j = 1, \dots, m$ independent model simulations at $\boldsymbol{\theta}^*$
    Compute the synthetic likelihood $L^* = \mathcal{N}\left[\eta(\mathbf{y}); \mu_m(\boldsymbol{\theta}^*), \Sigma_m(\boldsymbol{\theta}^*)\right]$ and $L^{i-1}$
    Compute the Metropolis-Hastings ratio:
        $r = \dfrac{L^* \pi(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^*)}{L^{i-1} \pi(\boldsymbol{\theta}^{i-1}) q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})}$
    if $\mathcal{U}(0,1) < r$ then
        Set $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$, $\mu_m(\boldsymbol{\theta}^i) = \mu_m(\boldsymbol{\theta}^*)$ and $\Sigma_m(\boldsymbol{\theta}^i) = \Sigma_m(\boldsymbol{\theta}^*)$
    else
        Set $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$, $\mu_m(\boldsymbol{\theta}^i) = \mu_m(\boldsymbol{\theta}^{i-1})$ and $\Sigma_m(\boldsymbol{\theta}^i) = \Sigma_m(\boldsymbol{\theta}^{i-1})$
    end if
end for
```

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, as $n \to \infty$, the ABC posterior $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ targeted by Algorithm 2: 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, the ABC posterior can be as efficient as the maximum likelihood estimator based on the likelihood $p(\eta(\mathbf{y})|\boldsymbol{\theta})$.

The required conditions on the tolerance, ε, for these results to be in evidence can be ordered in terms of the speed with which $\varepsilon \to 0$ 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 $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ to be an accurate estimate of $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$. Broadly speaking, the smaller is ε, the smaller is $|p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y})) - p(\boldsymbol{\theta}|\eta(\mathbf{y}))|$. However, a smaller choice of ε requires a larger number of simulations (i.e., a larger value of M) and, hence, a greater computational effort. For instance, if we wish for credible sets obtained by $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ to be valid in the frequentist sense, M is required to diverge faster than $n^{\dim(\boldsymbol{\eta})/2}$ (Corollary 1 in Frazier *et al.*, 2018).

In contrast to ABC, BSL is based on the Gaussian approximation to the likelihood $p(\eta(\mathbf{y})|\boldsymbol{\theta})$, and 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 frequentist 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 $m/\log(n)\rightarrow\infty$. 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 $\int |p_{a,m}(\boldsymbol{\theta}|\eta(\mathbf{y})) - p_{\varepsilon}(\boldsymbol{\theta}|\eta(\mathbf{y}))|d\boldsymbol{\theta} \overset{p}{\rightarrow} 0$ as $n\rightarrow\infty$. 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) The above asymptotic comparison between ABC and BSL abstracts from the actual sampling required to obtain draws from the posterior targets; that is, the large sample behavior discussed above is divorced from the actual practice of obtaining draws from the different posteriors, and thus ignores the computational efficiency of the two approaches. Once computational efficiency, is taken into account, 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.[8]

3.1.4 Illustrative example: ABC and BSL

We complete this section on simulation-based approximate methods with a brief discussion of an empirical example from Drovandi and Frazier (2021) in which both ABC and BSL methods are applied. We have selected this particular example as our illustration because it is has two features that are common to many empirical applications of ABC and BSL: *1)* The model does not enable a likelihood function to be computed analytically, but the model *can* be simulated; *2)* Despite the complexity of the model, the number of parameters of interest is small; hence a reasonably small number of summary statistics are able to be selected. The illustration also includes a comparison of summary-statistic based ABC with ABC based on full distances. We present certain graphical output (Figure 5 in their original paper) as Figure 1 below.

The empirical problem is one of conducting inference on the large-valued imperfections (or 'inclusions') in steel that can arise during the production process; or in general parlance, one of conducting

[8] 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.

inference for *stereological extremes.* We refer to Bortot *et al.* (2007) for all details of the physical and statistical problems. Suffice to say, for the illustrative purpose here, that a realistic model for explaining such extreme inclusions, namely an ellipsoid family for inclusion shapes, does not have an available likelihood function, but can be inexpensively simulated. Moreover, the particular model analyzed in Drovandi and Frazier (2021) is described by only three parameters: the rate parameter (λ) of a homogenous Poisson process describing the random number of inclusions per volume of steel, and the scale (σ) and shape (ξ) parameters of a generalized Pareto distribution related to the size of the inclusions.

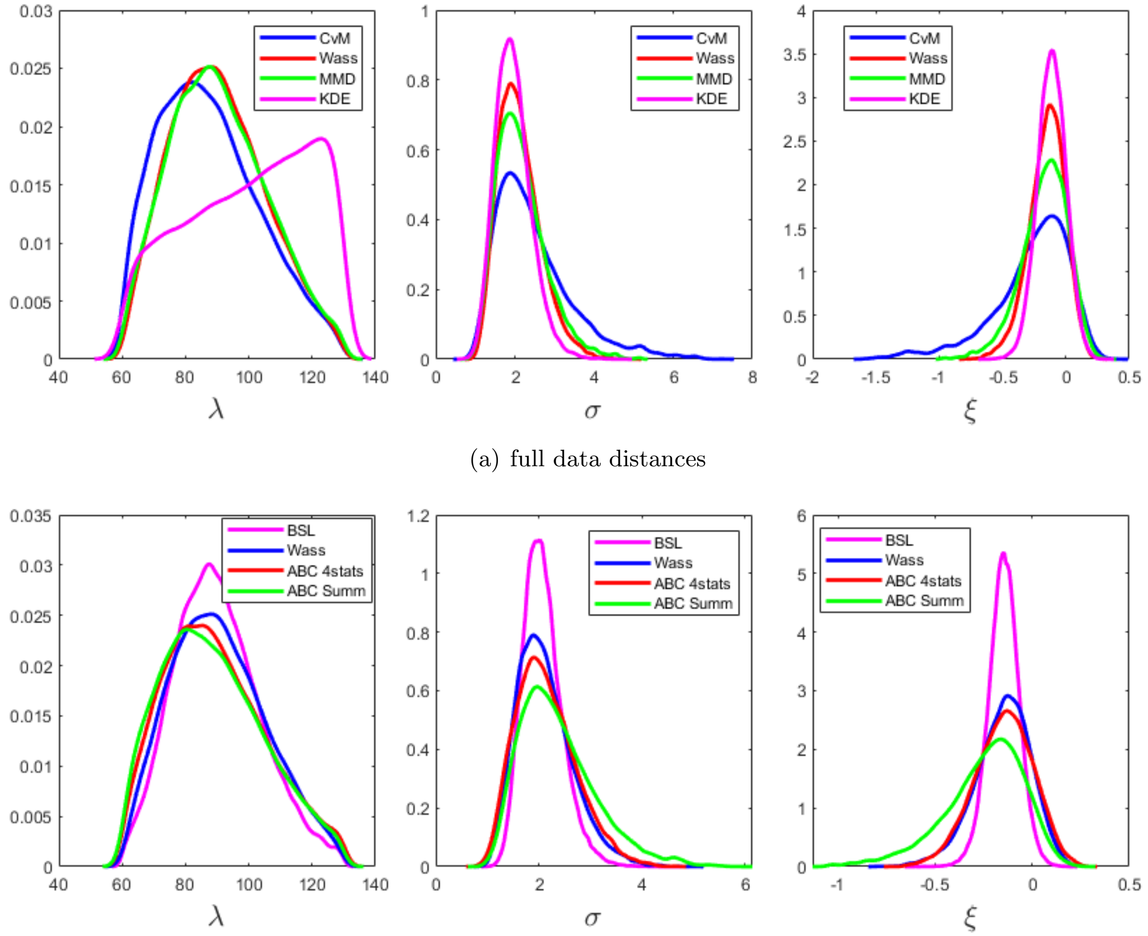
Drovandi and Frazier (2021) consider ABC based on two different sets of summary statistics. The first choice is based on a similar set of four statistics to that used in Bortot *et al.* (2007) ('ABC 4stats' in Figure 1), while the second set is based on the nine-dimensional score vector of an auxiliary Gaussian mixture model with three components ('ABC Summ' in Figure 1); BSL-based inference is based on this second set of summaries only ('BSL' in Figure 1). When applying the distance-based ABC approaches, Drovandi and Frazier (2021) note that the inclusion size, a continuous variable, and the number of inclusions, a discrete variable, both carry identifying information about the unknown parameters. To this end, the authors combine two distance functions, one for the number of inclusions, and one for the inclusion sizes. For the inclusion sizes, the authors use a range of distance functions including Cramer-von mises (CvM), Wasserstein (Wass), maximum mean discovery (MMD), and the simulation-based kernel density approach of Turner and Sederberg (2014) (KDE). Each distance is then combined with the absolute difference between the observed number of inclusions and the simulated number of inclusions from the model.

