Bayesian computation: a perspective on the current state, and sampling backwards and forwards

Peter Green, Krzysztof Latuszyski, Marcelo Pereyra, Christian Robert

Statistics and Computing manuscript No.
(will be inserted by the editor)

Bayesian computation: a perspective on the current state, and sampling backwards and forwards

Peter J. Green · Krzysztof Łatuszyński · Marcelo Pereyra · Christian P. Robert

Received: date / Accepted: date

Abstract The past decades have seen enormous improvements in computational inference based on statistical models, with continual enhancement in a wide range of computational tools, in competition. In Bayesian inference, first and foremost, MCMC techniques continue to evolve, moving from random walk proposals to Langevin drift, to Hamiltonian Monte Carlo, and so on, with both theoretical and algorithmic inputs opening wider access to practitioners. However, this impressive evolution in capacity is confronted by an even steeper increase in the complexity of the models and datasets to be addressed. The difficulties of modelling and then handling ever more complex datasets most likely call for a new type of tool for computational inference that dramatically reduce the dimension and size of the raw data while capturing its essential aspects. Approximate models and algorithms may thus be at the core of the next computational revolution.

Keywords Bayesian analysis · MCMC algorithms · ABC techniques · optimisation

Supported in part by "SuSTaIn", EPSRC grant EP/D063485/1, at the University of Bristol, and "i-like", EPSRC grant EP/K014463/1, at the University of Warwick. Krzysztof Łatuszyński holds a Royal Society University Research Fellowship, and Marcelo Pereyra a Marie Curie Intra-European Fellowship for Career Development. Peter Green also holds a Distinguished Professorship at UTS, Sydney, and Christian Robert a chair at Ceremade, Université Paris-Dauphine.

Peter Green and Marcelo Pereyra
School of Mathematics, University of Bristol
E-mail: P.J.Green, Marcelo.Pereyra@bristol.ac.uk

Krzysztof Łatuszyński and Christian P. Robert
Dept. of Statistics, University of Warwick
E-mail: K.G.Latuszynski@warwick.ac.uk, robert@ensae.fr

1 Introduction

One may reasonably balk at the terms "computational statistics" and "Bayesian computation" since, from its very start, statistics has always involved some computational step to extract information, something manageable like an estimator or a prediction, from raw data. This incomplete review of the recent past, current state, and immediate future of MCMC and related algorithms thus first requires us to explain what we mean by computation in a statistical context, before turning to what we perceive as medium term solutions and possible deadends.

Computations are an issue in statistics whenever processing a dataset becomes a difficulty, a liability, or even an impossibility. Obviously, the computational challenge varies according to the time when it is faced: what was an issue in the 19th century is most likely not so any longer so (take for instance the derivation of the moment estimates of a mixture of two normal distributions so painstakenly set by Pearson (1894) for estimating the ratio of "forehead" breadth to body length on a dataset of 1,000 crabs or the intense algebraic derivations found in the analysis of variance of the 1950's and 1960's (Searle et al. 1992)).

The introduction of simulation tools in the 1940's followed hard on the heels of the invention of the computer and certainly contributed an impetus towards faster and better computers, at least in the first decade of this revolution. This shows that these tools were both needed, and unavailable without electronic calculators. The introduction of Markov chain Monte Carlo is harder to pin down as some partial versions can be traced all the way back to 1944–45 and the Manhattan project at Los Alamos (Metropolis 1987). It is surprisingly much later, i.e., only by the early 1990's, that

such methods became part of the Bayesian toolbox, that is, some time after the devising of other computer-dependent tools like the bootstrap or the EM algorithm, and despite the availability of personal computers that considerably eased programming and experimenting (Robert and Casella 2010). It is presumably pointless to try to attribute this delay to a definite cause but a certain lack of probabilistic culture within the statistics community is probably partly to blame.

What makes this time-lag in MCMC methods becoming assimilated into the statistics community even more surprising is that fact that Bayesian inference having a significant role in statistical practice was really on hold pending the discovery of flexible computational tools that (implictly or explicitly) delivered values for the medium- to high-dimensional integrals that underpin the calculation of posterior distributions, in all but toy problems where conjugacy provided explicit answers. In fact, until Bayesians discovered MCMC, the only computational methodology that seemed to offer much chance of making practical Bayesian statistics practical was the portfolio of quadrature methods developed under Adrian Smith's leadership at Nottingham (Naylor and Smith 1982; Smith et al. 1985, 1987).

The very first article in the first issue of *Statistics and Computing*, whose quarter-century we celebrate in this special issue, was (to the journal's credit!) on Bayesian analysis, and was precisely in this direction of using clever quadrature methods to approach moderately high-dimensional posterior analysis (Dellaportas and Wright 1991). By the next (second) issue, sampling-based methods had started to appear, with three papers out of five in the issue on or related to Gibbs sampling (Verdinelli and Wasserman 1991; Carlin and Gelfand 1991; Wakefield et al. 1991).

Now, reflecting upon the evolution of MCMC methods over the 25 or so years they have been at the forefront of Bayesian inference, the focus has evolved a long way, from hierarchical models that extended the linear, mixed and generalised linear models (Albert 1988; Carlin et al. 1992; Bennett et al. 1996) which were initially the focus, and graphical models that stemmed from image analysis (Geman and Geman 1984) and artificial intelligence, to dynamical models driven by ODE's (Wilkinson 2011b) and diffusions (Roberts and Stramer 2001; Dellaportas et al. 2004; Beskos et al. 2006), hidden trees (Larget and Simon 1999; Huelsenbeck and Ronquist 2001; Chipman et al. 2008; Aldous et al. 2008) and graphs, aside with decision making in highly complex graphical models. While research on MCMC theory and methodology is still active and continually branching (Papaspiliopoulos et al. 2007; Andrieu and Roberts 2009; Łatuszyński et al. 2011; Douc and Robert 2011), progressively incorporating the capacities of parallel processors and GPUs (Lee et al. 2009; Jacob et al. 2011; Strid 2010; Suchard et al. 2010; Scott et al. 2013; Calderhead 2014), we wonder if we are not currently facing a new era where those methods are no longer appropriate to undertake the analysis of new models, and of new formulations where models are no longer completely defined. We indeed believe that imprecise models, incomplete information and summarised data will become, if not already, a central aspect of statistical analysis, due to the massive influx of data and the need to provide non-statisticians with efficient tools. This is why we also cover in this survey the notions of approximate Bayesian computation (ABC) and comment on the use of optimisation tools.

The plan of the paper is that in Sections 2 and 3 we discuss recent progress and current issues in Markov chain Monte Carlo and Approximate Bayesian Computation respectively. In Section 4, we highlight some araes of modern optimisation that, through lack of familiarity, are making less impact in the mainstream of Bayesian computation than we think justified. Our Discussion in Section 5 raises issues about data science and relevance to applications, and looks to the future.

2 MCMC, approximate simulations from an exact target

When MCMC techniques were introduced to the mainstream statistical (Bayesian) community in 1990, they were received with skepticism that they could one day become the central tool of Bayesian inference. For instance, despite the insurance provided by the ergodic theorem, many researchers thought at first that the convergence of those algorithms was a mere theoretical anticipation rather than a practical reality, in contrast to traditional Monte Carlo methods, and hence that they could not be trusted to provide "exact" answers. This perspective is obviously obsolete by now, when MCMC output is considered as "exact" as regular Monte Carlo, if possibly less efficient in some settings. Nowadays, MCMC is again attracting more attention (than in the past decade, say, where developments were more about alternatives, some of which described in the following sections), both because of methodological developments linked to better theoretical tools, for instance in the handling of stochastic processes, and because of new advances in accelerated computing via parallel and cloud computing.

2.1 Basics on MCMC

The introduction of Markov chain based methods within Monte Carlo thus took a certain degree of argument to reach the mainstream statistical community, when compared with other groups who were using MCMC methods 10 to 30 years earlier. It may sound unlikely at the current stage of our knowledge, but using methods that (a) generated correlated output, (b) required some burnin time to remove the impact of the initial distribution and (c) did not lead to a closed form expression for asymptotic variances were indeed met with resistance at first. As often, the immense computing advantages offered by this versatile tool soon overcame the reluctance to accept those methods as similarly "exact" as other Monte Carlo techniques, applications driving the move from the early 1990's. We reproduce below the generic version of the "all variables at once" Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970; Besag et al. 1995; Robert and Casella 2011) as it (still) constitutes in our opinion a fundamental advance in computational statistics, namely that, given a computable density π (up to a normalising constant) and a proposal Markov kernel $q(\cdot|\cdot)$, there exists a universal machine that returns a Markov chain with the proper stationary distribution, hence an associated operational MCMC algorithm.

Algorithm 1 Metropolis–Hastings algorithm (generic version)

Choose a starting value $\theta^{(0)}$)
for $t = 1$ to N **do**
Generate θ^* from a proposal $q(\cdot|\theta^{(t-1)})$
Compute the acceptance probability

$$\rho^{(t)} = 1 \wedge \pi(\theta^*)\, q(\theta^{(t-1)}|\theta^*))/\pi(\theta^{(t-1)} q(\theta^*|\theta^{(t-1)})$$

Generate $u_t \sim \mathcal{U}(0,1)$ and take $\theta^{(t)} = \theta^*$ if $u_t \leq \rho^{(t)}$, $\theta^{(t)} = \theta^{(t-1)}$ otherwise.
end for

The first observation about the Metropolis–Hastings is that the flexibility in choosing q is a blessing, but also a curse since the choice determines the performance of the algorithm. Hence a large part of the research on MCMC along the past 30 years (if we arbitrarily set the starting date at Geman and Geman (1984)) has been on choice of the proposal q to improve the efficiency of the algorithm, and in characterising its convergence properties. This typically requires gathering or computing additional information about π and we discuss some of the fundamental strategies in subsequent sections. Algorithm 1, and its variants in which variables are updated singly or in blocks according to some schedule, remains a keystone in standard use of MCMC methodology, even though the newer Hamiltonian Monte Carlo approach (see Section 2.3) may sooner or later come to replace it. While there is nothing intrinsically unique to the nature of this algorithm, or optimal in its convergence properties (other than the result of Peskun (1973) on the optimality of the acceptance ratio), attempts to bypass Metropolis–Hastings are few and limited. For instance, the birth-and-death process developed by Stephens (2000) used a continuous time jump process to explore a set of models, only to be later shown (Cappé et al. 2002) to be equivalent to the (Metropolis–Hastings) reversible jump approach of Green (1995).

Another aspect of the generic Metropolis–Hastings that became central more recently is that while the accept–reject step does overcome need to know the normalising constant, it still requires π, unnormalised, and this may be too expensive to compute or even intractable for complicated models and large datasets. Much recent research effort has been devoted to design and understanding of appropriate modifications that use estimators or approximations of π instead and we take the opportunity to summarise some of the progress in this direction.

2.2 MALA and Manifold MALA

Stochastic differential equations (SDEs) have been and still are informing Monte Carlo development in a number of seminal ways. A key insight is that the Langevin diffusion solving

$$dX_t = \frac{1}{2}\nabla \log \pi(X_t)dt + dB_t \tag{1}$$

has π as its stationary and limiting distribution. Here B_t is the standard Brownian motion and ∇ denotes gradient. The crude approach of sampling an Euler discretisation (Kloeden and Platen (1992)) of (1) and using it as an approximate sample from π was introduced in the applied literature (Ermak (1975); Doll and Dion (1976)). The method results in a Markov chain evolving according to the dynamics

$$\begin{aligned} X_{n+1}|X_n = x \sim\ & Q(X_n, \cdot) \\ & := x + \frac{h}{2}\nabla \log \pi(x) + h^{1/2} N(0, I_{d\times d}), \end{aligned} \tag{2}$$

for a chosen discretisation step h. There is a delicate tradeoff between accuracy of the approximation improving as $h \to 0$ and sampling efficiency (as measured e.g. by the effective sample size) improving when h increases. It was soon followed by its Metropolised version

(Rossky et al. (1978)) that uses the Euler approximation of (2) to produce a proposal in the Metropolis–Hastings algorithm 1, by letting $q(\cdot|\theta^{(t-1)}) := \theta^{(t-1)} + \frac{h}{2}\nabla \log \pi(\theta^{(t-1)}) + h^{1/2}N(0, I_{d\times d})$. While in the probability community Langevin diffusions and their equilibrium distributions had also been around for some time (Kent (1978)), it was the Roberts and Tweedie (1996a) paper that brought the approach to the centre of interest of the computational statistics community and sparked a systematic study, development and applications of Metropolis adjusted Langevin algorithms (hence MALA) and their cousins.

There is a large body of empirical evidence that at the extra price of computing the gradient, MALA algorithms typically provide a substantial speed-up in convergence on certain types of problems. However for very light-tailed distributions the drift term may grow to infinity and cause additional instability. More precisely, for distributions with sufficiently smooth contours, MALA is geometrically ergodic (c.f. Roberts and Rosenthal (2004)) if the tails of π decay as $\exp\{-|\theta|^\beta\}$ with $\beta \in [1, 2]$, while the random walk Metropolis algorithm is geometrically ergodic for all $\beta \geq 1$ (Roberts and Tweedie (1996a); Mengersen and Tweedie (1996)). This lack of geometrical ergodicity has been precisely quantified by Bou-Rabee and Hairer (2012).

Various refinements and extensions have been proposed. These include optimal scaling and choice of the discretisation step h, adaptive versions (both discussed in Section 2.4), combinations with proximal operators Pereyra (2013); Schreck et al. (2013), and applications and algorithm development for the infinite-dimensional context Pillai et al. (2012); Cotter et al. (2013). One particular direction of active research is considering a more general version of equation (1) with state dependent drift and diffusion coefficient

$$dX_t = \Big(\frac{\sigma(X_t)}{2}\nabla \log \pi(X_t) + \frac{\gamma(X_t)}{2}\Big)dt + \sqrt{\sigma}(X_t)dB_t \quad (3)$$
$$\gamma_i(X_t) = \sum_j \frac{\partial \sigma_{ij}(X_t)}{\partial X_j},$$

which also has π as invariant distribution (Xifara et al. (2014), c.f. Kent (1978)). The resulting proposals are

$$q(\cdot|\theta^{(t-1)}) := \frac{h}{2}\Big(\sigma(\theta^{(t-1)})\nabla \log \pi(\theta^{(t-1)}) + \gamma(\theta^{(t-1)})\Big) + h^{1/2}N(0, \sigma(\theta^{(t-1)})) + \theta^{(t-1)}.$$

Choosing appropriate σ for improved ergodicity is however nontrivial. The idea has been explored in Stramer and Tweedie (1999a,b); Roberts and Stramer (2002) and more recently Girolami and Calderhead (2010) introduced a mathematically-coherent approach of relating σ to a metric tensor on a Riemannian manifold of probability distributions. The resulting algorithms are termed Manifold MALA (MMALA), Simplified MMALA Girolami and Calderhead (2010), and position-dependent MALA (PMALA) Xifara et al. (2014), and differ in implementation cost, depending on how precise is the use they make of versions of equation (3). The approach still leaves the specification of the metric to be used in the space of probability distributions to the user, however there are some natural choices. For example, building the metric tensor from the spectrally-positive version of the Hessian of π and randomising the discretisation step size h results in an algorithm that is as robust as random walk Metropolis, in the sense that it is geometrically ergodic for targets with tail decay of $\exp\{-|\theta|^\beta\}$ for $\beta > 1$ (see Wolny (2014)).

2.3 Hamiltonian Monte Carlo

As with many improvements in the literature, starting with the very notion of MCMC, Hamiltonian (or hybrid) Monte Carlo (HMC) stems from Physics (Duane et al. 1987). After a slow emergence into the statistical community (Neal 1999), it is now central in statistical software like STAN (Stan Development Team 2014). For a complete account of this important flavour of MCMC, the reader is referred to Neal (2013), which inspired the description below; see also Betancourt et al. 2014 for a highly mathematical differential-geometry approach to HMC.

This method can be seen as a particular and efficient instance of auxiliary variables (see, e.g., Besag and Green 1993 and Rubinstein 1981), in which we apply a deterministic-proposal Metropolis method to the augmented target. In physical terms, the idea behind HMC is to add a "kinetic energy" term to the "potential energy" (negative log-target), leading to the Hamiltonian

$$H(q, p) = \log \pi(q) + p^{\mathrm{T}}M^{-1}p/2$$

where q denotes the object to be simulated (i.e., the parameter), p its speed or momentum and M the Hamiltonian matrix of π. In more statistical language, HMC creates an auxiliary variable q such that moving according to Hamilton's equations

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = \frac{\partial H}{\partial p} = M^{-1}p$$
$$\frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\frac{\partial \log \pi}{\partial q}$$

preserves the joint distribution with density $\exp -H(p, q)$, hence the marginal distribution of q, that is, $\pi(q)$. Hence, if we could simulate exactly this joint distribution of

(q, p), a sample from $\pi(q)$ would be a by-product. However, in practice, the equation is solved approximately and hence requires a Metropolis correction. As discussed in, e.g., Neal (2013), the dynamics induced by Hamilton's equations is reversible and volume-preserving in the (q, p) space, which means in practice that there is no need for a Jacobian in Metropolis updates. The practical implementation is called the *leapfrog approximation* (Girolami and Calderhead 2011) as it relies on a small step level ϵ, updating p and q via a modified Euler's method called the leapfrog that is reversible and preserves volume as well. This discretised update can be repeated for an arbitrary number of steps.

