Monte Carlo algorithms for model assessment via conflicting summaries

Oliver Ratmann · Pierre Pudlo · Sylvia Richardson · Christian Robert

Received: date / Accepted: date

Abstract The development of statistical methods and numerical algorithms for model choice is vital to many real-world applications. In practice, the ABC approach can be instrumental for sequential model design; however, the theoretical basis of its use has been questioned. We present a measure-theoretic framework for using the ABC error towards model choice and describe how easily existing rejection, Metropolis-Hastings and sequential importance sampling ABC algorithms are extended for the purpose of model checking. We considering a panel of applications from evolutionary biology to dynamic systems, and discuss the choice of summaries, which differs from standard ABC approaches. The methods and algorithms presented here may provide the workhorse machinery for an exploratory approach to ABC model choice, particularly as the application of standard Bayesian tools can prove impossible.

1 Introduction

Approximate Bayesian Computation (ABC) methods have appeared in the past ten years as a way to handle intractable likelihoods and posterior densities

\[ \pi(\theta|x_0) \propto f(x_0|\theta)\pi(\theta) \] (1)

that arise under high dimensional, data-generating models. For example, complex coalescent models are known to generate, currently, latent structures that are too high dimensional to bring a reliable numerical approximation in practical computer time. Originally developed for population genetics, ABC has since been applied to many applied problems where Bayesian analysis has long been contemplated but previously remained elusive (Beaumont, 2010; Marin et al, 2011).

As with other approximation methods like variational Bayes (Jaakkola and Jordan, 2000; MacKay, 2002) or indirect inference (Heggland and Frigessi, 2004), ABC suffers from a limited ability to quantify the uncertainty in the approximation of the posterior (1). Moreover, the loss of information brought by the ABC approximation implies that the application of parts of standard Bayesian machinery, such as the Bayes factor, is fraught with difficulties (Robert et al, 2011).

While much of Bayesian model checking is based on evaluating model predictions, ABC uses such model predictions for parameter inference. In this perspective, it is natural to attempt using the pseudo-data \( x \) that is generated by ABC Monte Carlo algorithms both for parameter inference and model assessment. There is no...
issue of bias in doing so because we are considering a simulation technique rather than an inferential method: the data itself is only "used once". Some of us have called for some technical refinement of this framework, called ABC under model uncertainty (ABCμ) (Robert et al 2009), particularly as this technique can generate further insight in practice (Drovandi et al 2011; Ratmann et al 2010). In opposition to more formal Bayesian model choice approaches (Toni et al 2008; Grelaud et al 2009), a key to the validation of ABC as a model assessment in Ratmann et al (2009) relies on the very fact that likelihood computations and comparisons under a model can be replaced by assessing the amount of fit between simulations x from that model and the observed data x₀.

In this paper, we first present technical modifications that reflect our consensus view on ABC under model uncertainty. We then describe basic, yet efficient Metropolis-Hastings and sequential importance sampling ABC algorithms for approximate parameter inference and model checking, which forms the main contribution of this paper. On purpose, these algorithms are closely related to existing, popular ABC algorithms to show how easily these methods can be extended to incorporate model checking at no or little additional computational cost. These algorithms are presented in Section 2 following a description of their theoretical foundations in Section 2. We illustrate these algorithms on a panel of applications, ranging from population genetics to network evolution and dynamic systems (Sections 4-6). We discuss the relative advantages of both algorithms, and compare these to a hybrid algorithm that seeks to combine the strengths of either method. Given the difficulties associated with using approximate Bayes factors for model choice, we conclude that the algorithms presented in this paper may provide the workhorse machinery for a viable, exploratory approach to model choice when the likelihood is computationally intractable (Section 7).

2 A measure-theoretic framework for ABC

To formalize the setting of ABC-led inference with an application to diagnostic model assessment in mind, we first present a measure-theoretic framework that updates the previous formulation in Ratmann et al (2009). The extended ABC algorithms in Section 3 also handle goodness-of-fit type analyses and follow immediately from this re-interpretation of ABC.