Some key messages to be taken from Figure 1 are as follows:

1. The ABC posteriors based on different summary statistics and distance functions produce different posteriors! More specifically, and as is reasonably typical, the posteriors for any given parameter are generally centred at similar points in the parameter space, but have varying degrees of dispersion. Of the posteriors based on summary statistics, plotted in the bottom row of the figure, ABC based on the nine summaries derived from the Gaussian mixture model ('ABC summ') has the largest dispersion in each case. This reflects the curse of dimensionality in the dimension of the summary statistics to which ABC is subject, as discussed in Section 3.1.1.

2. Following on from the above point, and with reference to Point *ii)* in Section 3.1.3, the BSL posterior based on the Gaussian mixture model summaries is notably less dispersed than the corresponding 'ABC summ'. This difference can be attributed to the approximate Gaussianity of the summary statistics in this example, which results in a BSL posterior that is less sensitive to the dimension of the summaries than ABC. Consequently, given the same computing budget for both methods, we would expect that BSL would produce more efficient posteriors since its acceptance rate does not decline as sharply as that of ABC when the dimension of the summaries is moderate or large.

3. With reference to the plots in the top row of Figure 1, not all distance functions produce reasonable posteriors. Like summaries, different distances capture different features of the data. Moreover,



(a) full data distances

(b) summary statistics

Figure 1: Figure 5 from Drovandi and Frazier (2021), with caption: “Comparison of estimates of the univariate ABC posterior distributions for the stereological extremes example based on real data. Shown are (a) comparisons with distance functions involving the full data and (b) comparisons with summary statistic-based approaches.”

as mentioned above, the use of a single distance alone may not be able to identify all models parameters in all circumstances. Therefore, careful preliminary analysis should be undertaken when using distance-based ABC.

4. Lastly, the least dispersed summary statistic method (i.e. BSL) has less dispersion than the best distance-based ABC approach (‘Wass’ in this case). Drovandi and Frazier (2021) find similar behavior in all the examples considered in their analysis, which suggests that, while distance-based ABC approaches are useful as they obviate the crucial choice of which summaries to select, they may not perform as well as methods based on informative summary statistics, at least in cases where a feasible informative and low-dimensional summary exists.

3.2 Optimization approaches

3.2.1 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, (2), 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 (2) – *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,

$$q^*(\boldsymbol{\theta}) := \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}\left[q(\boldsymbol{\theta})|p(\boldsymbol{\theta}|\mathbf{y})\right], \tag{11}$$

where

$$\mathrm{KL}\left[q(\boldsymbol{\theta})|p(\boldsymbol{\theta}|\mathbf{y})\right] = \int \log(q(\boldsymbol{\theta}))q(\boldsymbol{\theta})d\boldsymbol{\theta} - \int \log(p(\boldsymbol{\theta}|\mathbf{y}))q(\boldsymbol{\theta})d\boldsymbol{\theta} \equiv \mathbb{E}_q[\log(q(\boldsymbol{\theta}))] - \mathbb{E}_q[\log(p(\boldsymbol{\theta},\mathbf{y}))] + \log(p(\mathbf{y})) \tag{12}$$

and $p(\boldsymbol{\theta},\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. 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 (12) inaccessible as a result. Rather, the approach adopted is to define the so-called evidence lower bound (ELBO),

$$\mathrm{ELBO}[q(\boldsymbol{\theta})] := \mathbb{E}_q[\log(p(\boldsymbol{\theta},\mathbf{y}))] - \mathbb{E}_q[\log(q(\boldsymbol{\theta}))], \tag{13}$$

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 (11):

$$q^*(\boldsymbol{\theta}) := \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\max}\, \mathrm{ELBO}[q(\boldsymbol{\theta})]. \tag{14}$$

In practice, $q(\boldsymbol{\theta})$ is either explicitly or implicitly parameterized by a vector of 'variational parameters', $\boldsymbol{\lambda}$, and optimization occurs with respect to $\boldsymbol{\lambda}$.

The beauty of VB is that, for *certain* problems, including certain choices of the class $\mathcal{Q}$, the optimization problem in (14) 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 $p(\boldsymbol{\theta}|\mathbf{y})$ by $q^*(\boldsymbol{\theta})$ in (2) yields an expectation that is either available in closed form, or amenable to a relatively simple simulation-based solution. Moreover, the link between (12) and (13) makes it clear that maximizing (13) 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, $\text{ELBO}[q^*(\boldsymbol{\theta})]$ serves as an estimate of the quantity that underpins model choice.

The production of $q^*(\boldsymbol{\theta})$, and the associated estimate of an approximation of (2) as based on $q^*(\boldsymbol{\theta})$, is typically *much* faster (often orders of magnitude so) than producing an estimate of (2) via exact simulation of $p(\boldsymbol{\theta}|\mathbf{y})$. This is of particular import when both $\boldsymbol{\theta}$, and possibly $\mathbf{y}$ also, are high-dimensional. In such cases, the computational cost of simulating from $p(\boldsymbol{\theta}|\mathbf{y})$, via MCMC for example, may simply be prohibitive, given the need to both explore a high-dimensional and complex parameter space and – at each point in that search – evaluate $p(\mathbf{y}|\boldsymbol{\theta})$ at $\mathbf{y}$. In contrast, the variational family $\mathcal{Q}$, and the optimization algorithm, can be chosen in such a way that a VB approximation of $p(\boldsymbol{\theta}|\mathbf{y})$ can be produced within an acceptable timeframe, 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). The ability of VB to scale to large models and datasets also makes the method particularly suitable for exploring multiple models quickly, perhaps as a preliminary step to a more targeted analysis (Blei *et al.*, 2017).

We now give specific algorithmic details for two foundational VB algorithms: coordinate ascent variational inference (CAVI) (see Bishop, 2006, Chapter 10, for discussion) and stochastic variational inference (SVI) (Hoffman *et al.*, 2013), both of which seek to solve the optimization problem in (14), for given specifications of $p(\boldsymbol{\theta}, \mathbf{y})$ and choices of Q. These algorithms suit the intended purpose of this review as they both played a prominent role in the initial development of the VB literature, and allow us to discuss some of the mechanics of VB without getting needlessly bogged down in the details. For a review of more recent developments in VB, including details of implementation, we refer to Zhang *et al.* (2018).[9]

The CAVI algorithm is derived for the 'mean-field' variational family, where the elements of $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_p)'$ are specified as mutually independent, with joint density $q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)$ denoting a generic element of Q. The CAVI algorithm makes use of the fact that, under the mean-field family, the density $q_j^*(\theta_j)$, the solution to (14) for the j-th element of $\boldsymbol{\theta}$, has the closed form $q_j^*(\theta_j) \propto \exp(\mathbb{E}_{-j}[log(p(\theta_j|\boldsymbol{\theta}_{-j},\ \mathbf{y})])$ – where $\mathbb{E}_{-j}$ denotes the expectation with respect to the variational density over $\boldsymbol{\theta}_{-j}$, $\prod_{l\neq j} q_l(\theta_l)$ – which can be derived from (13) by exploiting the independence of the θ_j under the mean-field family (see Blei *et al.*, 2017, p.10). However, this solution is not explicit since $q_j^*(\theta_j)$ depends on expectations computed with respect to the other factors $q_{-j}(\boldsymbol{\theta}_{-j})$. Hence, given an initial solution, CAVI cycles through $q_j^*(\theta_j)$, $j = 1, \ldots, \dim(\boldsymbol{\theta})$, updating each factor in turn. The fact that we are able to calculate $\mathbb{E}_{-j}[\log p(\theta_j|\boldsymbol{\theta}_{-j}, \mathbf{y})]$ in closed form ensures, in turn, that the algorithm provides a very speedy solution to (14). In Algorithm 4, we provide pseudo-code for implementing CAVI, deferring to Blei *et al.* (2017) for further details.