When considering the implementation via a Metropolis algorithm, a new value of the momentum p is drawn from the pseudo-prior $\pi(p) \propto \exp\{-p^{\mathrm{T}} M^{-1} p/2\}$ and it is followed by a Metropolis step, which proposal is driven by the leapfrog approximation to the Hamiltonian dynamics on (q, p) and which acceptance is governed by the Metropolis acceptance probability. What makes the potential strength of this augmentation (or dis-integration) scheme is that the value of $H(q, p)$ hardly changes during the Metropolis move, which means that it is most likely to be accepted and that it may produce a very different value of $\pi(q)$ without modifying the overall acceptance probability. In other words, moving along level sets is almost energy-free, but if the move proceeds for long "enough", the chain can reach faraway regions of the parameter space, thus avoid the myopia of standard MCMC algorithms. As explained in Neal (2013), this means that Hamiltonian Monte Carlo avoids the inefficient random walk behaviour of most Metropolis–Hastings algorithms. What drives the exploration of the different values of $H(q, p)$ is therefore the simulation of the momentum, which makes its calibration both quite influential and delicate as it depends on the unknown normalising constant of the target. (By calibration, we mean primarily the choice of the time discretisation step ε in the leapfrog approximation and of the number L of leapfrog leaps, but also the choice of the precision matrix M.)

2.4 Optimal scaling and Adaptive MCMC

The convergence of the Metropolis-Hastings algorithm 1 depends crucially on the choice of the proposal distribution q, as does the performance of both more complex MCMC and SMC algorithms, that often are hybrids using Metropolis–Hastings as simulation substeps.

Optimising over all implementable q appears to be a "disaster problem" due to its infinite-dimensional character, lack of clarity about what is implementable, what is not, and the fact that this optimal q must depend in a complex way on the target π to which we have only a limited access. In particular MALA provides a particular approach to constructing π-tailored proposals and HMC can be viewed as a combination of Gibbs and specific Metropolis moves for an extended target.

In this optimisation context, it is thus reasonable to restrict ourselves to some parametric family of proposals q_ξ, or more generally of Markov transition kernels P_ξ, where $\xi \in \Xi$ is a tuning parameter, possibly high dimensional.

The aim of adaptive Markov chain Monte Carlo is conceptually very simple. One expects that there is a set $\Xi_\pi \subset \Xi$ of good parameters ξ for which the kernel P_ξ converges quickly to π, and one allows the algorithm to search for Ξ_π "on the fly" and redesign the transition kernel during the simulation as more and more information about π becomes available. Thus an adaptive MCMC algorithm would apply the kernel $P_{\xi^{(t)}}$ to sample $\theta^{(t)}$ given $\theta^{(t-1)}$, where the tuning parameter $\xi^{(t)}$ is itself a random variable which may depend on the whole history $\theta^{(0)}, \ldots, \theta^{(t-1)}$ and $\xi^{(t-1)}$. Adaptive MCMC rests on the hope that the adaptive parameter $\xi^{(t)}$ will find Ξ_π, stay there essentially forever and inherit good convergence properties.

There are at least two fundamental difficulties in executing this strategy in practice. First, standard measures of efficiency of Markovian kernels, like the total variation convergence rate (c.f. Meyn and Tweedie (2009); Roberts and Rosenthal (2004)), $L^2(\pi)$ spectral gap (Diaconis and Stroock (1991); Roberts (1996); Saloff-Coste (1997); Levin et al. (2009)) or asymptotic variance (Peskun (1973); Geyer (1992); Tierney (1998)) in the Markov chain central limit theorem will not be available explicitly, and their estimation from a Markov chain trajectory is often an even more challenging task than the underlying MCMC estimation problem itself.

Secondly, when executing an adaptive strategy and trying to improve the transition kernel on the fly, the Markov property of the process is violated, therefore standard theoretical tools do not apply, and establishing validity of the approach becomes significantly more difficult. While the approach has been successfully applied in some very challenging practical problems (Solonen et al. (2012); Richardson et al. (2010); Griffin et al. (2014)), there are examples of seemingly reasonable adaptive algorithms that fail to converge to the intended target distribution (Bai et al. (2011); Łatuszyński et al. (2013)), indicating that compared to standard MCMC even more care must be taken to ensure validity of inferential conclusions.

While heuristics-based adaptive algorithms have been considered already in Gilks et al. (1994), a remarkable result providing a tool to coherently address the dif-

ficulty of optimising Markovian kernels is the Roberts et al. (1997) paper that considers settings of increasing dimensionality and investigates efficiency of the random walk Metropolis algorithm as a function of its average acceptance rate. More specifically, given a sequence of targets with iid components constructed from conveniently smooth marginal f,

$$\pi_d(\theta) := \prod_{i=1}^{d} f(\theta_i), \qquad \text{for} \quad d = 1, 2, \dots \tag{4}$$

consider a sequence of Markov chains $\mathbb{X}_d$, $d = 1, 2, \dots$, where the chain $\mathbb{X}_d = (X_d^{(t)})_{t=0,1,\dots}$ is a random walk Metropolis targeting π_d with proposal increments distributed as $N(0, \sigma_n^2 I_{d\times d})$.

It then turns out that the only sensible scaling of the proposal as dimensionality increases is to take $\sigma_d^2 = l^2 d^{-1}$. In this regime the sequence of time-rescaled first coordinate processes

$$Z_d^{(t)} := X_{d,1}^{(\lfloor td \rfloor)}, \qquad \text{for} \quad d = 1, 2, \dots$$

converges in a suitable sense to the solution Z of a stochastic differential equation

$$dZ_t = h(l)^{1/2} dB_t + \frac{1}{2} h(l) \nabla \log f(Z_t) dt.$$

Hence maximising the speed of the above diffusion $h(l)$ is equivalent to maximising the efficiency of the algorithm as the dimension goes to infinity. Surprisingly, there is a one-to-one correspondence between the value $l_{opt} = \operatorname{argmax} h(l)$ and the mean acceptance probability of 0.234.

The magic number 0.234 does not depend on f and gives a universal tuning recipe to be used for example in adaptive algorithms: choose the scale of the increment so that approximately 23% of the proposals are accepted.

The result, established under restrictive assumptions, has been empirically verified to hold much more generally, for non iid targets and also in medium- and even low-dimensional examples with d as small as 5. It has been also combined with relative efficiency loss due to mismatch between the proposal and target covariance matrices (see Roberts and Rosenthal (2001)). A large body of theoretical work extends optimal scaling formally to different and more general scenarios. For example Metropolis for smooth non iid targets has been addressed e.g. by Bédard (2007), and in infinite dimensional settings by Beskos et al. (2009). Discrete and other discontinuous targets have been considered in Roberts (1998) and Neal et al. (2012). For MALA algorithms an optimal acceptance rate of 0.574 has been established in Roberts and Rosenthal (1998) and confirmed in infinite-dimensional settings in Pillai et al. (2012) along with the stepsize $\sigma_d^2 = l^2 d^{-1/3}$. Hybrid Monte Carlo (see Section 2.3) has been analysed in a similar spirit by Beskos et al. (2013) resulting in an optimal acceptance rate of 0.651 and leapfrog step size $h = l \times d^{-1/4}$. These results not only inform about optimal tuning, but also provide an efficiency ordering on the algorithms in $d-$dimensions. Metropolis algorithms need $\mathcal{O}(d)$ steps to explore the state space, while MALA and HMC need respectively $\mathcal{O}(d^{1/3})$ and $\mathcal{O}(d^{1/4})$.

Further extensions include the transient phase before reaching stationarity (Christensen et al. (2005); Jourdain et al. (2012, 2014)), scaling of multiple-try MCMC (Bédard et al. (2012)), delayed rejection MCMC (Bédard et al. (2014)), and temperature scale of parallel tempering type algorithms (Atchadé et al. (2011); Roberts and Rosenthal (2014)). Interestingly, the optimal scaling of the discussed in Section 2.5 pseudo-marginal algorithms as obtained in Sherlock et al. (2014) suggests the acceptance rate of just 0.07.

While each of these numerous optimal scaling results gives rise, at least in principle, to an adaptive MCMC design, the pioneering and most successful algorithm is the Adaptive Metropolis of Haario et al. (2001). With its increasing popularity in applications, this has fuelled the development of the field.

Here one considers a normal increment proposal that estimates the target covariance matrix from past samples and applies appropriate dimension-dependent scaling and covariance shrinkage. Precisely, the proposal takes the form

$$q(\cdot|\theta^{(t-1)}) = N(\theta^{(t-1)}, C^{(t)}), \tag{5}$$

with the covariance matrix

$$C^{(t)} = \frac{(2.38)^2}{d} \Big(\hat{\text{cov}}(\theta^{(0)}, \dots, \theta^{(t-1)}) + \varepsilon I_{d\times d} \Big) \tag{6}$$

which is efficiently computed using a recursive formula.

Versions and refinements of the adaptive Metropolis algorithm Roberts and Rosenthal (2009); Andrieu and Thoms (2008) have served well in applications and motivated much of the theoretical development. These include, among many other contributions, adaptive Metropolis, delayed rejection adaptive Metropolis (Haario et al. (2006)), regional adaptation and parallel chains Craiu et al. (2009), and the robust version of Vihola (2012) estimating the shape of the distribution rather than its covariance matrix and hence suitable for heavy tailed targets.

Analogous development of adaptive MALA algorithms in Atchadé (2006); Marshall and Roberts (2012) and of adaptive Hamiltonian and Riemannian Manifold Monte Carlo in Wang et al. (2013) building on the adaptive scaling theory, resulted in a similar drastic mixing improvement as the original Adaptive Metropolis.

Another substantial and still unexplored area where adaptive algorithms are applied for very high dimensional and multimodal problems is model and variable selection (Nott and Kohn (2005); Richardson et al. (2010); Lamnisos et al. (2013); Ji and Schmidler (2013); Griffin et al. (2014)). These algorithms can incorporate reversible jump moves (Green 1995) and are guided by scaling limits for discrete distributions as well as temperature spacing of parallel tempering to address multimodality. Successful implementations allow for fully Bayesian variable selection in models with over 20 000 variables for which otherwise only ad hoc heuristic approaches have been used in literature.

To address the second difficulty with adaptive algorithms, several approaches have been developed to establish their theoretical underpinning. While for standard MCMC convergence in total variation and law of large numbers are obtained almost trivially, and the effort concentrates on stronger results, like CLTs, geometric convergence, nonasymptotic analysis, and, maybe most importantly, comparison and ordering of algorithms, adaptive samplers are intrinsically difficult. The most elegant and theoretically-valid strategy is to change the underlying Markovian kernel at regeneration times only (Gilks et al. (1998)). Unfortunately, this is not very appealing for practitioners since regenerations are difficult to identify in more complex settings and are essentially impractically rare in high dimensions. The original Adaptive Metropolis of Haario et al. (2001) has been validated (under some restrictive additional conditions) by controlling the dependencies introduced by the adaptation and using convergence results for mixingales. The approach has been further developed in Atchadé and Rosenthal (2005) and Atchadé (2006) to verify its ergodicity under weaker assumptions and apply the mixingale approach to adaptive MALA. Another successful approach (Andrieu and Moulines (2006) refined in Saksman and Vihola (2010)) rests on martingale difference approximations and martingale limit theorems to obtain, under suitable technical assumptions, versions of LLN and CLTs. There are close links between analysing adaptive MCMC and stochastic approximation algorithms and in particular the adaptation step can be often written as a mean field of the stochastic approximation procedure; Andrieu and Robert (2001); Atchade et al. (2011); Andrieu et al. (2015) contribute to this direction of analysis. Fort et al. (2011) develop an approach where both adaptive and interacting MCMC algorithms can be treated in the same framework. This allows to address "external adaptation" algorithms such as the interacting tempering algorithm (a simplified version of the celebrated equi-energy sampler of Kou et al. (2006)) or adaptive parallel tempering in Miasojedow et al. (2013).

We present here the rather general but fairly simple coupling approach (Roberts and Rosenthal (2007)) to establishing convergence. Successfully applied to a variety of adaptive Metropolis samplers under weak regularity conditions (Bai et al. (2011)), adaptive Gibbs and adaptive Metropolis within adaptive Gibbs samplers (Łatuszyński et al. (2013)), it shows that two properties *Diminishing Adaptation* and *Containment* are sufficient to guarantee that an adaptive MCMC algorithm will converge asymptotically to the correct target distribution. To this end recall the total variation distance between two measures defined as $\|\nu(\cdot) - \mu(\cdot)\| := \sup_{A\in\mathcal{F}} |\nu(A) - \mu(A)|$, and for every Markov transition kernel P_ξ, $\xi \in \Xi$ and every starting point $\theta \in \Theta$ define the ε convergence function $M_\varepsilon : \Theta \times \Xi \to \mathbb{N}$ as

$$M_\varepsilon(\theta, \xi) := \inf\{t \geq 1 : \|P^t_\xi(\theta, \cdot) - \pi(\cdot)\| \leq \varepsilon\}.$$

Let $\{(\theta^{(t)}, \xi^{(t)})\}_{t=0}^\infty$ be the corresponding adaptive MCMC algorithm and by $A^{(t)}((\theta, \xi), \cdot)$ denote its marginal distribution at time t, i.e.

$$A^{(t)}((\theta, \xi), B) := \mathbb{P}(\theta^{(t)} \in B | \theta^{(0)} = \theta, \xi^{(0)} = \xi).$$

The adaptive algorithm is ergodic for every starting values of θ and ξ if $\lim_{t\to\infty} \|A^{(t)}((\theta, \xi, \cdot) - \pi(\cdot)\| = 0$. The two conditions guaranteeing ergodicity are

Definition 1 (Diminishing Adaptation) The adaptive algorithm with starting values $\theta^{(0)} = \theta$ and $\xi^{(0)} = \xi$ satisfies Diminishing Adaptation, if

$$\lim_{t\to\infty} D^{(t)} = 0 \quad \text{in probability, where}$$
$$D^{(t)} := \sup_{\theta\in\Theta} \|P_{\xi^{(t+1)}}(\theta, \cdot) - P_{\xi^{(t)}}(\theta, \cdot)\|.$$

Definition 2 (Containment) The adaptive algorithm with starting values $\theta^{(0)} = \theta$ and $\xi^{(0)} = \xi$ satisfies Containment, if for all $\varepsilon > 0$ the sequence $\{M_\varepsilon(\theta^{(t)}, \xi^{(t)})\}_{t=0}^\infty$ is bounded in probability.

While diminishing adaptation is a standard requirement, Containment is subject to some discussion. On one hand, it may seem difficult to verify in practice; on the other, it may appear restrictive in the context of ergodicity results under some weaker conditions (c.f. Fort et al. (2011)). However, it turns out (Łatuszyński and Rosenthal (2014)) that if Containment is not satisfied, then the algorithm may still converge, but with positive probability it will be asymptotically less efficient than any nonadaptive ergodic MCMC scheme. Hence algorithms that do not satisfy Containment are termed AdapFail and are best avoided. The condition has been further studied in Bai et al. (2011) and is in particular

implied by simultaneous geometric or polynomial drift conditions of the adaptive kernels.

Given that adaptive algorithms may be incorporated in essentially any sampling scheme, their introduction seems to be one of the most important innovations of the last two decades. However, despite substantial effort and many ingenious contributions, the theory of adaptive MCMC lags behind practice even more than may be the case in other computational areas. While theory always matters, the numerous unexpected and counterintuitive examples of transient adaptive algorithms suggest that in this area theory matters even more for healthy development.

For adaptive MCMC to become a routine tool, a clear-cut result is need saying that under some easily verifiable conditions these algorithms are valid and perform not much worse than their nonadaptive counterpart with fixed parameters. Such a result is yet to be established and may require deeper understanding of how to construct stable adaptive MCMC, rather than aiming heavy technical artillery at algorithms currently in use without modifying them.