A

Algorithm rejABC

on Θ × X to sample from Eq.[2]

rejABC1 Sample θ ∼ π(θ), simulate x ∼ f(θ) and compute ε = ρ(S(x), S(x₀)).

rejABC2 Accept (θ, x) with probability proportional to κ(ε; τ), and go to rejABC1.

Algorithm rejABCμ

on Θ × ℝ^K to sample from Eq.[3]

rejABCμ1 Sample θ ∼ π(θ), simulate x ∼ f(θ) and compute ε_k = ρ_k(S_k(x), S_k(x₀)), k = 1,..., K.

rejABCμ2 Accept (θ, ε₁,..., K) with probability proportional to \prod_k κ(ε_k; τ_k), and go to rejABCμ1.

Table 1 Rejection samplers for ABC and ABCμ.

2.1 The ABC approach

As in most ABC settings, we suppose that pseudo-data x ∈ X can be efficiently simulated for any vector of model parameters θ ∈ Θ from a data-generating process f that defines, perhaps implicitly, the likelihood. We also consider a set of K summaries S = \{S₁,...,S_K\}, a real-valued distance function ρ, and the one-dimensional ABC kernel

κ(ε; τ) = 1/τ \mathbb{I} \{|ε| ≤ τ/2\}

with tolerance τ > 0. To circumvent likelihood evaluations, Pritchard et al (1999) first proposed the rejection sampler rejABC in Table 1A.

The target density of rejABC on the augmented space Θ × X is therefore

\[ \pi(θ, x | x₀) \propto \kappa(ρ(S(x), S(x₀)); τ) f(x | θ) π(θ). \] (2)

In typical applications, the auxiliary variable x is extremely high-dimensional, and lower-dimensional summaries are used to compare the simulated data with the observed data. The realized error ε = ρ(S(x), S(x₀)) is then accepted by the ABC algorithm when within a prescribed tolerance τ. ABC is a valid, non-parametric estimation method in that, as τ → 0, the marginal density π(θ | x₀) of (2) approaches the true posterior distribution \( \pi(θ | x₀) \) if the summaries are sufficient for θ under the model. Otherwise, (2) converges to the posterior distribution \( π(θ | S(x₀)) \) when τ → 0. Fearnhead and Prangle (2010) and Dean et al (2011) show that an ABC-based inference is converging (in the number of observations) if the parameter θ is identifiable in the distribution of S(x).
2.2 ABC on error space

The error \( \varepsilon \) computed in algorithm rejABC (Table 1B) is, in fact, a compound error term that may reflect both stochastic fluctuations in simulating from \( f \) as well as systematic biases between \( f \) and the data. To exploit this information, we reformulate ABC as providing simulations on the joint space of model parameters \((\theta)\) and summary errors \((\varepsilon_{1:K})\). Algorithm rejABC\( \mu \) in Table 1B uses the projection

\[
\xi_{x_0,\theta} : \mathcal{K} \to \mathbb{R}^K, \quad x \mapsto \varepsilon_{1:K} = (\varepsilon_1, \ldots, \varepsilon_K),
\]

(3)

\( \varepsilon_k = \rho_k (S_k(x), S_k(x_0)) \), which induces the image measure (abusively denoted by)

\[
\xi_{x_0,\theta} (E_1 \times \cdots \times E_K) = \mathbb{P}_f \left( \xi^{-1}_{x_0} (E_1 \times \cdots \times E_K) \big| \theta \right)
\]

(4)

on the associated Borel image \( \sigma \)-algebra, conditional on \( x_0, \theta \). The density of (4) with respect to a suitable measure on the \( K \)-dimensional error space will be denoted (again abusively) by

\[
\xi_{x_0,\theta} : \mathbb{R}^K \to \mathbb{R}_0^+, \quad \varepsilon_{1:K} \to \xi_{x_0,\theta}(\varepsilon_{1:K}).
\]

(5)

This multi-dimensional error density is thus the image of the sampling density \( f(\cdot|\theta) \) by the transform \( \xi_{x_0,\theta} \). It can be interpreted as the prior predictive error density conditional on \( \theta \).