In contrast to CAVI, SVI is applicable to a broader range of scenarios for both $p(\boldsymbol{\theta}, \mathbf{y})$ and $\mathcal{Q}$ (see Hoffman *et al.*, 2013, Section 5, on this point). In addition, it scales better to very large data sets as,

[9] We note that, unlike approximate Bayesian methods based on simulation, the diverse, and complex, nature of the problems to which VB methods are applied make it somewhat less well-suited to generating well-behaved, and reliable, software products that can be used to implement the methods across a wide range of problems. That being said, the automatic differentiation variational inference (ADVI) approach of Kucukelbir *et al.* (2017) can be implemented in many different problems, and is the default method for variational inference in the popular probabilistic programming language STAN (Carpenter *et al.*, 2017).

unlike CAVI, it does not require the full vector $\mathbf{y}$ to be processed on each iteration. In Algorithm 5, we provide pseudo-code for implementing SVI for the case of a 'conditionally conjugate model' and a mean-field variational family, in which we now exploit the breakdown of $\boldsymbol{\theta}$ into a vector of global parameters, $\boldsymbol{\phi}$, and an n-dimensional vector of local parameters, $\mathbf{x}$ (see Section 2.2). Referring to Blei *et al.* and Hoffman *et al.* for further details (and noting the differing notation), we assume the following structure for the joint distribution:

$$p(\boldsymbol{\phi}, \mathbf{x}, \mathbf{y}) = p(\boldsymbol{\phi}|\boldsymbol{\alpha}) \prod_{i=1}^{n} p(x_i, y_i|\boldsymbol{\phi}), \tag{15}$$

where $p(x_i, y_i|\boldsymbol{\phi})$ is a member of the linear exponential family, and $p(\boldsymbol{\phi}|\boldsymbol{\alpha})$ is the appropriate natural conjugate prior, with hyperparameter vector, $\boldsymbol{\alpha}$. In the algorithm, ψ_i denotes the variational parameter for each local parameter, x_i, $\boldsymbol{\lambda}$ the vector of variational parameters associated with $\boldsymbol{\phi}$, and $\eta(\cdot)$ and $t(\cdot)$ are specific functional forms that define the member of the exponential family underlying the specification in (15) (see Blei *et al.*). The key implication of the assumed exchangeable structure in (15) is that this structure permits the use of stochastic optimization routines to search the variational parameters $\boldsymbol{\lambda}$ that deliver the best approximation in the class $\mathcal{Q}$. That is, in contrast with Algorithm 4, the full vector $\mathbf{y}$ need not be processed at each iteration. Instead, a single observation, y_i, or batches of y_i, can be randomly selected and used to optimize the ELBO over both the local and global variational parameters. This simplification allows the algorithm to successfully scale to problems in which $\mathbf{y}$ is truly massive, at the cost of assuming the exchangeable structure in (15).

Algorithm 4 Vanilla CAVI Algorithm

```
Input: A joint probability distribution, $p(\boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$
Output: A variational density from the mean-field class, $q^*(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j^*(\theta_j)$
Initialize: Variational factors $q_j(\theta_j), j = 1, \ldots, p$
while the ELBO in (13) has not converged do
  for $j = 1, \ldots, p$ do
    Set $q_j^*(\theta_j) \propto exp(\mathbb{E}_{-j}[log(p(\theta_j|\boldsymbol{\theta}_{-j}, \mathbf{y}))])$
  end for
end while
Return: $q^*(\boldsymbol{\theta})$
```

Recently, several authors have analyzed the asymptotic properties of VB methods; see, for example, Wang and Blei (2019a,b), and Zhang and Gao (2020). 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

Algorithm 5 SVI Algorithm for a Conditionally Conjugate Model

```
Input: A joint probability distribution, $p(\boldsymbol{\phi}, \mathbf{x}, \mathbf{y})$ of the form specified in (15)
Output: A variational density for the global parameters $q_{\boldsymbol{\lambda}}(\boldsymbol{\phi})$
Initialize: A variational parameter vector, $\boldsymbol{\lambda}^{(0)}$
Set: The step-size schedule $\rho^{(t)}$
while TRUE do
    Choose a data point, $y_i$, uniformly at random, $i \sim Unif(1, ..., n)$
    Optimize its local variational parameter, $\psi_i = \mathbb{E}_{\boldsymbol{\lambda}^{(t-1)}}[\eta(\boldsymbol{\phi}, y_i)]$
    Compute the intermediate global variational parameter vector as though $y_i$ is replicated $n$ times, $\widehat{\boldsymbol{\lambda}} = \boldsymbol{\alpha} + n[\mathbb{E}_{\psi_i}[t(x_i, y_i)]', 1]'$
    Update the global variational parameter vector, $\boldsymbol{\lambda}^{(t)} = (1 - \rho^{(t)})\boldsymbol{\lambda}^{(t-1)} + \rho^{(t)}\widehat{\boldsymbol{\lambda}}$
end while
Return: $q_{\boldsymbol{\lambda}}(\boldsymbol{\phi})$
```

the algorithm's optimization step. As a final point, we note that Yao *et al.* (2018) and Huggins *et al.* (2019) propose methods for validating the accuracy of VB posterior approximations using alternative (nonasymptotic) principles.[10]

3.2.2 Integrated nested Laplace approximation (INLA)

In 1774, Pierre Simon Laplace published one of his many remarkable papers, *'Mémoire sur la probabilité des causes par les événemens'*, in which he produced the first asymptotic (in n) approximation to a posterior probability.[11] In brief, and using a scalar θ for the purpose of illustration, his original method can be explained as follows. Begin by expressing an arbitrary posterior probability as

$$\mathbb{P}(a < \theta < b|\mathbf{y}) = \int_a^b p(\theta|\mathbf{y})d\theta = \int_a^b \exp\{nf(\theta)\}\, d\theta, \tag{16}$$

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 $\sigma^2 = -[nf''(\widehat{\theta})]^{-1}$, and substituting the expansion into (16) then yields

$$\begin{aligned} \mathbb{P}(a < \theta < b|\mathbf{y}) &\approx \exp\left\{nf(\widehat{\theta})\right\} \int_a^b \exp\left\{-\tfrac{1}{2\sigma^2}(\theta - \widehat{\theta})^2\right\} d\theta \\ &= \exp\left\{nf(\widehat{\theta})\right\} \sqrt{2\pi\sigma^2} \times \{\Phi[\tfrac{b-\widehat{\theta}}{\sigma}] - \Phi[\tfrac{a-\widehat{\theta}}{\sigma}]\}, \end{aligned} \tag{17}$$

where $\Phi(.)$ denotes the standard Normal cumulative distribution function (cdf); and where, buried within the symbol '$\approx$' in (17), is a rate of convergence that is a particular order of n, and is probabilistic if randomness in $\mathbf{y}$ is acknowledged.

Not only did the result in (17) 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

[10] See also Yu *et al.* (2019) (and earlier references therein) for practical validation approaches that are relevant to approximate posteriors in general.

[11] See Stigler (1975), Stigler (1986a), Stigler (1986b) and Fienberg (2006) for various details about Laplace's role in the development of 'inverse probability', or Bayesian inference as it is now known.

like that in (2). Two centuries later, Tierney and Kadane (1986) and Tierney *et al.* (1989) 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 (2), including (in the multiple parameter case) marginal posterior densities.