2.5 Estimated likelihoods and pseudo-marginals

There are numerous settings of interest where the target density $\pi(\cdot|x)$ is not available in closed form. For instance, in latent variable models, the likelihood function $\ell(\theta|x)$ is often only available as an intractable integral

$$\ell(\theta|x) = \int_{\mathcal{Z}} g(z, x|\theta)\,\mathrm{d}z\,,$$

which leads to

$$\pi(\theta|x) \propto \pi(\theta) \int_{\mathcal{Z}} g(z, x|\theta)\,\mathrm{d}z$$

being equally intractable. A solution proposed from the early days of MCMC (Tanner and Wong 1987) is to consider z as an auxiliary variable and to simulate the joint distribution $\pi(\theta, z|x)$ by a standard method, leading to simulating the marginal density $\pi(\cdot|x)$ as a by-product. However, when the dimension of the auxiliary variable z grows with the sample size, this technique may run into difficulties as induced MCMC algorithms are more and more likely to have convergence issues. An illustration of this case is provided by hidden Markov models, which have eventually to resort to particle filters as Markov chain algorithms become ineffective (Chopin 2007; Fearnhead and Clifford 2003). Another situation where the target density $\pi(\cdot|x)$ cannot be directly computed is the case of the "doubly intractable" likelihood (Murray et al. 2006a), when the likelihood function itself contains a term that is intractable $\ell(\theta|x) \propto g(x|\theta)$ and makes the normalising constant

$$\mathfrak{Z}(\theta) = \int_{\mathcal{X}} g(x|\theta)\,\mathrm{d}x$$

impossible to compute. Examples of this kind abound in Markov random fields models, as for instance for the Ising model (Murray et al. 2006b; Møller et al. 2006).

Both the approaches of Murray et al. (2006a) and Møller et al. (2006) require sampling data from the likelihood, which limits their applicability. The latter uses in addition an importance sampling function and retrospectively can be reinterpreted as Grouped Independence Metropolis-Hastings (GIMH of Andrieu and Roberts (2009), see below) with sample size 1. When perfect sampling from the likelihood is impossible, Girolami et al. (2013) develop an approach, also in the framework of GIMH, where the likelihoods are unbiasedly estimated by random truncation of their series expansions.

Andrieu and Roberts (2009) propose a more general resolution of such problems by designing a Metropolis–Hastings algorithm that replaces the intractable target density $\pi(\cdot|x)$ with an unbiased estimator, following an idea of Beaumont (2003). Proper changes to the Metropolis–Hastings acceptance ratio are sufficient to guarantee that the stationary density of the corresponding Markov chain remains equal to $\pi(\cdot|x)$. Indeed, if $\hat{\pi}(\theta|x, z)$ is an unbiased estimator of $\pi(\theta|x)$ when $z \sim q(\cdot|\theta)$, it is rather straightforward to check that the acceptance ratio

$$\frac{\hat{\pi}(\theta^*|x, z^*)}{\hat{\pi}(\theta|x, z)} \frac{q(\theta^*, \theta) q(z|\theta)}{q(\theta, \theta^*) q(z^*|\theta^*)}$$

preserves stationarity with respect to an extended target (see Andrieu and Roberts (2009)) for details) when $z \sim q(\cdot|\theta)$, $z^* \sim q(\cdot|\theta)$, and $\theta^*|\theta \sim q(\theta, \theta^*)$.

The performance of the approach will depend on the quality of the estimators $\hat{\pi}$ and so both stabilising them as well as understanding this relationship is an active area of current development. In particular, the improvements from using multiple samples of z to estimate π can be concluded from Andrieu and Vihola (2012) where the efficiency of the algorithm is studied in terms of its spectral gap and CLT asymptotic variance. Sherlock et al. (2014), on the other hand, investigate the efficiency as a function of the acceptance rate and variance of the noise.

Design and understanding of pseudo-marginal algorithms is a dynamic direction of methodological development that in the coming years will be further fuelled not only by complex models with intractable likelihoods, but also by the need of MCMC algorithms

for Big Data in contexts where the likelihood function can not be evaluated for the whole dataset (Korattikara et al. 2013; Bardenet et al. 2014; Teh et al. 2014; Maclaurin and Adams 2014; Minsker et al. 2014).

2.6 Particle MCMC

While we refrain from covering particle filters here, since others (names?!) are focussing on this technique, a recent advance at the interface between MCMC, pseudo-marginals, and particle filtering is the notion of particle MCMC (or *pMCMC*), developed by Andrieu et al. (2011). This is in facta specific case of a pseudo-marginal algorithm, taking advantage of the state-space models used by particle filters. And it differs from particle filters in that it targets (mostly) the marginal posterior distribution of the parameters.

The simplest setting in which pMCMC applies is one of a state-space model where a latent sequence $x_{0:T}$ is a Markov chain with joint density

$$p_0(x_0|\theta)p_1(x_1|x_0,\theta))\cdots p_T(x_T|x_{T-1},\theta)\,,$$

and is associated with an observed sequence y_{1+T} such that

$$y_{1+T}|X_{1:T},\theta \sim \prod_{i=1}^{T} q_i(y_i|x_i,\theta)\,.$$

The iterations of pMCMC are MCMC-like in that, at iteration t, a new value θ' of θ is proposed from an arbitrary transition kernel $\mathfrak{h}(\cdot|\theta^{(t)})$ and then a new value of the latent series $x'_{0:T}$n is generated from a particle filter approximation of $p(x_{0:T}|\theta',y_{1:T})$. Since the particle filter returns as a by-product (Del Moral et al. 2006) an unbiased estimator of the marginal posterior of $y_{1:T}$, $\hat{q}(y_{1:T}|\theta')$, this estimator can be used as such in the Metropolis–Hastings ratio

$$\frac{\hat{q}(y_{1:T}|\theta')\pi(\theta')\mathfrak{h}(\theta^{(t)}|\theta')}{\hat{q}(y_{1:T}|\theta)\pi(\theta^{(t)})\mathfrak{h}(\theta'|\theta^{(t)})}\wedge 1\,.$$

Its validation follows from the general argument of Andrieu and Roberts (2009), although some additional (notational) effort is needed to demonstrate all random variables used therein are correctly assessed (see Andrieu et al. 2011 and Wilkinson 2011a, the latter providing a very progressive introduction to the notions of pMCMC and particle Gibbs, which helped greatly in composing this section).

This approach is being used increasingly in complex dynamic models like those found in signal processing (Whiteley et al. 2010), dynamical systems like the PDEs in biochemical kinetics (Wilkinson 2011b) and probabilistic graphical models (Lindsten et al. 2014). An extension to approximating the sequential filtering distribution is found in Chopin et al. (2013).

2.7 Parallel MCMC

Since MCMC relies on local updating based on the current value of a Markov chain, opportunities for exploiting parallel resources, either CPU or GPU, would seem quite limited, In fact, the possibilities reach far beyond the basic notion of running independent or coupled MCMC chains on several processors. For instance, Craiu and Meng (2005) construct parallel antithetic coupling to create negatively correlated MCMC chains (see also Frigessi et al. 2000), while Craiu et al. (2009) use parallel exploration of the sample space to tune an adaptive MCMC algorithm. Jacob et al. (2011) exploit GPU facilities to improve by Rao-Blackwellisation the Monte Carlo approximations produced by a Markov chain, even though the parallelisation does not improve the convergence of the chain. See also Lee et al. (2009) and Suchard et al. (2010) for more detailed contributions on the appeal of using GPUs towards massive parallelisation, and Wilkinson (2005) for a general survey on the topic.

Another recently-explored direction is "prefetching". Based on Brockwell (2006) this approach computes the $2^2, 2^3, \ldots, 2^k$ values of the posterior that will be needed $2, 3, \ldots, k$ sweeeps ahead by simulating the possible "futures" of the Markov chain in parallel. Running a regular Metropolis–Hastings algorithm then means building a decision tree back to the current iteration and drawing $2, 3, \ldots, k$ uniform variates to go down the tree to the appropriate branch. As noted by Brockwell (2006), "in the case where one can guess whether or not acceptance probabilities will be 'high' or 'low', the tree could be made deeper down 'high' probability paths and shallower in the 'low' probability paths." This idea is exploited in Angelino et al. (2014), by creating "speculative moves" that consider the reject branch of the prefetching tree more often than not, based on some preliminary or dynamic evaluation of the acceptance rate. Using a fast but close-enough approximation to the true target (and a fixed sequence of uniforms) may also produce a "single most likely path" on which prefetched simulations can be run. The basic idea is thus to run simulations and costly likelihood computations on many parallel processors along a prefetched path, a path that has been prefetched for its high approximate likelihood. There are obviously instances where this speculative simulation is not helpful because the actual chain ends up following another path with the genuine target. Angelino et al. (2014) actually go further by constructing sequences of approximations for the precomputations. The proposition for the sequence found therein is to subsample the original data and use a normal approx-

imation to the difference of the log (sub-)likelihoods. See Strid (2010) for related ideas.

A different use of parallel capabilities is found in Calderhead (2014). At each iteration of Calderhead's algorithm, N proposed values are generated conditional on the "current" value of the Markov chain, which actually consists of $(N+1)$ components and from which one component is drawn at random to serve as a seed for the next proposal distribution and the simulation of N other values. In other words, this is a data-augmentation scheme with the index I on the one side and the N modified components on the other side. The neat trick in the proposal (and the reason for the gain in efficiency) is that the stationary distribution of the auxiliary variable can be determined and hence used $(N+1)$ times in updating the vector of $(N+1)$ components. (Note that picking the index at random means computing all $(N+1)$ possible transitions from one component to the N others, hence a potential increase in the computing cost, even though what costs the most is usually the likelihood computation, dispatched on the parallel processors.) While there are $(N+1)$ terms involved at each step, the genuine Markov chain is truly over a single chain and the N other proposed values are not recycled. An interesting feature in this approach is when the original Metropolis–Hastings algorithm is expressed as a finite state space Markov chain on the set of indices $\{1,\ldots,N+1\}$. Conditional on the values of the $(N+1)$ dimensional vector, the stationary distribution of that sub-chain is no longer uniform. Hence, picking $(N+1)$ indices from the stationary helps in selecting the most appropriate images, which explains why the rejection rate decreases. The paper indeed evaluates the impact of increasing the number of proposals in terms of effective sample size (ESS), acceptance rate, and mean squared jump distance, based two examples. Since this proposal is an almost free bonus resulting from using N processors, when compared with more complex coupled chains, it sounds worth investigating and comparing with those more complex parallel schemes.

Neiswanger et al. (2013) introduced the notion of embarrassingly parallel MCMC, where "embarrassing" refers to the "embarrassingly simple" solution proposed therein, namely to solve the difficulty in handling very large datasets by running completely independent parallel MCMC samplers on parallel threads or computers and using the outcomes of those samplers as density estimates, pulled together as a product towards an approximation of the true posterior density. In other words, the idea is to break the posterior as

$$p(\theta|x) = \prod_{i=1}^{m} p_i(\theta|x) \tag{7}$$

and to use the estimate

$$\hat{p}(\theta|x) = \prod_{i=1}^{m} \hat{p}_i(\theta|x)$$

where the individual estimates are obtained, say, nonparametrically. The method is then "asymptotically exact" in the weak (and unsurprising) sense of converging in the number of MCMC iterations. Still, there is a theoretical justification that is not found in previous parallel methods that mixed all resulting samples without accounting for the subsampling. And the point is made that, in many cases, running MCMC samplers with subsamples produces faster convergence. The decomposition of $p(\cdot)$ into its components is done by partitioning the iid data into M subsets and taking a power $1/m$ of the prior in each case. (This may induce issues about impropriety.) However, the subdivision is arbitrary and can thus be implemented in cases other than the fairly restrictive iid setting. Because each (subsample) nonparametric estimate involves T terms, the resulting overall estimate contains Tm terms and the authors suggest using an independent Metropolis sampler to handle this complexity. This is in fact necessary for producing a final sample from the (approximate) true posterior distribution.

In a closely related way, Wang and Dunson (2013) start from the same product representation of the target (posterior), namely, (7). However, they criticise the choice made by Neiswanger et al. (2013) to use MCMC approximations for each component of the product for the following reasons:

1. Curse of dimensionality in the number of parameters p;
2. Curse of dimensionality in the number of subsets m;
3. Tail degeneration;
4. Support inconsistency and mode misspecification.

While point 1 is clearly relevant, although there may be other ways than kernel estimation to mix samples from the terms in the product, terms Neiswanger et al. (2013) called the subposteriors, which is also a drawback with the current method, point 2 is not such a clearcut drawback: while the Tm explosion in the number of terms in a product of m sums of T terms sounds self-defeating, but Neiswanger et al. (2013) use a clever device to avoid the combinatorial explosion, namely operating on one component at a time. Having non-manageable targets is not such an issue in the post-MCMC era. Point 3 is formally correct, in that the kernel tail behaviour induces the kernel estimate tail behaviour, most likely disconnected from the true target tail behaviour, but this feature is true for any non-parametric estimate, even for the Weierstrass transform defined below, and

hence maybe not so relevant in practice. In fact, by lifting the tails up, the simulation from the subposteriors should help in visiting the tails of the true target. Finally, point 4 does not seem to be serious: assuming the true target can be computed up to a normalising constant, the value of the target for every simulated parameter could be computed, eliminating those outside the support of the product and highlighting modal regions.

The Weierstrass transform of a density f is a convolution of f and of an arbitrary kernel K. Wang and Dunson (2013) propose to simulate from the product of the Weierstrass transform, using a multi-tiered Gibbs sampler. Hence, the parameter is only simulated once and from a controlled kernel, while the random effects from the convolution are related with each subposterior. While the method requires coordination between the parallel threads, the components of the target are separately computed on a single thread. The clearest perspective on the Weierstrass transform may possibly be the rejection sampling version where simulations from the subpriors are merged together into a normal proposal on θ, to be accepted with a probability depending on the subprior simulations.

VanDerwerken and Schmidler (2013) keeps with the spirit of parallel papers like consensus Bayes (Scott et al. 2013), embarrassingly parallel MCMC (Neiswanger et al. 2013), Weierstrass MCMC (Wang and Dunson 2013), namely that the computation of the likelihood can be broken into batches and MCMC run over those batches independently. The idea of the authors is to replace an exploration of the whole space operated via a single Markov chain (or by parallel chains acting independently which all have to "converge") with parallel and independent explorations of parts of the space by separate Markov chains. The motivation is that "Small is beautiful": it takes a shorter while to explore each set of the partition, hence to converge, and, more importantly, each chain can work in parallel to the others. More specifically, given a partition of the space, into sets A_i with posterior weights w_i, parallel chains are associated with targets equal to the original target restricted to those A_is. This is therefore an MCMC version of partitioned sampling. With regard to the shortcomings listed in the quote above, the authors consider that there does not need to be a bijection between the partition sets and the chains, in that a chain can move across partitions and thus contribute to several integral evaluations simultaneously. It is somewhat unclear (a) whether or not this impacts ergodicity (it all depends on the way the chain is constructed, i.e., against which target) as it could lead to an over-representation of some boundary regions and (b) whether or not it improves the overall convergence properties of the chain(s). A more delicate issue with the partitioned MCMC approach stands with the partitioning. Indeed, in a complex and high-dimension model, the construction of the appropriate partition is a challenge in itself as we often have no prior idea where the modal areas are. Waiting for a correct exploration of the modes is indeed faster than waiting for crossing between modes, provided all modes are represented and the chain for each partition set A_i has enough energy to explore this set. It actually sounds unlikely that a target with huge gaps between modes will see a considerable improvement from the partioned version when the partition sets A_i are selected on the go, because some of the boundaries between the partition sets may be hard to reach with an off-the-shelf proposal. A last comment about this innovative paper is that the adaptive construction of the partition has much in common with Wang-Landau schemes (Wang and Landau 2001; Atchadé and Liu 2004; Lee et al. 2005; Jacob and Ryder 2014).

3 ABC and other acronyms, exact simulations from an approximate target

Motivated by highly complex models where MCMC algorithms and other Monte Carlo methods were too inefficient by far, approximate methods have emerged where the output cannot be considered as simulations from the genuine posterior, even under idealised situations of infinite computing power. These methods include ABC techniques, described in more details below, but also variational Bayes (Jaakkola and Jordan 2000), empirical likelihood (Owen 2001), INLA (Rue et al. 2009) and other solutions that rely on pseudo-models, or on summarised versions of the data, or both. It is quite important to signal this evolution as we think that it may be a central feature of computational Bayesian statistics in the coming years. From a statistical perspective, it also induces a somewhat paradoxical attitude where loss of information loss is balanced by improvement in precision, for a given computational budget. This perspective is not only interesting at the computational level but forces us (as statisticians) to re-evaluate in depth the nature of a statistical model and could produce a paradigm shift in the near future by giving a brand new meaning to George Box's motto that "all models are wrong".