Example 1 In many applications, pseudo-data \( x \) is simulated on a finite space \( \mathcal{X} \). Then, \( f(dx|\theta) \) is a counting measure, say

\[
f(dx|\theta) = \sum_{i=1}^{N_e} f_i \delta_{x_i}(dx).
\]

Hence, the image measure \( \xi_{x_0,\theta}(dx) \) is again a counting measure, say

\[
\xi_{x_0,\theta}(dx) = \sum_{j=1}^{N_e} \xi_j \delta_{x_j}(dx).
\]

where \( N_e \) is the size of \( \xi_{x_0,\theta}(\mathcal{X}) \), \( 0 < N_e \leq N_e^2 < \infty \).

From a computational perspective, (5) is in practice intractable. In direct analogy to ABC, we circumvent numerical evaluations of (5) through simulating from this density as illustrated in algorithm rejABC\( \mu \) in Table 1B. The target density of rejABC\( \mu \) on the augmented space \( \Theta \times \mathbb{R}^K \) is

\[
\pi_\tau (\theta, \varepsilon_{1:K}|x_0) \propto \prod_k \kappa (\varepsilon_k; \tau_k) \xi_{x_0,\theta}(\varepsilon_{1:K}) \pi (\theta).
\]

(6)

By construction, the marginal target densities \( \pi_\tau (\theta|x_0) \) of rejABC and rejABC\( \mu \) coincide if \( \kappa (\rho (S(x), S(x_0); \tau) \) in rejABC can be written, up to a constant of proportionality, as \( \prod_k \kappa (\rho_k (S_k(x), \xi_{x_0,\theta}(\varepsilon_{1:K}); \tau_k) \) (to be used in rejABC\( \mu \)) for some choice of \( \tau_k \); see the Appendix. For example, if ABC is run with the standard indicator kernel, this is equivalent to using the Manhattan distance

\[
\rho (S(x), S(x_0)) = \max \{ \rho_k (S_k(x), S_k(x_0)) \}
\]

and \( \tau_k = \tau \). The utility of this reformulation was first discussed in Ratmann et al (2009): it is possible to relate the marginal density \( \pi_\tau (\varepsilon_{1:K}|x_0) \) of (6) to standard Bayesian error measures. The marginal ABC error density can be understood as the prior predictive error density (Box 1980) that is re-weighted by error magnitude

\[
\pi_\tau (\varepsilon_{1:K}|x_0) \propto \prod_k \kappa (\varepsilon_k; \tau_k) \pi_{x_0}(\varepsilon_{1:K}).
\]

(7)

3 Extending existing ABC algorithms

Existing ABC algorithms are easily extended to sample from \( \pi_\tau (\theta, \varepsilon_{1:K}|x_0) \) for the purpose of parameter inference and model assessment.

In Table 2 we contrast the Metropolis-Hastings ABC sampler (Marjoram et al 2003) to its extension that samples from (6). To demonstrate the validity of algorithm mhABC\( \mu \), let \( z = (\theta, \varepsilon_{1:K}) \) and note that, on the augmented space, the proposal density of mhABC\( \mu \) is

\[
q(z \to z') = \xi_{x_0,\theta}(\varepsilon_{1:K}) q(\theta \to \theta') \pi (\theta) \pi_{x_0}(\varepsilon_{1:K}).
\]

Therefore, detailed balance is satisfied precisely for \( \pi_\tau (\theta, \varepsilon_{1:K}|x_0) \):

\[
\frac{mh(z, z')}{mh(z', z)} = \frac{q(\theta' \to \theta) \pi (\theta') \prod_k \kappa (\varepsilon_k'; \tau_k)}{q(\theta \to \theta') \pi (\theta) \prod_k \kappa (\varepsilon_k; \tau_k)}
\]

\[
= \frac{q(z' \to z) \pi (\theta') \prod_k \kappa (\varepsilon_k'; \tau_k) \xi_{x_0,\theta}(\varepsilon_{1:K})}{q(z \to z') \pi (\theta) \prod_k \kappa (\varepsilon_k; \tau_k) \xi_{x_0,\theta}(\varepsilon_{1:K})}
\]

\[
= \frac{q(z' \to z) \pi_\tau (z'|x_0)}{q(z \to z') \pi_\tau (z|x_0)}.
\]