Two decades later, Rue *et al.* (2009) then took the method further: adapting it to approximate marginal posteriors (and general expectations like those in (2)) in latent Gaussian models. 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 latent Gaussian model 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), Rue *et al.* (2017), Martino and Riebler (2019), van Niekerk *et al.* (2019) and Wood (2019) for specific implementation details (including of the latent Gaussian model structure), we provide here the *key* steps of INLA. To enhance the reader's understanding, we avoid the use of a summary algorithmic presentation of the method. Consistent with our previous notational convention, we decompose the full set of unknowns, $\boldsymbol{\theta}$, into an m-dimensional vector of 'hyperparameters' (in the language of INLA) that characterize the latent Gaussian model, $\boldsymbol{\phi}$, and the full set of K unknowns in the latent Gaussian field, denoted by $\boldsymbol{x}$. Each observation, y_i, $i = 1, 2, , , , n$, is assumed to be independent, conditional on a linear predictor, η_i, which is modelled as a random function of $\boldsymbol{x}$. For computational convenience, the vector $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_n)'$ is also included as an element of $\boldsymbol{x}$ (see Martino and Riebler, 2019, for details). The dimension, K, of $\boldsymbol{x}$ – which contains observation-specific, plus common, elements – is larger, and potentially much larger, than the dimension of $\mathbf{y}$ itself. The model is then expressed as:

$$\mathbf{y}|\mathbf{x},\boldsymbol{\phi} \sim \prod\nolimits_{i=1}^{n} p(y_i|\eta_i(\mathbf{x}), \boldsymbol{\phi}) \qquad \mathbf{x}|\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, Q^{-1}(\boldsymbol{\phi})) \qquad \boldsymbol{\phi} \sim p(\boldsymbol{\phi}), \tag{18}$$

where $Q(\boldsymbol{\phi})$ is the precision matrix of the latent Gaussian field, assumed – for computational feasibility – to be sparse. The goal of the authors is to approximate the marginal posteriors; $p(\phi_j|\mathbf{y})$, $j = 1, 2, .., m$, and $p(x_k|\mathbf{y})$, $k = 1, 2, .., K$. The problems envisaged are those in which m, the dimension of the hyperparameters $\boldsymbol{\phi}$, is small and K is large (potentially in the order of hundreds of thousands), with MCMC algorithms deemed to be computationally burdensome as a consequence, due to the scale of the unknowns (and potentially $\mathbf{y}$ also), and the challenging geometry of the posterior. We refer the reader to the references cited above for the wide range of problems of this type to which INLA has been applied.

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

$$p(\boldsymbol{\phi}|\mathbf{y}) = \frac{p(\mathbf{x}, \boldsymbol{\phi}|\mathbf{y})}{p(\mathbf{x}|\boldsymbol{\phi}, \mathbf{y})} \propto \frac{p(\mathbf{x}, \boldsymbol{\phi}, \mathbf{y})}{p(\mathbf{x}|\boldsymbol{\phi}, \mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{x},\boldsymbol{\phi})p(\mathbf{x}|\boldsymbol{\phi})p(\boldsymbol{\phi})}{p(\mathbf{x}|\boldsymbol{\phi}, \mathbf{y})}, \tag{19}$$

and recognizing that the proportionality sign arises due to the usual lack of integrating constant (over $\mathbf{x}$ and $\boldsymbol{\phi}$), the steps of the method (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, $p(\boldsymbol{\phi}|\mathbf{y})$ in (19) is approximated as

$$\widetilde{p}(\boldsymbol{\phi}|\mathbf{y}) \propto \frac{p(\mathbf{y}|\widehat{\mathbf{x}}(\boldsymbol{\phi}),\boldsymbol{\phi})p(\widehat{\mathbf{x}}(\boldsymbol{\phi})|\boldsymbol{\phi})p(\boldsymbol{\phi})}{p_G(\widehat{\mathbf{x}}(\boldsymbol{\phi})|\boldsymbol{\phi},\mathbf{y})}. \tag{20}$$

The denominator in (20) represents a Gaussian approximation of $p(\mathbf{x}|\boldsymbol{\phi},\mathbf{y})$, $p_G(\mathbf{x}|\boldsymbol{\phi},\mathbf{y}) = \mathcal{N}(\widehat{\mathbf{x}}(\boldsymbol{\phi}), \widehat{\Sigma}(\boldsymbol{\phi}))$, evaluated at the mode, $\widehat{\mathbf{x}}(\boldsymbol{\phi})$, of $p(\mathbf{x},\boldsymbol{\phi},\mathbf{y})$ (at a given value of $\boldsymbol{\phi}$), where $\widehat{\Sigma}(\boldsymbol{\phi})$ is the inverse of the Hessian of $-\log p(\mathbf{x},\boldsymbol{\phi},\mathbf{y})$ with respect to $\mathbf{x}$, also evaluated at $\widehat{\mathbf{x}}(\boldsymbol{\phi})$. The expression in (20) can obviously be further simplified to

$$\widetilde{p}(\boldsymbol{\phi}|\mathbf{y}) \propto p(\mathbf{y}|\widehat{\mathbf{x}}(\boldsymbol{\phi}),\boldsymbol{\phi})p(\widehat{\mathbf{x}}(\boldsymbol{\phi})|\boldsymbol{\phi})p(\boldsymbol{\phi}) \left|\widehat{\Sigma}(\boldsymbol{\phi})\right|^{1/2}, \tag{21}$$

which, up to the integrating constant, is 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 latent Gaussian model setting; whilst Tang and Reid (2021) provide further approximation results pertaining to high-dimensional models.

With the marginal posterior for the kth element of $\mathbf{x}$ defined as

$$\widetilde{p}(x_k|\mathbf{y}) = \int_\Theta \widetilde{p}(x_k|\boldsymbol{\phi},\mathbf{y})\widetilde{p}(\boldsymbol{\phi}|\mathbf{y})d\boldsymbol{\phi}, \tag{22}$$

a second application of a Laplace approximation would yield

$$\widetilde{p}(x_k|\boldsymbol{\phi},\mathbf{y}) \propto p(\mathbf{y}|\widehat{\mathbf{x}}_{-k}(\boldsymbol{\phi},x_k),\boldsymbol{\phi})p(\widehat{\mathbf{x}}_{-k}(\boldsymbol{\phi},x_k)|\boldsymbol{\phi})p(\boldsymbol{\phi}) \left|\widehat{\Sigma}_{-k}(\boldsymbol{\phi},x_k)\right|^{1/2}, \tag{23}$$

where $\widehat{\mathbf{x}}_{-k}(\boldsymbol{\phi},x_k)$ is the mode of $p(\mathbf{x}_{-k},x_k,\boldsymbol{\phi},\mathbf{y})$ (at given values of $\boldsymbol{\phi}$ and x_k, with $\mathbf{x}_{-k}$ denoting all elements of $\mathbf{x}$ other than the kth); and where $\widehat{\Sigma}_{-k}(\boldsymbol{\phi},x_k)$ is the inverse of the Hessian of $-\log p(\mathbf{x}_{-k},x_k,\boldsymbol{\phi},\mathbf{y})$ with respect to $\mathbf{x}_{-k}$, also evaluated at $\widehat{\mathbf{x}}_{-k}(\boldsymbol{\phi},x_k)$. Computation of (23) for each x_k would, however, involve K optimizations (over $\mathbf{x}_{-k}$) plus K specifications of the high-dimensional matrix $\widehat{\Sigma}_{-k}(\boldsymbol{\phi},x_k)$. Rue *et al.* (2009) avoid this computational burden by modifying the approximation in (23) in a number of alternative ways, all details of which are provided in the references cited above. Once a representation of $\widetilde{p}(x_k|\boldsymbol{\phi},\mathbf{y})$ is produced, (22) is computed using a deterministic numerical integration scheme defined over a grid of values for the low-dimensional $\boldsymbol{\phi}$.

Defining the marginal posterior for the jth element of $\boldsymbol{\phi}$ as $\widetilde{p}(\phi_j|\mathbf{y}) = \int_{\Theta_{-j}} \widetilde{p}(\boldsymbol{\phi}|\mathbf{y})d\boldsymbol{\phi}_{-j}$, where $\boldsymbol{\phi}_{-j}$ denotes all elements of $\boldsymbol{\phi}$ excluding ϕ_j, this integral is computed using $m-$dimensional deterministic integration over $\boldsymbol{\phi}_{-j}$, once again on the maintained assumption that m is small. Finally, if required, the marginal likelihood, $p(\mathbf{y})$ can be approximated by computing the normalizing constant in (21), $\int_\Theta p(\mathbf{y}|\widehat{\mathbf{x}}(\boldsymbol{\phi}),\boldsymbol{\phi})p(\widehat{\mathbf{x}}(\boldsymbol{\phi})|\boldsymbol{\phi})p(\boldsymbol{\phi}) \left|\widehat{\Sigma}(\boldsymbol{\phi})\right|^{1/2} d\boldsymbol{\phi}$, using deterministic integration over $\boldsymbol{\phi}$.

All steps of the INLA algorithm can be implemented using the dedicated package, R-INLA (available at www.r-inla.org), 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. Gomez-Rubio and Rue (2018) and Berild *et al.* (2021) demonstrate how the INLA approach (and the R-INLA software) can also be applied to models beyond the LGM class by means of additional MCMC or IS sampling steps applied to models that are LGMs *conditional* on certain fixed parameters. Margossian *et al.* (2020) extend INLA principles to the case in which m is too large for treatment by deterministic integration, by 'embedding' INLA within an HMC sampling scheme. In this case $\widetilde{p}(\boldsymbol{\phi}|\mathbf{y})$ - produced as in (21) - serves as the target density for the HMC sampler, and each $\widetilde{p}(\phi_j|\mathbf{y})$ is estimated via the HMC draws. Finally, Stringer *et al.* (2021) have adapted the standard INLA methodology both to cater for an *extended* class of LGMs, in which the conditional independence assumption for y_i is eschewed, and to scale better to large data sets.