3.1 ABC per se

It seems important to discuss ABC (Approximate Bayesian computation) in this partial tour of Bayesian compu-

tational techniques as (a) they provide the only approach to their model for some Bayesians, (b) they deliver samples in the parameter space that are exact simulations from a posterior of some kind (Wilkinson 2013), $\pi^{\text{ABC}}(\theta|x)$ if not the original posterior $\pi(\theta|x)$, (c) they may be more intuitive to some researchers outside statistics, as they entail simulating from the inferred model, i.e., going forward from parameter to data, rather than backward, from data to parameter, as in traditional Bayesian inference, (d) they can be merged with MCMC algorithms, and (e) they allow drawing inference directly from summaries of the data rather than the data itself.

ABC techniques play a role in the 2000s that MCMC methods did in the 1990s, in that they handle new models for which earlier (e.g., MCMC) algorithms were at a loss, in the same way the latter (MCMC) were able to handle models that regular Monte Carlo approaches could not reach, such as latent variable models (Tanner and Wong 1987; Diebolt and Robert 1994; Richardson and Green 1997). New models for which ABC unlocked the gate include Markov random fields, Kingman's coalescent for phylogeographical data, likelihood models with an untractable normalising constant, and models defined by their quantile function or their characteristic function. While the ABC approach first appeared a "quick-and-dirty" solution, to be considered only until more elaborate representations could be found, those algorithms have been progressively incorporated into the statistician's toolbox as a novel form of generic nonparametric inference handling partly-defined statistical models. They are therefore attractive as much for this reason as for being handy computational solutions when everything else fails.

A statistically intriguing feature of those methods is that they customarily require—for greater efficiency—replacing the data with (much) smaller-dimension summaries[1] or summary statistics, because of the complexity of the former. In almost every case calling for ABC, those summaries are not sufficient statistics and the method thus implies from the start a loss of statistical information, at least at a formal level since relying on the raw data is out of the question and therefore the additional information it provides is moot. This imposed reduction of the statistical information raises many relevant questions, from the choice of summary statistics (Blum et al. 2013) to the consistency of the ensuing inference (Robert et al. 2011).

[1] Maybe due to their initial introduction in population genetics, the oxymoron 'summary statistics' is now prevalent in descriptions of ABC algorithms, included in the statistical literature, where the (linguistically sufficient) term 'statistic' would suffice.

Although it has now diffused into a wide range of applications, the technique of Approximate Bayesian Computation (ABC) was first introduced by and for population genetics (Tavaré et al. 1997; Pritchard et al. 1999) to handle ancestry models driven by Kingman's coalescent and with strictly intractable likelihoods Beaumont (2010). The likelihood function of such genetic models is indeed "intractable" in the sense that, while derived from a fully defined and parameterised probability model, this function cannot be computed (at all or within a manageable time) for a single value of the parameter and for the given data. Bypassing the original example to avoid getting mired into the details of population genetics, examples of intractable likelihoods include densities with intractable normalising constants, i.e., $f(x|\theta) = g(y|\theta)/Z(\theta)$ such as in Potts (Potts 1952) and auto-exponential (Besag 1972) models, and pseudo-likelihood models (Cucala et al. 2009).

Example 1 A very simple illustration of an intractable likelihood is provided by Bayesian inference based on the median and median absolute deviation statistics of a sample from an arbitrary location-scale family, $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \sigma^{-1} g(\sigma^{-1}\{x-\mu\})$, as the joint distribution of this statistic is not available in closed form. ◀

The concept at the core of ABC methods can be seen as both very naïve and intrinsically related to the foundations of Bayesian statistics as *inverse probability* (Rubin 1984). This concept is that data x simulated conditional on values of the parameter close to the "true" value of the parameter should look more similar to the actual data x_0 than data x simulated conditional on values of the parameter far from the "true" value. ABC actually involves an acceptance/rejection step in that parameters simulated from the prior are accepted only when

$$d(x, x_0) < \epsilon\,,$$

where $d(\cdot,\cdot)$ is a distance and $\epsilon > 0$ is called the tolerance. It can be shown that the algorithm exactly samples the posterior when $\epsilon = 0$, but this is very rarely achievable in practice (Grelaud et al. 2009). An algorithmic representation is as follows:

Algorithm 2 ABC (basic version)

for $t = 1$ to N **do**
  repeat
    Generate θ^* from the prior $\pi(\cdot)$
    Generate $\mathbf{x}^*$ from the model $f(\cdot|\theta^*)$
    Compute the distance $\rho(\mathbf{x}^0, \mathbf{x}^*)$
    Accept θ^* if $\rho(\mathbf{x}^0, \mathbf{x}^*) < \epsilon$
  until acceptance
end for
return N accepted values of θ^*

Calibration of the ABC method in Algorithm 2 involves selecting the distance $\rho(\cdot, \cdot)$ and deducing the tolerance from computational cost constraints. However, in realistic settings, ABC is never implemented as such because comparing raw data to simulated raw data is rarely efficient, noise dominating signal (see, e.g., Marin et al. (2011) for toy examples). It is therefore natural that one first considers dimension-reduction techniques to bypass this curse of dimensionality. For instance, if rudimentary estimates $S(x)$ of the parameter θ are available, they are good candidates. In the ABC literature, they are called *summary statistics*, a term that does not impose any constraint on their form and hence leaves open the question of performance, as discussed in Marin et al. (2011); Blum et al. (2013). A more practical version of the ABC algorithm is shown in Algorithm 3 below, with a different output for each choice of the summary statistic. We stress in this version of the algorithm the construction of the tolerance ϵ as a quantile of the simulated distances $\rho(S(\mathbf{x}^0), S(\mathbf{x}^{(t)}))$, rather than an additional parameter of the method.

Algorithm 3 ABC (version with summary)

for $t = 1$ to N_{ref} **do**
  Generate $\theta^{(t)}$ from the prior $\pi(\cdot)$
  Generate $\mathbf{x}^{(t)}$ from the model $f(\cdot|\theta^{(t)})$
  Compute $d_t = \rho(S(\mathbf{x}^0), S(\mathbf{x}^{(t)}))$
end for
Order distances $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(N_{ref})}$
return the values $\theta^{(t)}$ associated with the k smallest distances

An immediate question about this approximate algorithm is how much it remains connected with the original posterior distribution and in case it does not, where does it draw its legitimacy. A first remark in this connection is that it constitutes at best a convergent approximation to the posterior distribution $\pi(\theta|S(y_0))$. It can easily be seen that ABC generates outcomes from a genuine posterior distribution when the data is randomised with scale ϵ (Wilkinson 2013; Fearnhead and Prangle 2012). This interpretation indicates a decrease in the precision of the inference but it does not provide a universal validation of the method. A second perspective on the ABC output is that it is based on a nonparametric approximation of the sampling distribution (Blum 2010; Blum and François 2010), connected with both indirect inference (Drovandi et al. 2011) and k-nearest neighbour estimation (Biau et al. 2014). While a purely Bayesian nonparametric analysis of this aspect has not yet emerged, this brings an additional if cautious support for the method.

Example 2 Continuing from the previous example of a location-scale sample only monitored through the pair median plus mad statistic, we consider the special case of a normal sample $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$, with $n = 100$. Using a conjugate prior $\mu \sim \mathcal{N}(0, 10)$, $\sigma^{-2} \sim \mathcal{G}a(2, 5)$, we generated 10^6 parameter values, along with the corresponding pairs of summary statistics. When creating the distance $\rho(\cdot, \cdot)$, we used both following versions:

$$\rho_1(S(\mathbf{x}^0), S(\mathbf{x})) = |\mathrm{med}(\mathbf{x}^0) - \mathrm{med}(\mathbf{x})|/\mathrm{mad}(\mathrm{med}(\mathbf{X})) + |\mathrm{mad}(\mathbf{x}^0) - \mathrm{mad}(\mathbf{x})|/\mathrm{mad}(\mathrm{mad}(\mathbf{X}))$$

$$\rho_2(S(\mathbf{x}^0), S(\mathbf{x})) = |\mathrm{med}(\mathbf{x}^0) - \mathrm{med}(\mathbf{x})|/\mathrm{mad}(\mathrm{med}(\mathbf{X})) + |\log \mathrm{mad}(\mathbf{x}^0) - \log \mathrm{mad}(\mathbf{x})|/\mathrm{mad}(\log \mathrm{mad}(\mathbf{X}))$$

where the denominators are computed from the reference table in order to scale the components properly. Figure 1 shows the impact of the choice of this distance, but even more clearly the discrepancy between inference based on the ABC and the true inference on (μ, σ^2).

The discrepancy can however be completely eliminated by post-processing: Figure 2 reproduces Figure 1 by comparing the histograms of an ABC sample with the version corrected by Beaumont et al.'s (2002) local regression, as the latter is essentially equivalent to a regular Gibbs output. ◀

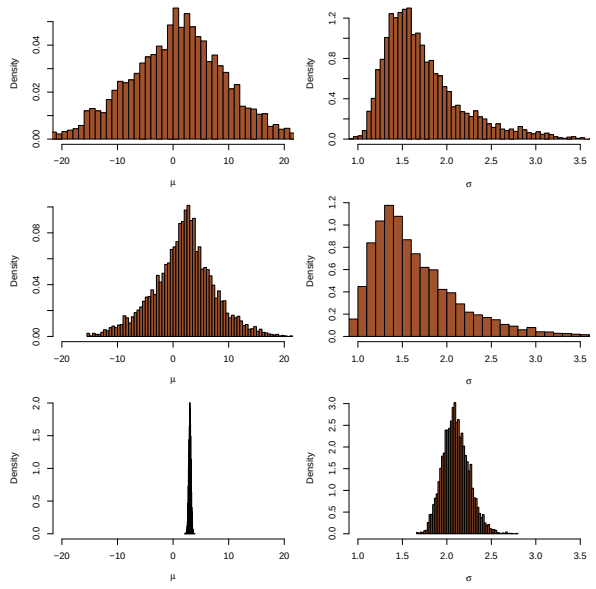

Fig. 1 Comparison of the posterior distributions on μ *(left)* and σ *(right)* when using an ABC algorithm 3 with distance ρ_1 *(top)* and ρ_2 *(central)*, and when using a standard Gibbs sampler *(bottom)*. All three samples are based on the same number of subsampled parameters. The dataset is a $\mathcal{N}(3, 2^2)$ sample and the tolerance value ϵ corresponds to $\alpha = .5\%$ of the reference table.

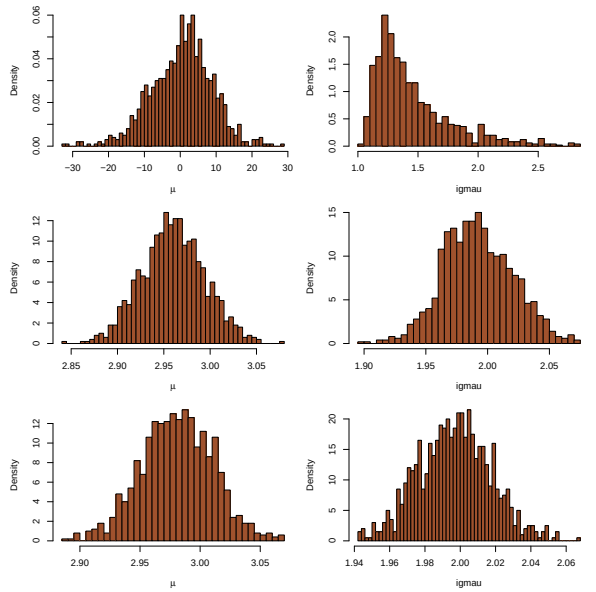

Fig. 2 Comparison of the posterior distributions on μ *(left)* and σ *(right)* when using an ABC algorithm 3 with distance ρ_1 *(top)*, a post-processed version by Beaumont et al.'s (2002) local regression *(central)*, and when using a standard Gibbs sampler *(bottom)*. The simulation setting is the same as in Figure 1.

Barber et al. (2013) studies the rate of convergence for ABC algorithms through the mean square error when approximating a posterior moment. They show the convergence rate is of order $\mathrm{O}(n^{2/q+4})$, when q is the dimension of the ABC summary statistic, associated with an optimal tolerance in $\mathrm{O}(n^{-1/4})$. Those rates are connected with the nonparametric nature of ABC, as already suggested in the earlier literature: for instance, Blum (2010), who links ABC with standard kernel density non-parametric estimation and find a tolerance (re-expressed as a bandwidth) of order $n^{-1/q+4}$) and an mse of order $^2/_{q+4}$ as well, while Fearnhead and Prangle (2012) obtain similar rates, with a tolerance of order $n^{-1/q+2}$ for noisy ABC. See also Calvet and Czellar (2014). Similarly, Biau et al. (2014) obtain precise convergence rates for ABC interpreted as a k-nearest-neighbour estimator.

Lee and Łatuszyński (2014) have also produced precise characterisations of the geometric ergodicity or lack thereof of ABC-MCMC algorithms. Among four versions of ABC algorithms, from the standard ABC-MCMC (with N replicates of the simulated pseudo-data to each simulated parameter value) to versions involving simulations of the replicates repeated at the subsequent step, use of a stopping rule in the generation of the pseudo data, and a "gold-standard algorithm based on the (unavailable) measure of an ϵ ball around the data. Based a result by Roberts and Tweedie (1996b), also used in Mengersen and Tweedie (1996), namely that an MCMC chain cannot be geometrically ergodic when there exist almost-absorbing states, they derive that (under some technical assumptions) the first two versions above cannot be variance-bounding (i.e., that the spectral gap is zero), while the last two versions can be both variance-bounding and geometrically ergodic under some appropriate conditions on the prior and the above ball measure. This result is thus rather striking in simulating a *random* number of auxiliary variables is sufficient to produce geometric ergodicity. We note that this result does not contradict the parallel result of Bornn et al. (2014), who establish that there is no efficiency gain in simulating $N > 1$ replicates of the pseudo-data, since there is no randomness involved in that approach. However, the latter approach only applies to functions with finite variances.

When testing hypotheses and selecting models, the Bayesian approach relies on modelling hypotheses and model indices as part of the parameter and hence ABC naturally operates as this level as well, as demonstrated in Algorithm 4 following Cornuet et al. (2008), Grelaud et al. (2009) and Toni et al. (2009). In fields like population genetics, model choice and hypotheses validation is presumably the primary motivation for using ABC methods as exemplified in Belle et al. (2008); Cornuet et al. (2010); Excoffier et al. (2009); Ghirotto et al.

(2010); Guillemaud et al. (2009); Leuenberger and Wegmann (2010); Patin et al. (2009); Ramakrishnan and Hadly (2009); Verdu et al. (2009); Wegmann and Excoffier (2010). It is also the area that concentrates most of the criticisms addressed against ABC: while some are easily dismissed (see, e.g., Templeton 2008, 2010; Beaumont et al. 2010; Berger et al. 2010), the impact of the choice of the summary statistics on the value of the posterior probability remains a delicate issue that prompted Pudlo et al. (2014) to advocate the alternative use of a posterior predictive error.

Algorithm 4 ABC (model choice)

for $i = 1$ to N **do**
  Generate $\mathfrak{M}$ from the prior $\pi(\mathcal{M} = m)$
  Generate $\theta_{\mathfrak{M}}$ from the prior $\pi_{\mathfrak{M}}(\theta_{\mathfrak{M}})$
  Generate $\mathbf{z}$ from the model $f_{\mathfrak{M}}(\mathbf{z}|\theta_{\mathfrak{M}})$
  Compute the distance $\rho\{S(\mathbf{z}), S(\mathbf{y})\}$
  Set $\mathfrak{M}^{(i)} = \mathfrak{M}$ and $\theta^{(i)} = \theta_{\mathfrak{M}}$
end for
return the values $\mathfrak{M}^{(i)}$ associated with the k smallest distances

Indeed, Robert et al. (2011) pointed out the potential irrelevance of ABC-based posterior probabilities, due to the possible ancillarity (for model choice) of summary statistics, as also explained in Didelot et al. (2011). Marin et al. (2014) consider for instance the comparison of normal and Laplace fits on both normal and Laplace samples and show that using sample mean and sample variance as summary statistics produces Bayes factors converging to values near 1, instead of the consistent 0 and $+\infty$.

Marin et al. (2014) analyses this phenomenon with the aim of producing a necessary and sufficient consistency condition on summary statistics. Quite naturally, the summaries that are acceptable must display different behaviour under both models, in the guise of ranges of means $\mathbb{E}_\theta[S(\mathbf{y})]$ that do not intersect for both models. (In the counter-example of the normal-Laplace test, the expectations of the sample mean and variance can be recovered under both models.) This characterisation then leads to a practical asymptotic test validating summary statistics and to the realisation that a larger number of summaries helps in achieving this goal (while degrading the estimated tolerance). More importantly, it shows that the reduction of information represented by an ABC approach may prevent discriminating between models, at least when trying to recover the Bayes factor. In the end, this is a natural consequence of simplifying the description of both the data and the model, and can be found in most limited information settings.

3.2 More fish in the alphabet soup

Besides ABC, approximation techniques have spread wide and far towards analysing more complex or less completely defined models. Rather than a confusion, this multiplicity of available approximations is beneficial both to the understanding of the underlying model and to the calibration of those different methods.