Bortot et al (2007) proposed a Metropolis-Hastings sampler on the space \( \Theta \times \mathcal{X} \times [0, \infty) \) for the purpose of parameter inference when the tolerance \( \tau \) is by design a random variable. For clarity, we note that this algorithm requires an extra proposal density \( q(\tau \to \tau') \), and has a target density different to both (2) and (6).

In Table 3, we contrast the popular sequential importance sampler (SIS) for ABC (Toni et al 2008) to its extension that samples from (6). To keep the particle system at the \( n \)th stage alive, these algorithms augment the state space with an additional random variable \( I_n \in \{1, \ldots, N\} \). This is the ancestor index of the particle at
Algorithm sisABC
on \(N \times \Theta \times X\) to sample, marginally, from Eq.\[2\]
Set the initial particle system at \(n = 1\) for
\(i = 1, \ldots, N\) compute \(\theta_1^i \sim \pi(\theta), x_1^i \sim f(\cdot; \theta_1^i),\)
\(\varepsilon_1^i = \rho(S(x_1^i), S(x_0))\), and then, for \(i = 1, \ldots, N,\)
\[W_1^i = \kappa(\varepsilon_1^i; \tau_1) / \sum_{j=1}^N \kappa(\varepsilon_1^j; \tau_1).\]

For \(n = 2, \ldots, n^*,\) do:
Set \(i = 1, c_n = \kappa(0; \tau_n)\) and repeat:

sisABC1 Propose the \(i\)th ancestor index \(I'\) from \(i = 1, \ldots, N\) with probabilities \(W_{i-1}\)
\[M_n(\theta_{I'-1}^i; \cdot), x' \sim f(\cdot; \theta')\) and compute \(\varepsilon' = \rho(S(x'), S(x_0)).\)

sisABC2 With probability \(\kappa(\varepsilon'; \tau_n)/c_n,\) set \((I_n^i, \theta_n^i, x_n^i) \leftarrow (I', \theta', x').\)
Compute the unnormalized weight
\[w_n^i = \pi(\theta_n^i) / \sum_{j=1}^N W_{n-1}^j M_n(\theta_{j-1}^i; \theta_n^i)\]
and increment \(i \leftarrow i + 1.\) If \(i = N,\) go to sisABC3.
Else return to sisABC1.

sisABC3 Compute the normalized weights
\[W_n^i = w_n^i / \sum_{i=1}^N w_n^i\]
and update \(n \leftarrow n + 1.\) If \(n < n^*,\) go to sisABC1.

Algorithm sisABC\(\mu\)
on \(N \times \Theta \times \mathbb{R}^K\) to sample, marginally, from Eq.\[6\]
Set the initial particle system at \(n = 1\) for
\(i = 1, \ldots, N\) compute \(\theta_1^i \sim \pi(\theta), x_1^i \sim f(\cdot; \theta_1^i),\)
\(\varepsilon_{1k} = \rho_k(S_k(x_1^i), S_k(x_0))\), and then, for \(i = 1, \ldots, N,\)
\[W_1^i = \prod_k \kappa(\varepsilon_{1k}^i; \tau_{1k}) / \sum_{j=1}^N \prod_k \kappa(\varepsilon_{jk}^i; \tau_{jk}).\]