3.2.3 Illustrative example: VB and INLA

We complete this section on approximate Bayesian inference via optimization by displaying and discussing graphical output from Braun and McAuliffe (2010) and Margossian *et al.* (2020), in which, respectively, VB and INLA are used to conduct inference. The selected illustration from Braun and McAuliffe highlights the feasibility, speed and (comparable) predictive accuracy of VB, versus an MCMC comparator. The illustration extracted from Margossian *et al.* compares the accuracy and speed of the 'embedded' HMC method with a 'full' HMC algorithm, in which *both* the latent Gaussian field and the hyperparameters are inferred via simulation.

VB illustration

We record here certain output from a particular simulation exercise in Braun and McAuliffe (2010), in which VB is used to perform inference on a large-scale hierarchical model for consumer choice. This illustration shares characteristics common to many applications of this approximate method (and, indeed, of INLA too): *1)* Very high-dimensional $\boldsymbol{\theta}$ and $\mathbf{y}$; but, at the same time *2)* An analytical expression for the model, $p(\mathbf{y}|\boldsymbol{\theta})$.

The model in question is a 'random utility model' specified for H customers, each with heterogeneous preferences or 'tastes', and each having to select from J items (or choices), each with K choice-specific attributes. The total number of unknowns comprise the K-dimensional vectors of customer-specific preferences over the attributes, $\boldsymbol{\beta}_h$, which may be specific to each of the $h = 1, 2, ..., H$ customers, plus the mean vector ($\boldsymbol{\zeta}$) and variance-covariance matrix ($\boldsymbol{\Omega}$) of the K-dimensional Gaussian distribution that models the distribution of preferences across the population. Hence, in terms of our notation, the dimension of $\boldsymbol{\theta}$ is the combined dimensions of $\boldsymbol{\zeta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_H$. The vector of observed data $\mathbf{y}$ comprises T_h choice events across H customers and is thus of total length $H * T_h$. A $(J \times K)$ matrix of observed attributes encountered by customer $h = 1, 2, ..., H$, at choice event $t = 1, 2, .., T_h$ completes the observed data, where we denote the full (concatenated) matrix of observed attributes over agents and events simply by $\mathbf{X}$. For the design scenario with the largest specifications, $H = 25,000$, $T_h = 25$, $J = 12$ and $K = 10$.

A mean-field variational family $\mathcal{Q}$ is adopted, with a variational approximation to $p(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_H, \boldsymbol{\zeta}, \boldsymbol{\Omega}| \mathbf{X}, \mathbf{y})$ chosen from $\mathcal{Q}$ to maximize the ELBO, via a block coordinate ascent algorithm implemented with analytical expressions for the gradient and the Hessian of the criterion function (Braun and McAuliffe, 2010, Appendix A). Whilst MCMC is obviously challenging for this particular model, due to the scale of both $\boldsymbol{\theta}$ and $\mathbf{y}$, and, indeed, exhausts machine memory at a very small number of iterations (1000), it *is* feasible; hence, one aim of this simulation exercise is to illustrate the *relative speed* of VB versus MCMC, where the MCMC algorithm is that of Rossi and Allenby (2003). We display (as our own Figure 2) Figure 2 from Braun and McAuliffe (2010), retaining the original caption as, in tandem with the explanatory material above, it is sufficiently informative to allow the results to be interpreted without access to the paper. We note that in the body of the figure: 'items' refers to J; 'attrs' refers to K; and 'Low/High het' refers to magnitude of the diagonal elements of $\mathbf{\Omega}$ (i.e. the degree of heterogeneity in the preferences of the customer population). In the key, 'VB' refers to the method summarized herein, and 'VEB' to the use of VB to implement empirical Bayes (which we do not discuss here, for reasons of space)

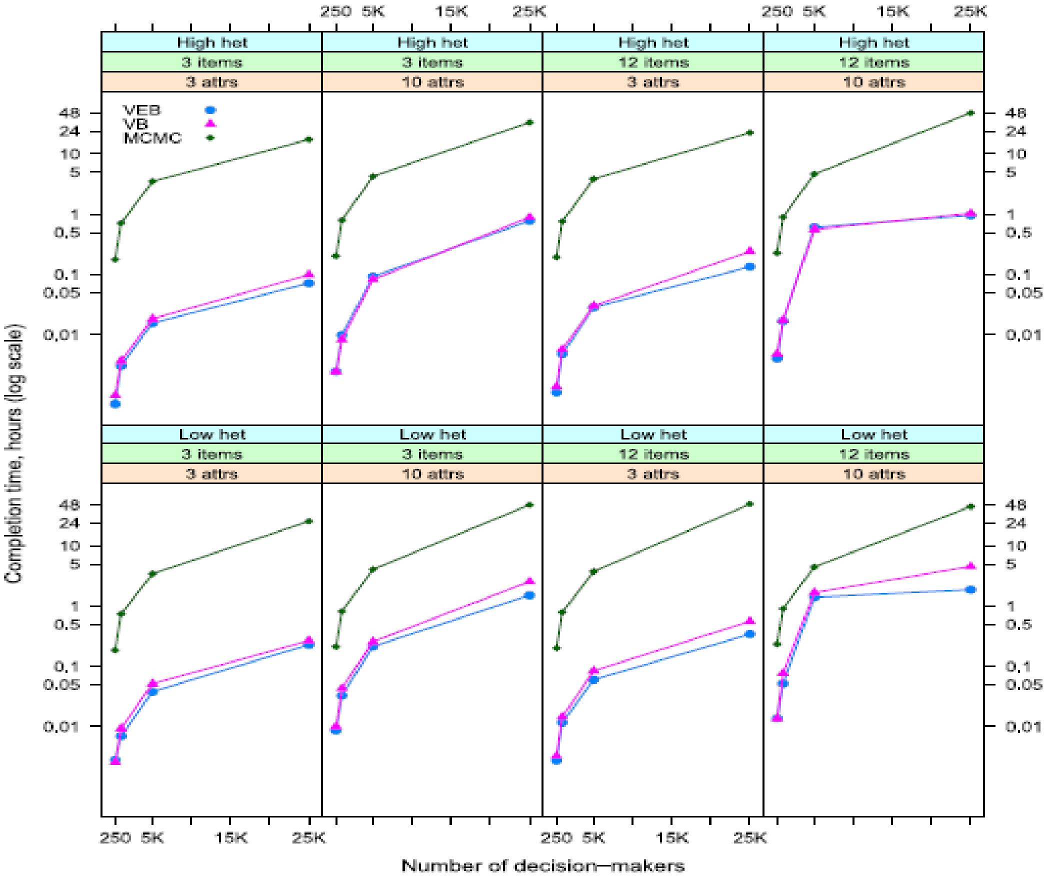


Figure 2: Figure 2 from Braun and McAuliffe (2010) with caption: "Timing results for Variational Empirical Bayes (VEB), Variational Hierarchical Bayes (VB), and MCMC. Within each panel, completion time is plotted on the log scale as a function of the number of agents, for fixed values of the other simulation parameters (shown at the top of each panel). In all simulated scenarios, variational methods complete more quickly than MCMC. With 25,000 decision-makers, the variational algorithms complete in five minutes to six hours, versus MCMC completion times of one to two days. In the 25,000 agent case, the figure shows the time to generate 6000 MCMC draws, based on a corresponding 1000-draw run (at which point the sampler exhausted memory resources)."

As in Section 3.1.4, we highlight the key messages to be gleaned from Figure 2:

1. For the scenario with $H = 25,000$, $T_h = 25$, $J = 12$ and $K = 10$ and high heterogeneity (top right-hand panel), MCMC uses two days of computation time to produce $6,000$ iterations, versus

one hour for VB. In the same setting, but with low heterogeneity (bottom right-hand panel), the comparison is two days versus 6 hours. That is, VB is between 8 and 48 times faster than MCMC.

2. For these two large-scale scenarios for J and K, as H (plotted on the horizontal axis) increases, VB also scales noticeably better to the consequent increase in $\mathbf{y}$ than does MCMC (i.e. the VB plots flatten more than do the MCMC plots).