Variational Bayes methods have been proposed for at least two decades to substitute exponential families $q(\theta|\lambda)$ for complex posterior distributions $\pi(\theta)$ (Jordan et al. 1999; MacKay 2002). The central notion in those methods is that the exponential family structure and a so-called mean-field representation of the approximation

$$q(\theta|\lambda) = \prod_{i=1}^{k} q_i(\theta_i|\lambda_i)$$

allows for a sometimes closed-form minimisation of the Kullback-Leibler distance $\mathrm{KL}(q(\theta|\lambda), \pi(\theta))$ between the true target and its approximation. If not, the setting is quite congenial to the use of EM algorithms (Paisley et al. 2012). See Salimans and Knowles (2013) for a contemporary view on this approach, which offers considerable gains in terms of computing time, while being difficult to assess in terms of discrepancy with the "truth", i.e., the outcome that would result from using the genuine posterior.

Another approach that has met with considerable interest in the past five years is Integrated nested Laplace approximation (INLA) (Rue et al. 2009). The method operates on latent Gaussian random fields, with likelihoods of the form

$$\prod_{i=1}^{n} f(y_i|\eta_i, \theta)\,,$$

where the y_i's are the observables and the η_i's are latent variables. Using Laplace approximations to the marginal distributions $\pi(\theta|\mathbf{y})$ and to $f(\eta|\mathbf{Y})$, INLA produces fast and accurate approximations of the true posterior distribution as well as of the marginal likelihood value. Thanks to the availability of a well-constructed package called R-INLA, this approach has gathered a large group of followers.

A somewhat exotic example of variational approximation is expectation-propagation (EP) (Minka 2001), which starts from an arbitrary decomposition of the target distribution

$$\pi(\theta) = \prod_{j=1}^{k} \pi_j(\theta)$$

(often inspired by a likelihood decomposition into groups of observations) and iteratively approximate each term

π_j in the product by a density member of an exponential family, $\nu(\cdot|\lambda)$m using the other approximations as a marginal. Given the current approximation of $\pi(\theta)$ at iteration t,

$$\nu(\theta|\lambda_t) = \prod_{j=1}^{k} \nu_j(\theta|\lambda_t)\,,$$

where λ_t is the current value of the hyperparameter, the t-th step in the expectation-propagation (EP) algorithm goes as follows:

1. Select $1 \leq j \leq k$ at random
2. Define the marginal

$$\nu_{-j}(\theta|\lambda_t) \propto \frac{\nu(\theta|\lambda_t)}{\nu_j(\theta|\lambda_t)}\,;$$

3. Update the hyperparameter λ_t by solving

$$\lambda_{t+1} = \underset{\lambda}{\operatorname{argmin}}\, \mathrm{KL}\left\{\pi_j(\theta)\nu_{-j}(\theta|\lambda_t), \nu(\theta|\lambda)\right\}$$

4. Update $\nu_j(\theta|\lambda_t)$ as

$$\nu_j(\theta|\lambda_{t+1}) \propto \frac{\nu(\theta|\lambda_{t+1})}{\nu_{-j}(\theta|\lambda_t)}\,.$$

(In the above, KL denotes the Kullback-Leibler divergence.) The algorithm stops at stationarity. The convergence of this approach is not yet fully understood, but Barthelmé and Chopin (2014) consider expectation-propagation as a practical substitute for ABC, avoiding the selection of summary statistics by using a local constraint

$$||y_i - y^{\mathrm{obs}}|| \leq \epsilon$$

on each element of the simulated pseudo-data vector, y^{obs} being the actual data. In addition, expectation-propagation provides an approximation of the evidence. In the ABC setting, when using a Normal distribution as the exponential family default, implementing EP means computing empirical mean and empirical variance, one observation at a time, under the above tolerance constraint. Obviously, using a Normal candidate means that the final approximation will also look much like a Normal distribution, which both links with other Normal approximations like INLA and variational methods, and signals a difficulty with EP in less smooth cases, such as ridge-like or multimodal posteriors.

While different approximations keep being developed and tested, with arguments ranging from efficient programming, to avoiding simulations, to having an ability to deal with more complex structures, their drawback is the overall incapacity to assess the amount of approximation involved. Bootstrap evaluations can be attempted in the simplest cases but cannot be extended to more realistic situations.

4 Optimisation in modern Bayesian computation

Optimisation methodology for high-dimensional maximum-a-posteriori (MAP) estimation is another area of Bayesian computation that has received a lot of attention over the last years, particularly for problems related to machine learning, signal processing and computer vision. One reason for this is that for many Bayesian models optimisation is significantly more computationally tractable than integration. This has generated a lot of interest in MAP estimators, especially for applications involving very high-dimensional parameter spaces or tight computing time constraints, for which calculating other summaries of the posterior distribution is not feasible. Here we review some of the major breakthroughs in this topic, which originated mainly outside the statistics community. We focus on developments related to high-dimensional convex optimisation, though many of the techniques discussed below are also useful for non-convex optimisation. In particular, in Section 4.1 we concentrate on *proximal optimisation algorithms*, a powerful class of iterative methods that exploit tools from convex analysis, monotone operator theory and theory of non-expansive mappings to construct carefully designed fixed-point schemes. We refer the reader to the excellent book by Bauschke and Combettes (2011) for the mathematics underpinning proximal optimisation algorithms, and to the recent tutorial papers by Combettes and Pesquet (2011), Cevher et al. (2014) and Parikh and Boyd (2014) for an overview of the field and applications to signal processing and machine learning.

4.1 Proximal algorithms

Similarly to many other computational methodologies that are widely used nowadays, proximal algorithms were first proposed several decades ago by Moreau (1962), Martinet (1970) and Rockafellar (1976), and regained attention recently in the context of large-scale inverse problems and "big data".

We consider the computation of maximisers of posterior densities $\pi(\boldsymbol{\theta}) = \exp\{-g(\boldsymbol{\theta})\}/\kappa$ that are high-dimensional and log-concave, which we formulate as

$$\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\operatorname{argmin}}\, g(\boldsymbol{\theta}) \tag{8}$$

where g belongs to the class $\Gamma_0(\mathbb{R}^n)$ of lower semicontinuous convex functions from $\mathbb{R}^n \to (-\infty, +\infty]$. Notice that g may be non-differentiable and take value $g(\boldsymbol{\theta}) = +\infty$, reflecting constraints in the parameter space. In order to introduce proximal algorithms we

first recall the following standard definitions and results from convex analysis: We say that $\boldsymbol{\varphi} \in \mathbb{R}^n$ is a subgradient of g at $\boldsymbol{\theta} \in \mathbb{R}^n$ if it satisfies $(\mathbf{u}-\boldsymbol{\theta})^T\boldsymbol{\varphi} + g(\boldsymbol{\theta}) \leq g(\mathbf{u}), \forall \mathbf{u} \in \mathbb{R}^n$. The set of all such subgradients defines the subdifferential set $\partial g(\boldsymbol{\theta})$, and $\hat{\boldsymbol{\theta}}_{MAP}$ is a minimiser of g if and only if $\mathbf{0} \in \partial g(\hat{\boldsymbol{\theta}}_{MAP})$. The (convex) conjugate of $g \in \Gamma_0(\mathbb{R}^n)$ is the function $g^* \in \Gamma_0(\mathbb{R}^n)$ defined as $g^*(\boldsymbol{\varphi}) = \sup_{\mathbf{u}\in\mathbb{R}^n} \mathbf{u}^T\boldsymbol{\varphi} - g(\mathbf{u})$. The subgradients of g and g^* satisfy the property $\boldsymbol{\varphi} \in \partial g(\boldsymbol{\theta}) \Leftrightarrow \boldsymbol{\theta} \in \partial g^*(\boldsymbol{\varphi})$.

Proximal algorithms take their name from the proximity mapping, defined for $g \in \Gamma_0(\mathbb{R}^n)$ and $\lambda > 0$ as (Moreau 1962)

$$\text{prox}_g^\lambda(\boldsymbol{\theta}) = \underset{\mathbf{u}\in\mathbb{R}^n}{\text{argmin}}\ g(\mathbf{u}) + \|\mathbf{u}-\boldsymbol{\theta}\|^2/2\lambda. \tag{9}$$

In order to gain intuition about this mapping it is useful to analyse its behaviour when $\lambda \in \mathbb{R}^+$ is either very small or very large. In the limit $\lambda \to \infty$, the quadratic penalty term vanishes and (9) maps all points to $\hat{\boldsymbol{\theta}}_{MAP}$. In the opposite limit $\lambda \to 0$, (9) becomes the identity operator and maps $\boldsymbol{\theta}$ to itself. For finite values of λ, $\text{prox}_g^\lambda(\boldsymbol{\theta})$ behaves similarly to a gradient mapping and moves points in the direction of $\hat{\boldsymbol{\theta}}_{MAP}$. Like gradients, proximity mappings have several properties that are useful for devising fixed-point methods (Bauschke and Combettes 2011).

Property 1: The proximity mapping of g is related to its subdifferential by the inclusion $\{\boldsymbol{\theta} - \text{prox}_g^\lambda(\boldsymbol{\theta})\}/\lambda \in \partial g\{\text{prox}_g^\lambda(\boldsymbol{\theta})\}$, which collapses to $\nabla g\{\text{prox}_g^\lambda(\boldsymbol{\theta})\}$ when $g \in \mathcal{C}^1$. As a result, for any $\lambda > 0$, the minimiser of g verifies the fixed point equation $\boldsymbol{\theta} = \text{prox}_g^\lambda(\boldsymbol{\theta})$.

Property 2: Proximity mappings are firmly non-expansive; that is, $\|\text{prox}_g^\lambda(\boldsymbol{\theta}) - \text{prox}_g^\lambda(\boldsymbol{\phi})\|^2 \leq (\boldsymbol{\theta}-\boldsymbol{\phi})^T\{\text{prox}_g^\lambda(\boldsymbol{\theta}) - \text{prox}_g^\lambda(\boldsymbol{\phi})\}, \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^n$.

Property 3: The proximity mappings of g and its conjugate g^* are related by Moreau's decomposition formula: $\boldsymbol{\theta} = \text{prox}_g^\lambda(\boldsymbol{\theta}) + \lambda\,\text{prox}_{g^*}^{1/\lambda}(\boldsymbol{\theta}/\lambda)$.

The simplest proximal method to solve (8) is the *proximal point algorithm* given by the iteration

$$\boldsymbol{\theta}^{k+1} = \text{prox}_g^\lambda(\boldsymbol{\theta}^k). \tag{10}$$

Every sequence $\{\boldsymbol{\theta}^k\}_{k\in\mathbb{N}}$ produced by this algorithm converges to $\hat{\boldsymbol{\theta}}_{MAP}$, even if proximity mappings are evaluated inexactly, as long as the errors are of certain types (e.g., summable). A more general proximal point algorithm includes relaxation, i.e.,

$$\boldsymbol{\theta}^{k+1} = (1-\alpha_k)\boldsymbol{\theta}^k + \alpha_k\,\text{prox}_g^\lambda(\boldsymbol{\theta}^k), \quad \alpha_k \in (0,2),$$

and with over-relaxation (i.e., $\alpha_k \in (1,2)$) often converges faster than (10). Notice from Property 1 that (10) can be interpreted as an implicit (backward) subgradient steepest descent to minimise g, i.e., $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \lambda\boldsymbol{\varphi}$, with $\boldsymbol{\varphi} \in \partial g\left(\boldsymbol{\theta}^{k+1}\right)$. Alternatively, proximal point algorithms can also be interpreted as explicit (forward) gradient steepest descent to minimise the *Moreau envelope* of g, $e_\lambda(\boldsymbol{\theta}) = \inf_{\mathbf{u}\in\mathbb{R}^n}\ g(\mathbf{u}) + \|\mathbf{u}-\boldsymbol{\theta}\|^2/2\lambda$, a convex lower bound on g that by construction is continuously differentiable and has the same minimiser as g.

Proximal point algorithms may appear little relevant because evaluating prox_g^λ can be as difficult as solving (8) in the first place (notice that (9) is a convex minimisation problem similar to (8)). Surprisingly, many advanced proximal optimisation methods can in fact be shown to be either applications of this simple algorithm, or closely related to it.

Most proximal methods operate by splitting g, e.g.,

$$\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{argmin}}\{g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta})\}, \tag{11}$$

such that $g_1 \in \Gamma_0(\mathbb{R}^n)$ and $g_2 \in \Gamma_0(\mathbb{R}^n)$ have gradients or proximity mappings that are easy to compute or approximate. For example, for many Bayesian models it is possible to find a decomposition $g(\boldsymbol{\theta}) = g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta})$ such that g_1 is β-Lipschitz[2] differentiable and $g_2 \in \Gamma_0(\mathbb{R}^n)$, possibly non-differentiable, has a proximity mapping that can be computed efficiently with a specialised algorithm. This decomposition is useful for instance in linear inverse problems, where g_1 is often related to a Gaussian observation model involving linear operators and g_2 to a log-prior promoting a parsimonious representation (e.g., sparsity on some appropriate dictionary, low-rankness) or enforcing convex constraints (e.g., positivity, positive definiteness). For models that admit this decomposition, it is possible to compute $\hat{\boldsymbol{\theta}}_{MAP}$ efficiently with a *forward-backward* algorithm, also known as the proximal gradient algorithm

$$\boldsymbol{\theta}^{k+1} = \text{prox}_{g_2}^{\lambda_n}(\boldsymbol{\theta}^k - \lambda_n\nabla g_1(\boldsymbol{\theta}^k)). \tag{12}$$

For $\lambda_n = \lambda \in (0, 1/\beta)$ the objective function $g(\boldsymbol{\theta}^k)$ converges to $g(\hat{\boldsymbol{\theta}}_{MAP})$ with rate $O(1/k)$. If the value of the Lipschitz constant β is unknown λ_n can be found by line-search.

A remarkable property of (12) is that it can be accelerated to converge with rate $O(1/k^2)$, which is optimal for this class of problems (Nesterov 2004). This can be achieved for instance by introducing an extrapolation step

$$\begin{aligned} \boldsymbol{\theta}^+ &= \boldsymbol{\theta}^k + \omega_k(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}), \\ \boldsymbol{\theta}^{k+1} &= \text{prox}_{g_2}^{\beta^{-1}}(\boldsymbol{\theta}^+ - \beta^{-1}\nabla g_1(\boldsymbol{\theta}^+)), \end{aligned} \tag{13}$$

[2] $g_1 \in \mathcal{C}^1$ has β-Lipschitz continuous gradient if $\|\nabla g_1(\boldsymbol{\theta}) - \nabla g_1(\mathbf{u})\| \leq \beta\|\boldsymbol{\theta}-\mathbf{u}\|, \quad \forall(\boldsymbol{\theta},\mathbf{u}) \in \mathbb{R}^N \times \mathbb{R}^N$

where $\{\omega_k\}_{k\in\mathbb{N}}$ is an appropriate sequence of extrapolation parameters. It was noticed by Combettes and Pesquet (2011) that several important convex optimisation algorithms can be derived as applications of the forward-backward algorithm, for example the projected gradient algorithm for minimising a Lipschitz differentiable function subject to a convex constraint (in this case the proximity mapping reduces to a projection onto the convex set). Notice that (12) can be interpreted as an implementation of the proximal point iteration (10) where $\text{prox}_g^\lambda(\boldsymbol{\theta}^k)$ is approximated by replacing g_1 with its first order Taylor series approximation around the point $\boldsymbol{\theta}^k$.

Moreover, in some cases it may be more efficient to compute $\hat{\boldsymbol{\theta}}_{MAP}$ by solving the dual of (11), for instance if g admits a decomposition $g(\boldsymbol{\theta}) = g_1(\boldsymbol{\theta}) + g_2(L\boldsymbol{\theta})$ for some linear operator $L \in \mathbb{R}^{n\times p}$, $g_1 \in \Gamma_0(\mathbb{R}^n)$ strongly convex and $g_2 \in \Gamma_0(\mathbb{R}^p)$ with efficient proximity mapping. In this case, the Fenchel–Rockafellar theorem states that $\hat{\boldsymbol{\theta}}_{MAP}$ can be computed by solving the dual problem (Bauschke and Combettes 2011, ch. 19)

$$\boldsymbol{\psi}^* = \underset{\boldsymbol{\psi}\in\mathbb{R}^p}{\text{argmin}}\, g_1^*(-L^T\boldsymbol{\psi}) + g_2^*(\boldsymbol{\psi}) \tag{14}$$

and setting $\hat{\boldsymbol{\theta}}_{MAP} = \nabla g_1^*(-L^T\boldsymbol{\psi}^*)$. This p-dimensional problem can be solved iteratively with a forward-backward algorithm $\boldsymbol{\psi}^{k+1} = \text{prox}_{g_2^*}^{\lambda_n}(\boldsymbol{\psi}^k - \lambda_n\nabla g_1^*(-L^T\boldsymbol{\psi}^k))$ that can also be accelerated to converge with rate $O(1/k^2)$, and where we note that the proximity mapping of g_2^* is typically evaluated by using Property 3, and that the strong convexity of g_1 implies Lipschitz differentiability of g_1^*. Computing $\hat{\boldsymbol{\theta}}_{MAP}$ via (14) can lead to important computational savings, in particular if $p \ll n$ or if g_2 is separable and has a proximity mapping that can be computed in parallel for each element of $\boldsymbol{\theta}$ (this is generally not possible for $g_2 \circ L$). We refer the reader to (Komodakis and Pesquet 2014) for an overview of recent dual and primal-dual algorithms and guidelines for parallel implementations.