3. Similar comparable relativities between the MCMC and VB computational burdens obtain for all other scenarios, although the superior scaling performance of VB is less noticeable.

4. One would expect the use of an SVI algorithm to greatly reduce the time taken to tackle the largest versions of the problem, and hence render the performance gains of VB over MCMC even more marked.

5. In addition to the speed comparison documented in Figure 2, the authors report (Braun and McAuliffe, 2010, Appendix A, Tables 1 and 2) that the accuracy with which the VB- and MCMC-based predictives match the true predictive choice distribution (known in this artificial data setting, and defined for an 'average agent' and a 'typical' item attribute) is almost identical. This result tallies with subsequent results in the VB literature (see, e.g. Quiroz *et al.*, 2018a, and Frazier *et al.*, 2021c), which demonstrate that predictive results obtained using VB are largely unaffected by the inferential inaccuracy of the VB posterior approximation.

INLA illustration

As a final illustration we report selected results from Margossian *et al.* (2020), in which a combination of INLA and HMC (referred to by the authors as the 'embedded' Laplace approximation) is applied to a spatial model for mortality counts in Finland. In brief, conditionally Poisson mortality counts (y_i), aggregated over 100 geographical regions ($i = 1, 2, ..., n = 100$), are modelled using a latent Gaussian process. Whilst the overarching aim of Margossian *et al.* is to adapt INLA to cater for a very high-dimensional hyperparameter vector ($\boldsymbol{\phi}$), and whilst INLA itself was developed for the case of a high-dimensional latent Gaussian field ($\mathbf{x}$), this illustrative example aims to compare the speed and accuracy of the embedded method with that of a full HMC algorithm; hence, both $\boldsymbol{\phi}$ and $\mathbf{x}$ are very low-dimensional. Specifically, for each region i, $y_i|\eta_i \sim Poisson(y_e^i \exp(\eta_i))$, where y_e^i is the standardized expected number of deaths, and η_i is a linear function of a two-dimensional vector of regional characteristics, $\mathbf{x}_i$. An exponentiated quadratic kernel defines the elements of $Q^{-1}(\boldsymbol{\phi})$ in (18), and the two-dimensional vector $\boldsymbol{\phi}$ comprises the standard deviation (α) and length scale (l) in the kernel function. (See Vanhatalo *et al.*, 2010, for all details of the general model structure in which the specification used by Margossian *et al.* is nested.)

We display (as our own Figure 3) Figure 2 from Margossian *et al.* (2020), including the original caption, which is sufficiently informative. We do note, however, that the authors use the notation θ_i to denote η_i, and they record – in addition to results for α and l – results for the first two elements, θ_1 ($\equiv \eta_1$) and θ_2 ($\equiv \eta_2$).

The key highlights of Figure 3 are as follows:

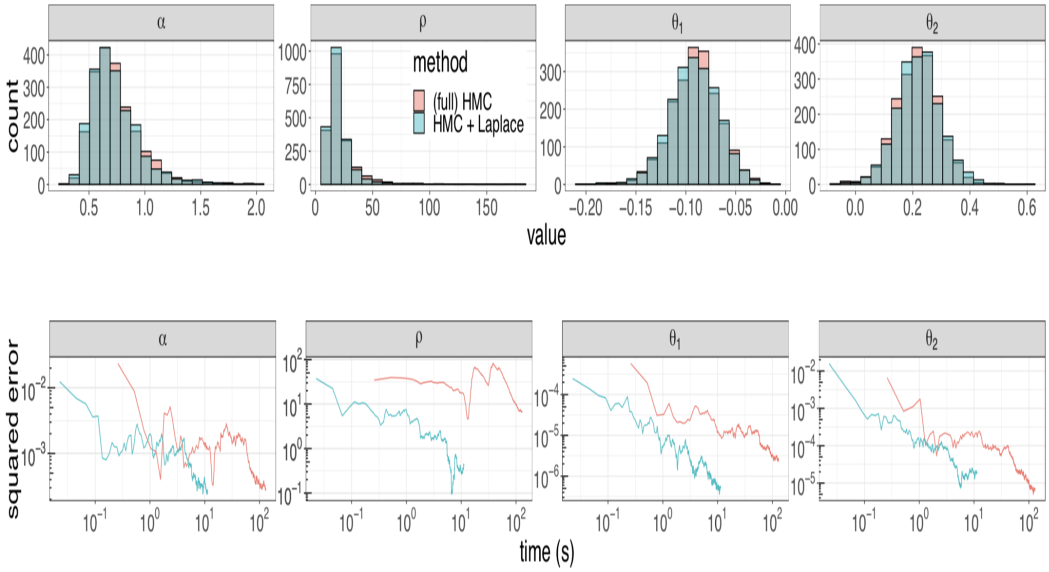


Figure 3: Figure 2 from Margossian *et al.* (2020) with caption: "(Up) Posterior samples obtained with full HMC and the embedded Laplace approximation when fitting the disease map. (Down) Error when estimating the expectation value against wall time. Unreported in the figure is that we had to fit full HMC twice before obtaining good tuning parameters."

1. For this example the marginal posteriors for the four unknowns (plotted in the top panel) produced by both the embedded Laplace approximation and the full HMC algorithm are very similar; with both algorithms based on 500 burn-in iterates and 500 subsequent draws.

2. At the same time, as documented in the bottom panel, the speed with which the embedded approach estimates the relevant posterior expectations produced from 18,000 HMC draws, to a certain level of precision, is an order of magnitude greater than the full HMC method.

3. Finally, on the matter of speed, the authors illustrate that the speed gains of the embedded method can be case-specific, depending, in part, on the relative dimensions of $\boldsymbol{\phi}$ and $\mathbf{x}$; and with particular gains to be had when the dimension of $\mathbf{x}$ is much larger than that of $\boldsymbol{\phi}$. Nevertheless, the authors do highlight that, even without dramatic speed gains, the use of INLA to integrate out $\mathbf{x}$ does avoid the delicate tuning required to implement HMC successfully in such a high-dimensional space.

3.3 Hybrid approximate methods

3.3.1 Overview

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 and INLA are much better equipped to deal with high-dimensional $\boldsymbol{\theta}$ (and/or $\mathbf{y}$), but require the evaluation of $p(\boldsymbol{\theta}, \mathbf{y})$ and, thus, $p(\mathbf{y}|\boldsymbol{\theta})$. Recently, hybrid algorithms that meld aspects of ABC, BSL and VB, along with so-called *pseudo-marginal* principles, have been used to deal with settings in which the likelihood is intractable *and* either $\boldsymbol{\theta}$ or $\mathbf{y}$, or both, are high-dimensional. A hybrid method that reduces the impact of dimensionality on ABC by introducing Gibbs steps has also been proposed. All such 'mixed' techniques are outlined below, after a very brief outline of pseudo-marginal MCMC.

3.3.2 A brief introduction to pseudo-marginal MCMC

The (combined) insight of Beaumont (2003) and Andrieu and Roberts (2009) began with the following observation. Use $\mathbf{u} \in \mathcal{U}$ to denote all of the canonical (problem-specific) random variables that may be used to produce an *unbiased* estimate of the likelihood function, $p(\mathbf{y}|\boldsymbol{\theta})$. Using the now standard concept of 'data augmentation' (Tanner and Wong, 1987), an MCMC scheme can then be applied to the joint space $(\boldsymbol{\theta}, \mathbf{u})$, in order to 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 $\mathbf{u}$ (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 $h(\boldsymbol{\theta}|\mathbf{y}) \propto \int_{\mathcal{U}} h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}) p(\boldsymbol{\theta}) h(\mathbf{u}) d\mathbf{u} = p(\boldsymbol{\theta}) E_{\mathbf{u}}[h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})] = p(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{y})$, as desired. That is, in yielding a chain with an invariant distribution equal to the correct marginal, $p(\boldsymbol{\theta}|\mathbf{y})$, a pseudo-marginal method produces an exact simulation-based estimate of (2).

Application of the pseudo-marginal principle to a Metropolis-Hastings MCMC algorithm involves substituting $h(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ for $p(\mathbf{y}|\boldsymbol{\theta})$ in the expression defining the acceptance probability (Chib and Greenberg, 1995), with the term pseudo-marginal Metropolis-Hastings used in this case. When the unbiased likelihood estimate is produced specifically via the use of *particle filtering* in a state space model, the term *particle* MCMC (PMCMC) has also been coined (Andrieu *et al.*, 2011).

Pseudo-marginal principles play a role in the hybrid methods in Sections 3.3.3 and 3.3.6 below.[12]

3.3.3 VB with intractable likelihoods

Tran *et al.* (2017) devise a hybrid VB/pseudo-marginal method for use when the likelihood function 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 $\boldsymbol{\lambda}$, so that $\mathcal{Q} := \{\boldsymbol{\lambda} \in \Lambda : q_{\boldsymbol{\lambda}}\}$. The variational approximation is then obtained by maximizing the ELBO, $\mathcal{L}(\boldsymbol{\lambda}) := \text{ELBO}[q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})]$, 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 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 $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, z) = q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) g(z|\boldsymbol{\theta})$. Whilst, in general, minimization of $\text{KL}[q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, z)|p(\boldsymbol{\theta}, z|\mathbf{y})]$ is not the same as minimization of $\mathcal{L}(\boldsymbol{\lambda})$, 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.* (2018) propose an alternative VB-based method for intractable likelihood problems. The authors begin with the recognition that establishing the conditions

[12] We refer to Doucet *et al.* (2015), Deligiannidis *et al.* (2018), Bardenet *et al.* (2017), Quiroz *et al.* (2018b), Quiroz *et al.* (2019) and Moores *et al.* (2020) for applications of pseudo-marginal MCMC methods (in their own right) to intractable problems.

under which the minimizers of $\mathrm{KL}[q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, z)|p(\boldsymbol{\theta}, z|\mathbf{y})]$ and $\mathcal{L}(\boldsymbol{\lambda})$ 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 (8). 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 $\mathrm{KL}[q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, z)|p(\boldsymbol{\theta}, z|\mathbf{y})]$ and $\mathcal{L}(\boldsymbol{\lambda})$ will always coincide.

While useful, it must be remembered that the approach of Ong *et al.* (2018) targets only the *partial* posterior $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$. Furthermore, given the discussion in Section 3.1.2, 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.

3.3.4 VB and ABC

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', $\mathbf{y}_1, \dots, \mathbf{y}_b$, 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:

$$p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y})) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{b} \left\{ \int p(\mathbf{z}_i|\mathbf{y}_{1:i-1}, \boldsymbol{\theta}) \mathbb{I}\left\{ \|\eta\left(\mathbf{z}_i\right) - \eta\left(\mathbf{y}_i\right)\| \leq \varepsilon \right\} d\mathbf{z}_i \right\} = p(\boldsymbol{\theta}) \prod_{i=1}^{b} \ell_i(\boldsymbol{\theta}), \tag{24}$$

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

The posterior in (24) is then approximated using expectation propagation (EP) (see Bishop, 2006, Chapter 10, for details). The EP approximation seeks to find a tractable density $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \in \mathcal{Q}$ that is close to $p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))$ by minimizing $\mathrm{KL}\left[p_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{y}))|q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right]$. The reader will note that this minimization problem is actually the reverse of the standard variational problem in (14), 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 mean vector and covariance matrix, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, respectively, $i = 1, \dots, b$, i.e., $\mathcal{Q} := \{\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_b) \in \Lambda : q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) := \prod_{i=1}^{b} q_{i,\boldsymbol{\lambda}_i}(\boldsymbol{\theta})\}$, this minimization problem is solved iteratively by minimizing the KL divergence between $\ell_i(\boldsymbol{\theta})$ and $q_{i,\boldsymbol{\lambda}_i}(\boldsymbol{\theta})$ 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 vector and covariance matrix of $q_{i,\boldsymbol{\lambda}_i}(\boldsymbol{\theta})$, 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 (24) 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.

3.3.5 ABC and Gibbs sampling

As described in Section 3.1.1, ABC methods suffer from a curse of dimensionality. Whilst this is typically expressed in terms of the dimensionality of the summary statistics, there is obviously an intrinsic link between the dimension of the summaries and that of $\boldsymbol{\theta}$ itself, with the dimension of $\boldsymbol{\theta}$ necessarily imposing a lower bound on the dimension of $\eta(\mathbf{y})$ that can be used to guarantee identification (see Frazier *et al.*, 2018). Nott *et al.* (2014) and Martin *et al.* (2019) both provide ways of alleviating this particular issue by advocating a separate selection process for individual elements (or blocks) of $\boldsymbol{\theta}$, with corresponding elements (or blocks) of $\eta(\mathbf{y})$ used in the matching process. Different criteria are used in each paper to define what is meant by 'corresponding'. This component-by-component approach is shown to produce more effective algorithms, within the confines of the particular examples explored, but no formal investigation or validation of the principle is undertaken in either piece of work.

Clarté *et al.* (2020) have attempted to formalize the approach, by combining the principles of ABC and Gibbs sampling, which they refer to as ABC-Gibbs (see also Kousathanas *et al.*, 2016, and Rodrigues *et al.*, 2019, for related work). In short, the vector $\boldsymbol{\theta}$ is blocked in a suitable way, and conditional posteriors defined as in a standard Gibbs sampling algorithm. Simulation from each posterior then occurs via an ABC step (along the lines of Algorithm 2, for example) but with a summary statistic chosen to be informative about the component of $\boldsymbol{\theta}$ that is the argument of the conditional posterior - a choice that is deemed to be easier than choosing informative summaries about the full $\boldsymbol{\theta}$. In the case where a conditional can be simulated from directly, the approximation step is not required.

While questions remain regarding the theoretical behaviour of the hybrid algorithm, the authors do establish some sufficient conditions for the convergence of the algorithm, with convergence being to the limiting distribution of reject/accept ABC in certain cases. They also demonstrate notable improvement in the numerical efficiency of the algorithm, in comparison with both reject/accept ABC and a particular ABC-SMC method.

3.3.6 ABC and PMCMC

Thus far we have discussed the use of ABC to conduct direct inference on the fixed, static or global parameters. In state space settings, in which both global and local parameters (or latent states) feature, ABC principles have also been used to the implement the particle filtering that is often required as an intermediate step towards conducting inference on the global unknowns. This has been particularly useful in cases where the measurement density has no closed form and, hence, cannot be used to define the particle weights in the usual way. In this case, the matching principle that underpins ABC is applied at a single observation level, one time point at a time, and without summarization if the data is one-dimensional. This process of 'ABC filtering' then provides a simulation-based estimate of the likelihood function. This can, in turn, be used either as a basis for producing frequentist point estimates of the parameters (Jasra *et al.*, 2012; Calvet and Czellar, 2015) or – in the spirit of this section on Bayesian hybrids – as an input into a PMCMC scheme (Dean *et al.*, 2014; Jasra, 2015).

4 Future Directions for Approximate Methods

We end our review of approximate Bayesian methods by documenting work that addresses the following questions: *1)* What are the implications for approximate computation if an assumed parametric model is misspecified?; *2)* What are the implications for approximate computation if the conventional likelihood-based paradigm is eschewed altogether, and a *generalized, robust* Bayesian, or *moment-based* approach to inference (e.g. Bissiri *et al.*, 2016; Chib *et al.*, 2018; Miller and Dunson, 2019; Loaiza-Maya *et al.*, 2021a) is adopted?; and *3)* What role can approximate computation play in Bayesian prediction?

1) Papers that address the first question (with reference to ABC, BSL and VB respectively) are as follows. *First*, 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. *Second,* similar to ABC, Frazier *et al.* (2021a) demonstrate that BSL displays non-standard behavior under model misspecification: depending on the nature and level of model misspecification, the BSL posterior may be approximately Gaussian, mixed-Gaussian, or concentrate onto the boundary of the parameter space. In a similar vein, 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. *Third*, Alquier and Ridgway (2020) and Zhang and Gao (2020) investigate posterior concentration of VB methods under model misspecification. Both pairs of authors demonstrate that the VB posterior concentrates onto the value that minimizes the Kullback-Leibler (KL) divergence from the true DGP.

2) With reference to the second question, 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 VB computational tool within that setting. In a somewhat similar spirit, Schmon *et al.* (2021) extend ABC to accommodate general loss functions, with Pacchiardi and Dutta (2021) applying a similar approach to deal with intractable likelihoods in the context of scoring rules.