Another important proximal optimisation method is the Douglas–Rachford splitting algorithm given by

$$\begin{aligned}\boldsymbol{\theta}^{k+\frac{1}{2}} &= \text{prox}_{g_1}^{\lambda}(\boldsymbol{\theta}^k),\\ \boldsymbol{\theta}^{k+1} &= \boldsymbol{\theta}^k - \boldsymbol{\theta}^{k+\frac{1}{2}} + \text{prox}_{g_2}^{\lambda}(2\boldsymbol{\theta}^{k+\frac{1}{2}} - \boldsymbol{\theta}^k).\end{aligned} \tag{15}$$

From a theoretical viewpoint this algorithm is more general than the forward-backward algorithm because it does not require g_1 or g_2 to be continuously differentiable. However, its practical application is limited to problems for which both g_1 and g_2 have efficient proximity mappings. Similarly to the forward-backward algorithm, (15) includes many proximal algorithms that been proposed in the literature for specific models, and can also be interpreted as an application of the proximal point algorithm.

The proximal method that is arguably most widely used in Bayesian inference is the *alternating direction method of multipliers* (ADMM), which operates by formulating (11) as a constrained optimisation problem

$$\begin{aligned}&\underset{\boldsymbol{\theta}\in\mathbb{R}^n,\, \mathbf{z}\in\mathbb{R}^n}{\text{argmin}}\; g_1(\boldsymbol{\theta}) + g_2(\mathbf{z})\\ &\quad\text{subject to}\quad \boldsymbol{\theta} = \mathbf{z},\end{aligned} \tag{16}$$

and then using augmented Langrangian techniques to express (16) as an unconstrained saddle point problem with saddle function $g_1(\boldsymbol{\theta}) + g_2(\mathbf{z}) + \lambda\boldsymbol{\varphi}^T(\boldsymbol{\theta} - \mathbf{z}) + ||\boldsymbol{\theta} - \mathbf{z}||^2/2\lambda$ (Boyd et al. 2011). ADMM solves this problem with the iteration

$$\begin{aligned}\boldsymbol{\theta}^{k+1} &= \text{prox}_{g_1}^{\lambda}(\mathbf{z}^k - \boldsymbol{\varphi}^k),\\ \mathbf{z}^{k+1} &= \text{prox}_{g_2}^{\lambda}(\boldsymbol{\theta}^{k+1} + \boldsymbol{\varphi}^k),\\ \boldsymbol{\varphi}^{k+1} &= \boldsymbol{\varphi}^k + \boldsymbol{\theta}^{k+1} - \mathbf{z}^{k+1},\end{aligned} \tag{17}$$

that also involves the proximity mappings of g_1 and g_2. This basic ADMM iteration can be tailored to specific models in many ways (e.g., to exploit decompositions of the form $g_1 = \tilde{g}_1 \circ L_1$ and $g_2 = \tilde{g}_2 \circ L_2$ so that proximal updates can be performed in parallel for all components of $\boldsymbol{\theta}$, $\mathbf{z}$ and $\boldsymbol{\varphi}$). Interestingly, ADMM can be interpreted as an application of the Douglas–Rachford algorithm to the dual of (16), and is therefore also a special case of the proximal point algorithm. For more details about the ADMM algorithm, see the recent tutorial by Boyd et al. (2011).

Furthermore, an important characteristic of proximal optimisation algorithms is that they can be massively parallelised to take advantage of parallel computer architectures. Suppose for instance that g admits the decomposition $g(\boldsymbol{\theta}) = \sum_{m=1}^M g_m(L_m\boldsymbol{\theta})$ with $g_m \in \Gamma(\mathbb{R}^{p_m})$ and $L_m \in \mathbb{R}^{n\times p_m}$ such that the mappings of g_m are easy to compute and $Q = \sum_{m=1}^M L_m^T L_m$ is invertible. Then, in a manner akin to (16), we express (8) as

$$\begin{aligned}&\underset{\mathbf{z}_1\in\mathbb{R}^n,\ldots,\,\mathbf{z}_M\in\mathbb{R}^n}{\text{argmin}} \sum\nolimits_{m=1}^M g_m(\mathbf{z}_m)\\ &\text{subject to}\quad \mathbf{z}_m = L_m\boldsymbol{\theta},\ \forall m = 1,\ldots,M,\end{aligned} \tag{18}$$

and compute $\hat{\boldsymbol{\theta}}_{MAP}$ with the following iteration

$$\begin{aligned}\boldsymbol{\theta}^{k+1} &= Q^{-1}\sum\nolimits_{m=1}^M L_m^T(\mathbf{z}_m^k - \boldsymbol{\varphi}_m^k),\\ \mathbf{z}_m^{k+1} &= \text{prox}_{g_m}^{\lambda}(L_m\boldsymbol{\theta}^{k+1} - \boldsymbol{\varphi}_m^k),\ \forall m = 1,\ldots,M,\\ \boldsymbol{\varphi}_m^{k+1} &= \boldsymbol{\varphi}_m^k + L_m\boldsymbol{\theta}^{k+1} - \mathbf{z}_m^{k+1},\ \ \forall m = 1,\ldots,M,\end{aligned} \tag{19}$$

that can be parallelised with factor M at a coarse level (e.g., on a multi-processor system). Further parallelisation may be possible at a finer scale (e.g., on a vectorial

processor such as GPU or FPGA) by taking advantage of the structure of $\text{prox}_{g_m}^{\lambda}$ or by using specialised algorithms. This algorithm, known as the *simultaneous direction method of multipliers*, is also closely related to the ADMM, Douglas–Rachford and proximal point algorithms. Notice that splitting g not only allows the exploitation of parallel computer architectures, but may also significantly simplify the computation of proximity mappings; often $\text{prox}_{g_m}^{\lambda}$ has a closed-form expression. Lastly, it is worth mentioning that there are other modern proximal optimisation algorithms that can be massively parallelised, for example the *generalised forward backward* algorithm (Raguet et al. 2013), the *parallel proximal* algorithms (Combettes and Pesquet 2008; Pesquet and Pustelnik 2012), and the parallel primal-dual algorithm (Combettes and Pesquet 2012).

Finally, main current topics of research in proximal optimisation include theory and methodology for: 1) randomised and stochastic algorithms that operate with estimators of gradients and proximity mappings to reduce computational complexity and allow for errors in the update rules, 2) adaptive and variable metric algorithms (e.g. Riemannian and Newton-type) that exploit the model's geometry to improve convergence speed, and 3) proximal methods for non-convex problems. We anticipate that in the future new and stronger connections will emerge between proximal optimisation and stochastic simulation, in particular through developments in stochastic optimisation and high-dimensional MCMC sampling. For example, one connection is through the integration of modern stochastic convex optimisation and Markovian stochastic approximation (Combettes and Pesquet 2014; Andrieu et al. 2015), and of proximal optimisation and high-dimensional MCMC sampling (Pereyra 2013).

4.2 Convex relaxations

It is worth mentioning that modern proximal optimisation was greatly motivated by important theoretical results on the recovery of partially-observed sparse vectors and low-rank matrices through convex minimisation (Candès et al. 2006; Candès and Tao 2009) and on *compressive sensing* (Candès and Wakin 2008). A key idea underlying these works is that of approximating a combinatorial optimisation problem, whose solution is NP-hard, with a "relaxed" convex problem that is computationally tractable, and whose solution is in some sense close to the solution of the original problem. Reciprocally, the development of modern convex optimisation has in turn generated much interest in log-concave models, convex regularisers, and "convexifications" (i.e., convex relaxations for intractable or poorly tractable models) for statistical inference problems involving high-dimensionality, large datasets and computing time constraints (Chandrasekaran et al. 2012; Chandrasekaran and Jordan 2013).

4.3 Beyond MAP to approximating the posterior

We think it is vital to insist that, at the same time as asserting that modern optimisation methodology represents a much-underused opportunity in Bayesian inference, it nevertheless in its raw form inevitably fails to deliver essential elements of the Bayesian paradigm. The vision is not to deliver a point estimate of an unknown structure, but the full richness of Bayesian inference in its coherence, its proper treatment of uncertainty, its intrinsic treatment of model uncertainty, and so on. Bayesian statistics does not boil down to optimisation with penalisation (Lange et al. 2014). We need to express the uncertainty associated with decisions and estimation, stemming from the stochastic nature of the data, and our lack of knowledge about relevant mechanisms.

The challenge is to use the awesome capacity of fast optimisation in a high-dimensional parameter space to focus on local regions of that space where a combination of analytic and numerical investigation can deliver at least approximations to full posterior distributions and derived quantities. The community has barely risen to this challenge, with only isolated examples such as the discussion in (Green 2015) of a problem in unlabelled shape analysis.

4.4 Illustrative example

For illustration, we show an application of proximal optimisation to Bayesian image resolution enhancement. The goal is to recover a high-resolution image $\boldsymbol{\theta} \in \mathbb{R}^n$ from a blurred and noisy observed image $\mathbf{y} \sim \mathcal{N}(H\theta, \sigma^2 \boldsymbol{I}_n)$, where $H \in \mathbb{R}^{n \times n}$ is a linear operator representing the blur point spread function of the low resolution acquisition system and σ^2 is the system's noise power. This inverse problem is ill-posed, a difficulty that Bayesian image processing methods address by exploiting prior knowledge about $\boldsymbol{\theta}$. Here we use the following hierarchical Bayesian model (Oliveira et al. 2009)

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\theta}) &= (2\pi\sigma^2)^{-n/2} \exp\{-\|\mathbf{y} - H\boldsymbol{\theta}\|_2^2 / 2\sigma^2\}, \\ \pi(\boldsymbol{\theta}|\alpha) &\propto \alpha^{-n} \exp\left(-\alpha \|\nabla_d \boldsymbol{\theta}\|_{1-2}\right), \\ \pi(\alpha) &= e^{-\alpha} \mathbf{1}_{\mathbb{R}^+}(\alpha), \end{aligned} \tag{20}$$

where $\pi(\boldsymbol{\theta}|\alpha)$ is the (improper) *total-variation* Markov random field, $\|\cdot\|_{1-2}$ denotes the composite $\ell_1 - \ell_2$ norm

and ∇_d is the discrete gradient operator that computes the vertical and horizontal differences between neighbour image pixels. This prior is log-concave and models the fact that differences between neighbouring image pixels are usually very small but occasionally take large values; it is arguably the most widely used prior in modern statistical image processing. The values of H and σ^2 are typically determined during the system's calibration process and are here assumed known.

We compute the MAP estimator of $\boldsymbol{\theta}$ associated with the marginal posterior $\pi(\boldsymbol{\theta}|\mathbf{y}) = \int_0^\infty \pi(\boldsymbol{\theta}, \alpha|\mathbf{y})\mathrm{d}\alpha$, which is unimodal but not log-concave,

$$\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\operatorname{argmin}} \quad \|\mathbf{y} - H\boldsymbol{\theta}\|_2^2/2\sigma^2 + (n+1)\log\left(\|\nabla_d\boldsymbol{\theta}\|_{1-2} + 1\right). \tag{21}$$

Problem (21) is not convex, but can nevertheless be solved efficiently with proximal algorithms by using a *majorisation–minimisation* strategy. To be precise, starting from some initial condition $\boldsymbol{\theta}^{(0)}$, e.g., $\boldsymbol{\theta}^{(0)} = \mathbf{y}$, we iteratively minimise the following sequence of strictly convex majorants (Oliveira et al. 2009)

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\operatorname{argmin}} \|\mathbf{y} - H\boldsymbol{\theta}\|_2^2/2\sigma^2 + \alpha_{\text{eff}}^{(t)}\|\nabla_d\boldsymbol{\theta}\|_{1-2}, \quad \text{with} \quad \alpha_{\text{eff}}^{(t)} = (n+1)(\|\nabla_d\boldsymbol{\theta}^{(t)}\|_{1-2} + 1). \tag{22}$$

Iteration (22) involves a convex subproblem that can easily be solved using most modern proximal optimisation techniques. For example, here we use the state-of-the-art ADMM algorithm *SALSA* (Afonso et al. 2011) implemented with $g_1(\boldsymbol{\theta}) = \|\mathbf{y} - H\boldsymbol{\theta}\|_2^2/2\sigma^2$, $g_2(\mathbf{u}) = \alpha_{eff}^{(t)}\|\nabla_d\mathbf{u}\|_{1-2}$, and the constraint $\boldsymbol{\theta} = \mathbf{u}$ (though we could have also used other modern algorithms (Pesquet and Pustelnik 2012; Combettes and Pesquet 2012; Raguet et al. 2013)). To compute the proximity mapping of g_1 we use the fact that H is block-circulant to compute matrix products and pseudo-inverses with the FFT algorithm. We compute the proximity mapping of g_2 with a highly parallelised implementation of the specialised algorithm of Chambolle (2004).

Fig. 3 Observed blurred noisy image $\mathbf{y}$.

Fig. 4 Resolution enhanced image $\hat{\boldsymbol{\theta}}_{MAP}$ obtained by solving (21) with the majorisation-minimisation strategy (22).

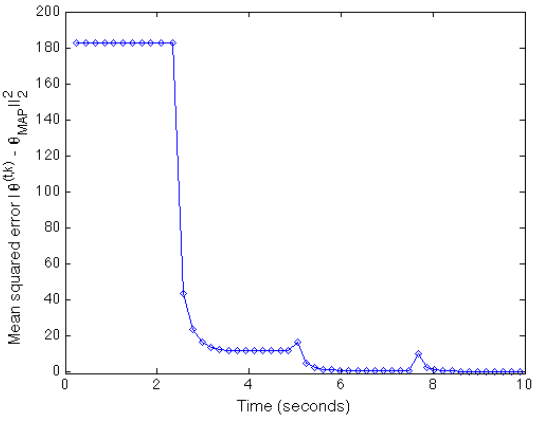

Fig. 5 Convergence of the estimate $\boldsymbol{\theta}$ to $\hat{\boldsymbol{\theta}}_{MAP}$ vs computing time (seconds)

Figure 3 presents a blurred and noisy observation $\mathbf{y}$ of the popular "boats" image of size 512×512 pixels, generated with a uniform 9×9 blur and a noise power

of $\sigma^2 = 0.5^2$ (blurred-signal-to-noise ratio $BRSN = 10\log_{10}\{\|H\boldsymbol{\theta}_0\|_2^2/\sigma^2\} = 40$dB). Figure 4 below shows MAP estimate $\hat{\boldsymbol{\theta}}_{MAP}$ obtained by solving (21) using 4 iterations of (22) and a total of 51 ADMM iterations. We observe that this resolution enhancement process has produced a remarkably sharp image with very noticeable fine detail. Finally, Figure 5 shows the convergence of the estimates $\boldsymbol{\theta}^{(t,k)}$ produced by each ADMM iteration to $\hat{\boldsymbol{\theta}}_{MAP}$ (as measured by the mean squared error $\|\boldsymbol{\theta}^{(t,k)} - \hat{\boldsymbol{\theta}}_{MAP}\|_2^2$) . Notice that computing this estimate only required 10 seconds (experiment conducted on an Apple Macbook Pro computer running MATLAB 2013, a C++ implementation would certainly produce even faster results). This is remarkably fast given the high dimensionality of the problem ($n = 262\,144$); estimating the posterior mean with a state-of-the-art MCMC algorithm would require at least 10 hours.

5 Discussion

5.1 Bayesian computation in the era of data science

Is there a revolution taking place right now and have we missed the train, standing on the platform, only concerned with small-print on the train schedules – apart, that is, from the obvious but not-so-new requirement to handle massive datasets (and the mistakes that come with them)?!