Frazier *et al.* (2021c) also apply an approximate method in a setting in which a general loss function is specified, but with predictive accuracy dictating the form of the loss. A VB approximation of the resultant 'Gibbs posterior' (Zhang, 2006a; Zhang, 2006b; Jiang and Tanner, 2008) is adopted due the high dimensionality of the problems tackled. The authors prove theoretically, and illustrate numerically, that for a large enough value of n there is no reduction in predictive accuracy as a result of approximating the posterior via VB.

Finally, Tran *et al.* (2019) extend VB to manifolds, rather than using VB to approximate the conventional likelihood-based posterior.

3) The work by Frazier *et al.* (2021c) cited above continues in the vein of other work in which approximate computation plays a role in Bayesian prediction. Frazier *et al.* (2019a), for instance, produce an approximation of $p(y_{n+1}^{*}|\mathbf{y})$ in (3) by using an ABC-based posterior to replace $p(\boldsymbol{\theta}|\mathbf{y})$. The approximate predictive is numerically indistinguishable from the exact predictive (in the cases investigated), and yields equivalent out-of-sample accuracy as a consequence. Further, under the regularity that ensures Bayesian consistency for both the exact and ABC posteriors, the exact and approximate predictives are shown to be asymptotically equivalent. (See also Canale and Ruggiero, 2016, and Kon Kam King *et al.*, 2019.) Other work produces an approximate predictive by using a VB approximation of the (likelihood-based) posterior (Tran *et al.*, 2017; Quiroz *et al.*, 2018a; Koop and Korobilis, 2018; Chan and Yu, 2020; Loaiza-Maya *et al.*, 2021b). The tenor of this work is somewhat similar to that of Frazier *et al.* (2019a) and Frazier *et al.* (2021c); that is, computing the posterior via an approximate method does not necessarily reduce predictive accuracy. In contrast, Frazier *et al.* (2021b) document an important case where the approximation *can* matter. In brief, the use of a VB approximation to the posterior of the local parameters in a state space model *is* found to impinge on predictive accuracy in some cases, due to the lack of Bayesian consistency of the posterior for the global parameters that can arise.

In summary, approximate Bayesian methods are beginning to confront – and adapt to – the reality of misspecified DGPs, and the generalizations beyond the standard likelihood-based update that are increasingly adopted. Their good performance in many predictive settings is also encouraging. Being able to tackle intractable problems via an approximate method without compromising predictive accuracy is an attractive prospect for investigators, and suggests that approximate computation may play an increasingly large role in complex predictive settings, over and above its critical role in inference.

References

Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497.

An, Z., South, L. F., and Drovandi, C. (2019). BSL: An R package for efficient parameter estimation for simulation-based models via Bayesian synthetic likelihood.

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

Ardia, D., Baştürk, N., Hoogerheide, L., and van Dijk, H. K. (2012). A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis*, 56(11):3398–3414.

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.

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.

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.

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.

Beaumont, M. A. (2019). Approximate Bayesian computation. *Annual Review of Statistics and Its Application*, 6(1):379–403.

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

Berild, M. O., Martino, S., Gómez-Rubio, V., and Rue, H. (2021). Importance sampling with the integrated nested Laplace approximation.

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. 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*.

Bilodeau, B., Stringer, A., and Tang, Y. (2021). Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference.

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. and François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statist. Comput.*, 20:63–73.

Blum, M. G. (2017). Regression approaches for approximate Bayesian computation. *arXiv preprint arXiv:1707.01254*.

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.

Bortot, P., Coles, S. G., and Sisson, S. A. (2007). Inference for stereological extremes. *Journal of the American Statistical Association*, 102(477):84–92.

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.

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.

Carpenter, B., Gelman, A., Hoffman, M. D., 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*, 76(1):1–32.

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

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

Chan, J. C. and Yu, X. (2020). Fast and accurate variational inference for large Bayesian VARs with stochastic volatility. *CAMA Working Paper*.

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.

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

Clarté, G., Robert, C. P., Ryder, R. J., and Stoehr, J. (2020). Componentwise approximate Bayesian computation via Gibbs-like steps. *Biometrika*, 108(3):591–607.

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

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.

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

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

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. and Frazier, D. T. (2021). A comparison of likelihood-free methods with and without summary statistics. *arXiv preprint arXiv:2103.02407*.

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.

Elvira, V. and Martino, L. (2021). Advances in importance sampling.

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. 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.

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

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., Drovandi, C., and Nott, D. J. (2021a). Synthetic likelihood in misspecified models: Consequences and corrections. *arXiv preprint arXiv:2104.03436*.

Frazier, D. T., Loaiza-Maya, R., and Martin, G. M. (2021b). A note on the accuracy of variational Bayes in state space models: Inference and prediction.

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. (2021c). Loss-based variational Bayes prediction. *arXiv preprint arXiv:2104.14054*.

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*.

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

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

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.

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

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

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

Geyer, C. J. (2011). 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.

Gomez-Rubio, V. and Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28(5).

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.

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.

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.

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.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347.

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.

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

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, B. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1711–1721. PMLR.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and datamining. *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.

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*.

Kousathanas, A., Duchen, P., and Wegmann, D. (2019). A guide to general-purpose ABC software. In *Handbook of approximate Bayesian computation*, pages 369–413. Chapman and Hall/CRC.

Kousathanas, A., Leuenberger, C., Helfer, J., Quinodoz, M., Foll, M., and Wegmann, D. (2016). Likelihood-free inference in high-dimensional models. *Genetics*, 203(2):893–904.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

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.

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.

Llorente, F., Martino, L., Delgado, D., and Lopez-Santiago, J. (2021). Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *https://arXiv:2005.08334*.

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2021a). Focused Bayesian prediction. *Journal of Applied Econometrics*, 36(5):517–543.

Loaiza-Maya, R., Smith, M. S., Nott, D. J., and Danaher, P. J. (2021b). Fast and accurate variational inference for models with many latent variables. *Journal of Econometrics*.

Margossian, C. C., Vehtari, A., Simpson, D., and Agrawal, R. (2020). Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond. *https://arXiv:2004.12550*.

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

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., Frazier, D. T., and Robert, C. P. (2020). Computing Bayes: Bayesian computation from 1763 to the 21st century. *https://arXiv:2004.06425*.

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

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.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *J. American Statist. Assoc.*, 44:335–341.

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

Moores, M. T., Pettitt, A. N., and Mengersen, K. (2020). Bayesian computation with intractable likelihoods.

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.

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., Fan, Y., Marshall, L., and Sisson, S. A. (2014). Approximate Bayesian computation and Bayes' linear analysis: Toward high-dimensional ABC. *Journal of Computational and Graphical Statistics*, 23(1):65–86.

O'Hagan, A. and West, M. (2010). *The Oxford Handbook of Applied Bayesian Analysis.* OUP.

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

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

Pacchiardi, L. and Dutta, R. (2021). Generalized bayesian likelihood-free inference using scoring rules estimators.

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Artificial Intelligence and Statistics*, pages 398–407. PMLR.

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.

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

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.

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.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Robert, C. (2001). *The Bayesian Choice.* 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., Elvira, V., Tawn, N., and Wu, C. (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1435.

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

Rossi, P. E. and Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, 22(3):304–328.

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.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with inla: A review. *Annual Review of Statistics and Its Application*, 4(1):395–421.

Schmon, S. M., Cannon, P. W., and Knoblauch, J. (2021). Generalized posteriors in approximate Bayesian computation.

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. and Fan, Y. (2019). ABC samplers. *Handbook of Approximate Bayesian Computation*, pages 88–123. Chapman & Hall/CRC. Eds. Sisson, S., Fan, Y., Beaumont, M.

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.

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.

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

Stringer, A., Brown, P., and Stafford, J. (2021). Fast, scalable approximations to posterior distributions in extended latent Gaussian models.

Tang, Y. and Reid, N. (2021). Laplace and saddlepoint approximations in high dimensions.

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.

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.

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

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.

Turner, B. M. and Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic bulletin & review*, 21(2):227–250.

van Niekerk, J., Bakka, H., Rue, H., and Schenk, O. (2019). New frontiers in Bayesian modeling using the INLA package in R.

Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15):1580–1607.

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, 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.

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.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Yes, but did it work?: Evaluating variational inference. *Proceedings of the 35th International Conference on Machine Learning*, 80:5581–5590.

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*.

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. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207.

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.