As with other areas of statistical science, the Bayesian computation community has to decide whether data science is an opportunity or a threat. Inevitably if we do not treat it as an opportunity, it will become a threat. Thanks to the ubiquity of "big data" (as an over-hyped phrase mostly useful for attracting research funding, but also to at least some extent in reality), a new potentially multi-disciplinary field of data science is rapidly opening up. This field is attracting huge material resources, and will absorb much human talent. Statistical science has to be a part of this, for its own survival, but also for the sake of society. As Tim Harford has cogently argued (Harford 2014):

> Recall big data's four articles of faith. Uncanny accuracy is easy to overrate if we simply ignore false positives [...]. The claim that causation has been "knocked off its pedestal" is fine if we are making predictions in a stable environment but not if the world is changing [...] or if we ourselves hope to change it. The promise that "N = All", and therefore that sampling bias does not matter, is simply not true in most cases that count. As for the idea that "with enough data, the numbers speak for themselves" – that seems hopelessly naïve in data sets where spurious patterns vastly outnumber genuine discoveries.
>
> "Big data" has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers – without making the same old statistical mistakes on a grander scale than ever.

It is a mistake to think that Bayes has no part to play in these developments, but more of us need to get more involved, and learn new tools, as in the way the Consensus Monte Carlo algorithm (Scott et al. 2013) exploits the Hadoop environment (White 2012) and the MapReduce programming model (Dean and Ghemawat 2008). Another direction that can prevent a potential schism between Bayesian modelling and highly complex models is to aim for modularity and local learning, that it, to abandon the goal of modelling big universes for analysing a series of small worlds, in spite of the loss of coherence, amd hence compromise to the Bayesian paradigm, that this entails. The curious case of the cut models presented in Plummer (2014) is an illustration of the potential for developing partial-information Bayesian inference tools where "small is beautiful" because this is the only viable solution.

5.2 Do we care enough about applications?

Bayesian computation began in order to answer rather practical problems – how can we perform a Bayesian analysis of these data using this model? – or the corresponding meta-problems – how can Bayesian analysis be performed generally and reliably for this class of models? The focus was applied methodology (although since the methods were new, they tended to be published in premier theory/methodology journals). Because the research community wanted to understand (the advantages, performance and limitations of) the methods they were advocating, more theoretical work started to be conducted, and, for example, many probabilists were attracted to study the Markov chains that MCMC methodologists created. The centre of mass of research activity drifted away from the original motivations, just as has happened in other areas of mathematically-rigorous computation.

At the same time, those working with data became more ambitious with regard to the scale of data, the complexity of modelling and the sophistication of analysis, all factors that have in principle (and often in fact) stimulated new developments in Bayesian computation. But to a large extent this is a rich, self-stimulating and self-supporting area of research; new applications may or may not need new computational techniques, but

new techniques don't seem to need applications to justify themselves. It is apposite to ask to what extent is cutting-edge computational methodology research really delivering answers to questions that application domains are posing. And to what extent is cutting-edge computational methodology research successfully answering real questions?

We may not be unanimous about answers to these questions, except we can probably all agree they are "not entirely". We will also disagree about how much this matters, but again there may be something to agree about, that we have failed if methodological innovations disconnect completely from applications. Legitimate differences in research goals partially explain the trend in this direction, but it is fair to say that there is a big communication problem between the computational statistics community and many of the communities where Bayesian computational methods are applied. Unfortunately people in these communities do not always keep up with the state of the art in computational statistics. At the same time, statisticians are often not aware of important developments arising in other fields. (ABC is a good illustration: it took more than five years of development within the population genetics community before statisticians became aware the technique existed and a few more years before they realised this was proper Bayesian inference applied on approximate models.) We can perhaps blame the fact that there are not enough people working at the interface of the different communities, but life at the interface is not easy because multidisciplinary and interdisciplinary research is often seen as "marginal" by both communities and is thus difficult to publish, communicate, etc. Then there are of course problems in dissemination, related to the different writing styles, journals, computing languages, software, etc. of each community.

We strongly encourage those developing new techniques always to find a way to disseminate them in such a way that at least *somebody* else could use them, preferably someone without the ability to have invented the technique for themselves! – and advocate, of course, that successful dissemination be properly rewarded in our career structures.

In a somewhat parallel path, we have seen over the past decades the emergence of new languages and metalanguages intended to handle complexity both of problems and of solutions towards a wider audience of users. BUGS (Lunn et al. 2010) is the archetypal example of such languages and it has been successful to the extent that a large proportion of the users has a fairly limited statistical background and often even less of a computational background. However, the population of BUGS users and sympathisers is tiny compared to that of SAS or other corporate statistical systems. In this respect, we have failed to disseminate concepts like Bayesian analysis and wonderful tools like MCMC algorithms, because most people are unable to turn them into codes by themselves. (Perusing one of the numerous statistics and machine-learning on-line forums like Cross Validated quickly exposes the methodological gap between academics and the masses!) It is unclear how novel programming developments like STAN (Stan Development Team 2014) are going to modify this picture, in that they still assume a decent understanding of both modelling and simulation issues. In that respect, network-based approaches as those covered by BUGS sound more promising towards "modelling locally to learn globally". Similarly, ABC software is either too specific, like DIYABC (Cornuet et al. 2008) which addresses only population genetic questions, or too dependent on the ability of the modeller to program simulated outcomes from the model under study.

5.3 Anticipating the future

In which of the areas we discuss do we expect a particular emphasis of effort, or significant progress, or do we see particular needs for new efforts or new directions?

One expectation is that in the future computational methodologies will be more flexible and malleable. Over the past 25 years Bayesian modelling and inference techniques have been applied successfully to thousands of problems across a wide range application domains. Each application brings its own constraints in terms of model dimensionality and complexity, data, inferences, accuracy and computing times. These constraints also vary significantly within specific applications. For example, in hyperspectral remote sensing, when a new Bayesian model is introduced it is often first explored and validated by MCMC sampling, then approximated with a variational Bayes method, and then approximated again so that it can be applied to gigabyte-large datasets by using optimisation techniques. Similarly, an interesting result revealed by a fast inference technique can be analysed more deeply with more reliable and accurate methods. Therefore we expect that in the future the different main computational methodologies will become more adaptable and that the boundaries between them will be less well defined, with many algorithms developed that combine simulation, variational approximations and optimisation. These will be able to handle a wide spectrum of models, degrees of accuracy and computing times, as well as models that have some parts that are simple but high-dimensional and others that are more complex but that only involve low-dimensional

components. This can be achieved by using approximations and optimisation to improve stochastic sampling, by using simulation within deterministic algorithms to handle specific parts of the model that are difficult to compute analytically, or in completely new and original ways.

We also anticipate that computational methodologies will continue to be challenged by larger and larger datasets. There is of course a threat that the whole field turns into a variety of machine-learning techniques, with limited validation on reference learning sets and a quick turnover of methods, which would both impoverish the field and fail to reach a general audience of practitioners. We must retain a sense of the stochastic elements in data collection, data analysis, and inference, recognising uncertainty in data and models, to preserve the inductive strength of data science – seeing beyond the data we have to what it might have been, what it might be next time, and where it came from.

References

Afonso, M., Bioucas-Dias, J. and Figueiredo, M. (2011). An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE. Trans. on Image Process.*, **20** 681–695.

Albert, J. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *J. American Statist. Assoc.*, **83** 1037–1044.

Aldous, D., Krikun, M. and Popovic, L. (2008). Stochastic models for phylogenetic trees on higher-order taxa. *J. Math. Biology*, **56** 525–557.

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

Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, **16** 1462–1505.

Andrieu, C. and Robert, C. (2001). *Controlled MCMC for optimal sampling.*

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

Andrieu, C., Tadić, V. B. and Vihola, M. (2015). On the stability of some controlled Markov chains and its applications to stochastic approximation with Markovian dynamic. *The Annals of Applied Probability*, **25** 1–45.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, **18** 343–373.

Andrieu, C. and Vihola, M. (2012). Convergence properties of pseudo-marginal markov chain monte carlo algorithms. *arXiv preprint arXiv:1210.1484; Annals of Applied Probability, to appear.*

Angelino, E., Kohler, E., Waterland, A., Seltzer, M. and Adams, R. (2014). Accelerating MCMC via parallel predictive prefetching. *arXiv preprint arXiv:1403.7265.*

Atchadé, Y. (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in Applied Probability*, **8** 235–254.

Atchade, Y., Fort, G., Moulines, E. and Priouret, P. (2011). *Adaptive Markov chain Monte Carlo: theory and methods*, vol. Bayesian Time Series Models. Cambridge University Press.

Atchadé, Y. and Rosenthal, J. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, **11** 815–828.

Atchadé, Y. F. and Liu, J. S. (2004). The Wang-Landau algorithm for Monte Carlo computation in general state spaces. *Technical Report.*

Atchadé, Y. F., Roberts, G. O. and Rosenthal, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Stat. Comput.*, **21** 555–568.

Bai, Y., Roberts, G. and Rosenthal, J. (2011). On the containment condition for adaptive Markov chain Monte Carlo algorithms. *Advances and Applications in Statistics*, **21** 1–54.

Barber, S., Voss, J. and Webster, M. (2013). The Rate of Convergence for Approximate Bayesian Computation. *ArXiv e-prints.* 1311.2038.

Bardenet, R., Doucet, A. and Holmes, C. (2014). Towards scaling up markov chain monte carlo: an adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (T. Jebara and E. P. Xing, eds.). JMLR Workshop and Conference Proceedings, 405–413.

Barthelmé, S. and Chopin, N. (2014). Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, **109** 315–333.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer New York.

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., Nielsen, R., Robert, C., Hey, J., Gaggiotti, O., Knowles, L., Estoup, A., Mahesh, P., Coranders, J., Hickerson, M., Sisson, S., Fagundes, N., Chikhi, L., Beerli, P., Vitalis, R., Cornuet, J.-M., Huelsenbeck, J., Foll, M., Yang, Z., Rousset, F., Balding, D. and Excoffier, L. (2010). In defense of model-based inference in phylogeography. *Molecular Ecology*, **19(3)** 436–446.

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

Bédard, M. (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.*, **17** 1222–1244.

Bédard, M., Douc, R. and Moulines, E. (2012). Scaling analysis of multiple-try MCMC methods. *Stochastic Process. Appl.*, **122** 758–786.

Bédard, M., Douc, R. and Moulines, E. (2014). Scaling analysis of delayed rejection MCMC methods. *Methodol. Comput. Appl. Probab.*, **16** 811–838.

Belle, E., Benazzo, A., Ghirotto, S., Colonna, V. and Barbujani, G. (2008). Comparing models on the genealogical relationships among Neandertal, Cro-Magnoid and modern Europeans by serial coalescent simulations. *Heredity*, **102** 218–225.

Bennett, J., Racine-Poon, A. and Wakefield, J. (1996). MCMC for nonlinear hierarchical models. In *Markov chain Monte Carlo in Practice* (W. Gilks, S. Richard-

son and D. Spiegelhalter, eds.). Chapman and Hall, New York, 339–358.

Berger, J., Fienberg, S., Raftery, A. and Robert, C. (2010). Incoherent phylogeographic inference. *Proc. National Academy Sciences*, **107** E57.

Besag, J. and Green, P. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Royal Statist. Society Series B*, **55** 25–38.

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10** 3–66.

Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *J. Roy. Statist. Soc. Ser. B*, **34** 75–83.

Beskos, A., Papaspiliopoulos, O., Roberts, G. and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. Royal Statist. Society Series B*, **68** 333–382.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M. and Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, **19** 1501–1534.

Beskos, A., Roberts, G. and Stuart, A. (2009). Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Probab.*, **19** 863–898.

Betancourt, M. J., Byrne, S., Livingstone, S. and Girolami, M. (2014). The geometric foundations of Hamiltonian Monte Carlo. *ArXiv e-prints*. 1410.5110.

Biau, G., Cérou, F. and Guyader, A. (2014). New insights into Approximate Bayesian computation. *Annales de l'IHP (Probability and Statistics)*, **51** 376–403.

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

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

Blum, M., Nunes, M., Prangle, D. and Sisson, S. (2013). A comparative review of dimension reduction methods in Approximate Bayesian computation. *Stat Sci*, **28** 189–208.

Bornn, L., Pillai, N., Smith, A. and Woodard, D. (2014). A pseudo-marginal perspective on the ABC algorithm. *ArXiv e-prints*. 1404.6298.

Bou-Rabee, N. and Hairer, M. (2012). Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis* drs003.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3** 1–122.

Brockwell, A. (2006). Parallel Markov chain Monte Carlo simulation by pre-fetching. *J. Comput. Graphical Stat.*, **15** 246–261.

Calderhead, B. (2014). A general construction for parallelizing Metropolis-Hastings algorithms. *Proceedings of the National Academy of Sciences*, **111** 17408–17413.

Calvet, C. and Czellar, V. (2014). Accurate methods for Approximate Bayesian computation filtering. *J. Financial Econometrics*. (to appear).

Candès, E. J., Romberg, J. K. and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, **59** 1207–1223.

Candès, E. J. and Tao, T. (2009). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, **56** 2053–2080.

Candès, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Process. Mag.*, **25** 21–30.

Cappé, O., Robert, C. and Rydén, T. (2002). Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *J. Royal Statist. Society Series B*, **65** 679–700.

Carlin, B. and Gelfand, A. (1991). An iterative Monte Carlo method for nonconjugate Bayesian analysis. *Statistics and Computing*, **1** 119–28.

Carlin, B., Gelfand, A. and Smith, A. (1992). Hierarchical Bayesian analysis of change point problems. *Applied Statistics (Series C)*, **41** 389–405.

Cevher, V., Becker, S. and Schmidt, M. (2014). Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *Signal Processing Magazine, IEEE*, **31** 32–43.

Chambolle, A. (2004). An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, **20** 89–97.

Chandrasekaran, V. and Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *PNAS*, **110** 1181–1190.

Chandrasekaran, V., Recht, B., Parrilo, P. and Willsky, A. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.*, **12** 805–849.

Chipman, H., George, E. and McCulloch, R. (2008). BARTf: Bayesian additive regression trees. Tech. rep., Acadia University. ArXiv:0806.3286v1.

Chopin, N. (2007). Inference and model choice for time-ordered hidden Markov models. *J. Royal Statist. Society Series B*, **69(2)** 269–284.

Chopin, N., Jacob, P. E. and Papaspiliopoulos, O. (2013). SMC2: an efficient algorithm for sequential analysis of state space models. *J. Royal Statist. Society Series B*, **75** 397–426.

Christensen, O., Roberts, G. and Rosenthal, J. (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67** 253–268.

Combettes, P. L. and Pesquet, J.-C. (2008). A proximal decomposition method for solving convex variational inverse problems. *Inverse Problems*, **24** 065014.

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* (H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke and H. Wolkowicz, eds.). Springer New York, 185–212.

Combettes, P. L. and Pesquet, J.-C. (2012). Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. *Set-Valued A.*, **20** 307–330.

Combettes, P. L. and Pesquet, J.-C. (2014). Stochastic Quasi-Fejér block-coordinate fixed point iterations with random sweeping. *ArXiv e-prints*. 1404.7536.

Cornuet, J.-M., Ravigné, V. and Estoup, A. (2010). Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, **11** 401.

Cornuet, J.-M., Santos, F., Beaumont, M., Robert, C., Marin, J.-M., Balding, D., Guillemaud, T. and Estoup, A. (2008). Inferring population history with

DIYABC: a user-friendly approach to Approximate Bayesian computation. *Bioinformatics*, **24** 2713–2719.

Cotter, S., Roberts, G., Stuart, A., White, D. et al. (2013). MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, **28** 424–446.

Craiu, R., Rosenthal, J. and Yang, C. (2009). Learn from thy neighbour: Parallel-chain and regional adaptive MCMC. *J. American Statist. Assoc.*, **104** 1454–1466.

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

Cucala, L., Marin, J.-M., Robert, C. and Titterington, D. (2009). Bayesian inference in k-nearest-neighbour classification models. *J. American Statist. Assoc.*, **104 (485)** 263–273.

Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, **51** 107–113.

Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *J. Royal Statist. Society Series B*, **68** 411–436.

Dellaportas, P., Papaspiliopoulos, O. and Roberts, G. (2004). Bayesian inference for non-Gaussian Ornstein-uhlenbeck stochastic volatility processes. *J. Royal Statist. Society Series B*, **66** 369–393.

Dellaportas, P. and Wright, D. (1991). Positive embedded integration in Bayesian analysis. *Statistics and Computing*, **1** 1–12.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *The Annals of Applied Probability* 36–61.

Didelot, X., Everitt, R., Johansen, A. and Lawson, D. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, **6** 48–76.

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

Doll, J. and Dion, D. (1976). Generalized langevin equation approach for atom/solid–surface scattering: Numerical techniques for gaussian generalized langevin dynamics. *The Journal of Chemical Physics*, **65** 3762–3766.

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

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

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B*, **195** 216–222.

Ermak, D. (1975). A computer simulation of charged particles in solution. I. Technique and equilibrium properties. *The Journal of Chemical Physics*, **62** 4189–4196.

Excoffier, C., Leuenberger, D. and Wegmann, L. (2009). Bayesian computation and model selection in population genetics. ArXiv:0901.2231.

Fearnhead, P. and Clifford, P. (2003). On-line inference for hidden Markov models via particle filters. *J. Royal Statist. Society Series B*, **65** 887–899.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for Approximate Bayesian Computation: semi-automatic Approximate Bayesian Computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74** 419–474. (With discussion.).

Fort, G., Moulines, E. and Priouret, P. (2011). Convergence of adaptive and interacting markov chain monte carlo algorithms. *The Annals of Statistics*, **39** 3262–3289.

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

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.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, **7** 473–483.

Ghirotto, S., Mona, S., Benazzo, A., Paparazzo, F., Caramelli, D. and Barbujani, G. (2010). Inferring genealogical processes from patterns of bronze-age and modern DNA variation in Sardinia. *Mol. Biol. Evol.*, **27** 875–886.

Gilks, W., Roberts, G. and George, E. (1994). Adaptive direction sampling. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **43** 179–189.

Gilks, W., Roberts, G. and Sahu, S. (1998). Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association*, **93** 1045–1054.

Girolami, M. and Calderhead, B. (2010). An object-oriented random-number package with many long streams and substreams. *J. Royal Statist. Society Series B*, **73** 1–37.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73** 123–214.

Girolami, M., Lyne, A.-M., Strathmann, H., Simpson, D. and Atchade, Y. (2013). Playing russian roulette with intractable likelihoods. *arXiv preprint arXiv:1306.4032*.

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

Green, P. J. (2015). MAD-Bayes matching and alignment for labelled and unlabelled configurations. In *Geometry driven statistics* (I. L. Dryden and J. T. Kent, eds.), chap. 19. Wiley, Chichester, 365–375.

Grelaud, A., Marin, J.-M., Robert, C., Rodolphe, F. and Tally, F. (2009). Likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, **3(2)** 427–442.

Griffin, J., Łatuszyński, K. and Steel, M. (2014). Individual adaptation: an adaptive MCMC scheme for variable selection problems. *submitted*.

Guillemaud, T., Beaumont, M., Ciosi, M., Cornuet, J.-M. and Estoup, A. (2009). Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, **104** 88–99.

Haario, H., Laine, M., Mira, A. and Saksman, E. (2006). DRAM: efficient adaptive MCMC. *Statistics and Computing*, **16** 339–354.

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7** 223–242.

Harford, T. (2014). Big data: Are we making a big mistake? *Significance*, **11** 14–19.

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

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, **10** 25–37.

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

Jacob, P. E. and Ryder, R. J. (2014). The Wang–Landau algorithm reaches the flat histogram criterion in finite time. *Ann. Appl. Probab.*, **24** 34–53.

Ji, C. and Schmidler, S. C. (2013). Adaptive Markov chain Monte Carlo for Bayesian variable selection. *Journal of Computational and Graphical Statistics*, **22** 708–728.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, **37** 183–233.

Jourdain, B., Lelièvre, T. and Miasojedow, B. (2012). Optimal scaling for the transient phase of the random walk Metropolis algorithm: the mean-field limit. *arXiv preprint arXiv:1210.7639; Annals of Applied Probability, to appear.*

Jourdain, B., Lelièvre, T. and Miasojedow, B. (2014). Optimal scaling for the transient phase of Metropolis Hastings algorithms: the longtime behavior. *Bernoulli*, **20** 1930–1978.

Kent, J. (1978). Time-reversible diffusions. *Adv. in Appl. Probab.*, **10, no. 4** 819–835.

Kloeden, P. E. and Platen, E. (1992). *Numerical solution of stochastic differential equations*, vol. 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin.

Komodakis, N. and Pesquet, J.-C. (2014). Playing with Duality: An Overview of Recent Primal-Dual Approaches for Solving Large-Scale Optimization Problems. *ArXiv e-prints*. 1406.5429.

Korattikara, A., Chen, Y. and Welling, M. (2013). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. *arXiv preprint arXiv:1304.5299*.

Kou, S. C., Zhou, Q. and Wong, W. H. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.*, **34** 1581–1652. With discussions and a rejoinder by the authors.

Lamnisos, D., Griffin, J. E. and Steel, M. F. (2013). Adaptive Monte Carlo for Bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics*, **22** 729–748.

Lange, K., Chi, E. C. and Zhou, H. (2014). A brief survey of modern optimization for statisticians. *International Statistical Review*, **82** 46–70.

Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, **16** 750–759.

Łatuszyński, K., Kosmidis, I., Papaspiliopoulos, O. and Roberts, G. O. (2011). Simulating events of unknown probabilities via reverse time martingales. *Random Structures & Algorithms*, **38** 441–452.

Łatuszyński, K., Roberts, G. and Rosenthal, J. (2013). Adaptive Gibbs samplers and related MCMC methods. *Ann. Appl. Probab.*, **23(1)** 66–98.

Łatuszyński, K. and Rosenthal, J. S. (2014). The containment condition and AdapFail algorithms. *Journal of Applied Probability*, **51** 1189–1195.

Lee, A. and Łatuszyński, K. (2014). Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for Approximate Bayesian computation. *Biometrika*, **101** 655–671.

Lee, A., Yau, C., Giles, M., Doucet, A. and Holmes, C. (2009). On the utility of graphics cards to perform massively parallel simulation with advanced Monte Carlo methods. *Arxiv preprint arXiv:0905.2441*.

Lee, H. K., Okabe, Y. and Landau, D. P. (2005). Convergence and refinement of the Wang-Landau algorithm. *Technical Report*.

Leuenberger, C. and Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. *Genetics*, **184** 243–252.

Levin, D. A., Peres, Y. and Wilmer, E. L. (2009). *Markov chains and mixing times*. American Mathematical Soc.

Lindsten, F., Jordan, M. I. and Schön, T. B. (2014). Particle Gibbs with ancestor sampling. *ArXiv e-prints*. 1401.0604.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2010). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC Press.

MacKay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Cambridge, UK.

Maclaurin, D. and Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. *arXiv preprint arXiv:1403.5693*.

Marin, J., Pillai, N., Robert, C. and Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *J. Royal Statist. Society Series B.* (to appear).

Marin, J., Pudlo, P., Robert, C. and Ryder, R. (2011). Approximate Bayesian computational methods. *Statistics and Computing* 1–14.

Marshall, T. and Roberts, G. (2012). An adaptive approach to Langevin MCMC. *Statistics and Computing*, **22** 1041–1057.

Martinet, B. (1970). Regularisation d'inéquations variationelles par approximations successives. *Revue Fran. d'Automatique et Infomatique Rech. Opérationelle*, **4** 154–159.

Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, **24** 101–121.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science*, **15** 125–130.

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.

Meyn, S. and Tweedie, R. (2009). *Markov chains and stochastic stability*. Cambridge University Press.

Miasojedow, B., Moulines, E. and Vihola, M. (2013). An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, **22** 649–664.

Minka, T. (2001). Expectation propagation for approximate Bayesian inference. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence* (D. K. Jack S. Breese, ed.). University of Washington, Seattle, 362–369.

Minsker, S., Srivastava, S., Lin, L. and Dunson, D. B. (2014). Robust and scalable Bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.

Møller, J., Pettitt, A. N., Reeves, R. and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, **93** 451–458.

Moreau, J.-J. (1962). Fonctions convexes duales et points proximaux dans un espace Hilbertien. *C. R. Acad. Sci. Paris Sér. A Math.*, **255** 2897–2899.

Murray, I., Ghahramani, Z., and MacKay, D. (2006a). Mcmc for doubly-intractable distributions. In *Uncertainty in Artificial Intelligence*. UAI-2006.

Murray, I., MacKay, D. J., Ghahramani, Z. and Skilling, J. (2006b). Nested sampling for Potts models. In *Advances in Neural Information Processing Systems 18* (Y. Weiss, B. Schölkopf and J. Platt, eds.). MIT Press, Cambridge, MA, 947–954.

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

Neal, P., Roberts, G. and Yuen, W. K. (2012). Optimal scaling of random walk Metropolis algorithms with discontinuous target densities. *Ann. Appl. Probab.*, **22** 1880–1927.

Neal, R. (1999). *Bayesian Learning for Neural Networks*, vol. 118. Springer–Verlag, New York. Lecture Notes.

Neal, R. (2013). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds.). Chapman & Hall/CRC Press, 113–162.

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

Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course*, vol. 87 of *Applied optimization*. Kluwer Academic Publishers.

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

Oliveira, J., Bioucas-Dias, J. and Figueiredo, M. (2009). Adaptive total variation image deblurring: A majorization-minimization approach. *Signal Process.*, **89** 1683–1693.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall.

Paisley, J., Blei, D. M. and Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. 1367–1374.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, **22** 59–73.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, **1** 123–231.

Patin, E., Laval, G., Barreiro, L., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K., Kidd, J., Van Der Veen, L., Hombert, J. et al. (2009). Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genetics*, **5** e1000448.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Proc. Trans. Royal Society A*, **185** 71–110.

Pereyra, M. (2013). Proximal Markov chain Monte Carlo algorithms. *ArXiv e-prints*. 1306.0187.

Peskun, P. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika*, **60** 607–612.

Pesquet, J.-C. and Pustelnik, N. (2012). A parallel inertial proximal optimization method. *Pac. J. Optim.*, **8** 273–305.

Pillai, N. S., Stuart, A. M. and Thiéry, A. H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, **22** 2320–2356.

Plummer, M. (2014). Cuts in Bayesian graphical models. *Statistics and Computing* to appear.

Potts, R. B. (1952). Some generalized order-disorder transitions. *Proceedings of Cambridge Philosophical Society*, **48** 106–109.

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.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M. and Robert, C. P. (2014). ABC model choice via random forests. *ArXiv e-prints*. 1406.6288.

Raguet, H., Fadili, J. and Peyré, G. (2013). A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, **6** 1199–1226.

Ramakrishnan, U. and Hadly, E. (2009). Using phylochronology to reveal cryptic population histories: review and synthesis of 29 ancient DNA studies. *Molecular Ecology*, **18** 1310–1330.

Richardson, S., Bottolo, L. and Rosenthal, J. (2010). Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics*, **9**.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B*, **59** 731–792.

Robert, C. and Casella, G. (2010). A history of Markov chain Monte Carlo-subjective recollections from incomplete data. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, X. Meng and G. Jones, eds.). Chapman and Hall, New York, 49–66. ArXiv0808.2902.

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

Robert, C., Cornuet, J.-M., Marin, J.-M. and Pillai, N. (2011). Lack of confidence in ABC model choice. *Proceedings of the National Academy of Sciences*, **108(37)** 15112–15117.

Roberts, G., Gelman, A. and Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, **7** 110–120.

Roberts, G. and Rosenthal, J. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. Roy. Stat. Soc. B*, **60** 255–268.

Roberts, G. and Rosenthal, J. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, **16** 351–367.

Roberts, G. and Rosenthal, J. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, **1** 20–71.

Roberts, G. and Rosenthal, J. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, **44** 458.

Roberts, G. and Rosenthal, J. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18** 349–367.

Roberts, G. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, **4** 337–357.

Roberts, G. and Tweedie, R. (1996a). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, **2** 341–363.

Roberts, G. and Tweedie, R. (1996b). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83** 95–110.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. *Markov chain Monte Carlo in practice*, **57**.

Roberts, G. O. (1998). Optimal Metropolis algorithms for product measures on the vertices of a hypercube. *Stochastics and Stochastic Reports*, **62** 275–283.

Roberts, G. O. and Rosenthal, J. S. (2014). Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *Ann. Appl. Probab.*, **24** 131–149.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the metropolis–hastings algorithm. *Biometrika*, **88** 603–621.

Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, **14** 877–898.

Rossky, P., Doll, J. and Friedman, H. (1978). Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, **69** 4628–4633.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12** 1151–1172.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. J. Wiley, New York.

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

Saksman, E. and Vihola, M. (2010). On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. *The Annals of Applied Probability*, **20** 2178–2203.

Salimans, T. and Knowles, D. A. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Anal.*, **8** 837–882.

Saloff-Coste, L. (1997). Lectures on finite Markov chains. In *Lectures on probability theory and statistics (Saint-Flour, 1996)*, vol. 1665 of *Lecture Notes in Math.* Springer, Berlin, 301–413.

Schreck, A., Fort, G., Corff, S. L. and Moulines, E. (2013). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. *arXiv preprint arXiv:1312.5658*.

Scott, S., Blocker, A., Bonassi, F., Chipman, H., George, E. and McCulloch, R. (2013). Bayes and big data: The consensus Monte Carlo algorithm. *EFaBBayes 250 conference*, **16**.

Searle, S., Casella, G. and McCulloch, C. (1992). *Variance Components*. John Wiley, New York.

Sherlock, C., Thiery, A. H., Roberts, G. O. and Rosenthal, J. S. (2014). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics*, **43** 238–275.

Smith, A., Skene, A., Shaw, J., Naylor, J. and Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Comm. Statist.-Theory Methods*, **14** 1079–1102.

Smith, A., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987). Progress with numerical and graphical methods for practical Bayesian statistics. *J. Roy. Statist. Soc. Series D*, **36** 75–82.

Solonen, A., Ollinaho, P., Laine, M., Haario, H., Tamminen, J. and Järvinen, H. (2012). Efficient MCMC for climate model parameter estimation: Parallel adaptive chains and early rejection. *Bayesian Analysis*, **7** 715–736.

Stan Development Team (2014). STAN: A C++ library for probability and sampling, version 2.5.0. URL http://mc-stan.org/.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.*, **28** 40–74.

Stramer, O. and Tweedie, R. (1999a). Langevin-type models I: Diffusions with given stationary distributions and their discretizations. *Methodology and Computing in Applied Probability*, **1** 283–306.

Stramer, O. and Tweedie, R. (1999b). Langevin-type models II: Self-targeting candidates for MCMC algorithms. *Methodology and Computing in Applied Probability*, **1** 307–328.

Strid, I. (2010). Efficient parallelisation of Metropolis–Hastings algorithms using a prefetching approach. *Computational Statistics & Data Analysis*, **54** 2814–2835.

Suchard, M., Wang, Q., Chan, C., Frelinger, J., Cron, A. and West, M. (2010). Understanding GPU programming for statistical computation: studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, **19** 418–438.

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

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

Teh, Y. W., Thiéry, A. and Vollmer, S. (2014). Consistency and fluctuations for stochastic gradient Langevin dynamics. *arXiv preprint arXiv:1409.0578*.

Templeton, A. (2008). Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. (a)pproximate Bayesian computation. *Molecular Ecology*, **18(2)** 319–331.

Templeton, A. (2010). Coherent and incoherent inference in phylogeography and human evolution. *Proc. National Academy of Sciences*, **107(14)** 6376–6381.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, **8** 1–9.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6** 187–202.

VanDerwerken, D. N. and Schmidler, S. C. (2013). Parallel Markov chain Monte Carlo. *ArXiv e-prints*. 1312.7479.

Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. *Statist. Comput.*, **1** 105–117.

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Théry, S., Froment, A., Le Bomin, S., Gessain, A., Hombert, J.-M., Van der Veen, L., Quintana-Murci, L., Bahuchet, S. and Heyer, E. (2009). Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, **19** 312–318.

Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, **22** 997–1008.

Wakefield, J., Gelfand, A. and Smith, A. (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing*, **1** 129–133.

Wang, F. and Landau, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, **86** 2050–2053.

Wang, X. and Dunson, D. (2013). Parallizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.

Wang, Z., Mohamed, S. and Freitas Nando, D. (2013). Adaptive Hamiltonian and Riemann manifold Monte Carlo. In *Proceedings of The 30th International Con-*

ference on Machine Learning. 1462–1470.

Wegmann, D. and Excoffier, L. (2010). Bayesian inference of the demographic history of chimpanzees. *Molecular Biology and Evolution*, **27** 1425–1435.

White, T. (2012). *Hadoop: the definitive guide*. O'Reilly Media.

Whiteley, N., Andrieu, C. and Doucet, A. (2010). Efficient Bayesian inference for switching state-space models using discrete particle Markov chain Monte Carlo methods. *ArXiv e-prints*. 1011.2437.

Wilkinson, D. (2005). Parallel Bayesian computation. In *Handbook of Parallel Computing and Statistics* (E. J. Kontoghiorghes, ed.). Marcel Dekker/CRC Press, New York, 481–512.

Wilkinson, D. (2011a). The particle marginal Metropolis–Hastings (PMMH) particle MCMC algorithm. https://darrenjw.wordpress.com/2011/05/17/the-particle-marginal-metropolis-hastings-pmmh-particle-mcmc-algorithm/. Darren Wilkinson's research blog.

Wilkinson, D. J. (2011b). *Stochastic modelling for systems biology*. CRC press, New York. (Second edition).

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

Wolny, K. (2014). PhD thesis. *University of Warwick*.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S. and Girolami, M. (2014). Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, **91** 14–19.