For \(n = 2, \ldots, n^*,\) do:
Set \(i = 1, c_n = \prod_k \kappa(0; \tau_{nk})\) and repeat:

sisABC\(\mu\)1 Propose the \(i\)th ancestor index \(I'\) from \(i = 1, \ldots, N\) with probabilities \(W_{i-1}\)
\[M_n(\theta_{I'-1}^i; \cdot), x' \sim f(\cdot; \theta')\) and compute \(\varepsilon' = \rho_k(S_k(x'), S_k(x_0))\) for all \(k.\)

sisABC\(\mu\)2 With probability \(\prod_k \kappa(\varepsilon_{k}'; \tau_{nk})/c_n,\) set \((I_n^i, \theta_n^i, \varepsilon_n; \kappa^i) \leftarrow (I', \theta', \varepsilon'_{1K}).\)
Compute the unnormalized weight
\[w_n^i = \pi(\theta_n^i) / \sum_{j=1}^N W_{n-1}^j M_n(\theta_{j-1}^i; \theta_n^i)\]
and increment \(i \leftarrow i + 1.\) If \(i = N,\) go to sisABC\(\mu\)3.
Else return to sisABC\(\mu\)1.

sisABC\(\mu\)3 Compute the normalized weights
\[W_n^i = w_n^i / \sum_{i=1}^N w_n^i\]
and update \(n \leftarrow n + 1.\) If \(n < n^*,\) go to sisABC\(\mu\)1.

Table 3 Vanilla sequential importance samplers for ABC and ABC\(\mu\).
Both algorithms require to specify a decreasing sequence of tolerances \(\tau_n\) for \(n = 1, \ldots, n^*.\) The first tolerance \(\tau_1\) is here set large enough such that \(W_1^i > 0\) for all \(i = 1, \ldots, N,\) and subsequent ones can be set automatically [Del Moral et al 2008] or according to an annealing scheme. Using the indicator kernel, the acceptance probability in sisABC2 and sisABC\(\mu\)2 is either zero or one [Toni et al 2009]. Typically, the variance of the proposal kernel \(M_n\) is modified at each stage to improve the convergence of the algorithm [Beaumont et al 2010].
Fig. 1 Discrepancies between several network evolution models and protein network data from *T. pallidum*. We show results for two alternative evolution models and two different observation models, (A) DD+LNK+PA-L, (B) DD+PA-L, (C) DD+PA-BP; results for DD+LNK+PA-BP are similar to (A) and not shown. Left column: trajectories of the CC error for four Markov chains that are generated in parallel. Convergence was most difficult to achieve for the model in (A) and involved re-parameterization because the support of some parameters spanned several orders of magnitude. Right columns: two dimensional estimates of the seven dimensional ABC error density, as reproduced with average shifted histograms.
When large-scale, protein-protein interaction data became available, it came as a surprise that their topological features deviate markedly from those expected under standard random graphs. A variety of simple mathematical models were subsequently proposed to explain some of these unexpected topological features (Stumpf et al., 2007). These models grow networks iteratively node by node until a given finite network size is reached. The preferential attachment model (PA) was able to reproduce, roughly, the observed fat-tailed empirical distribution of node degrees (number of outgoing edges of a node). Subsequently, more biologically plausible models based on the duplication of nodes and (local) link divergence among the duplicates (DD), as well as (global) link rewiring were proposed (LNK).

We analyzed mixes of these mechanisms on protein network data that was obtained with high-throughput bait-prey technologies (Ratmann et al., 2009). Briefly, the DD+PA mixture model has three parameters, and DD+LNK+PA has five parameters. Available network data reflects only a subset of the true interaction network. To interface the evolution models with data, several observation models have been formulated. Given a simulated complete network, we can randomly sample links until the observed number of links is matched (L). Alternatively, we can mimic the experimental design by randomly sampling bait and prey proteins until the observed numbers are matched, and record associated links (BP); see the Appendix for details. These sampling schemes do not add any parameters. All prior densities are uninformative.

To compare observed network data $x_0$ to the corresponding simulations $x$, we use seven summary statistics that reflect local and global properties of the network topology: average node degree ($\delta$); within-reach distribution ($WR$), the mean probability of how many edges to nodes outside the box. The choice of these summaries is discussed in (Ratmann et al., 2007).
We applied algorithm mhABCµ in Table 2B with annealing on the ABC thresholds and the variance of the Gaussian proposal density \( q(\theta \rightarrow \theta') \) across a set of network evolution models. The sampler converged rapidly to the target density \( f_\theta \), as assessed by four parallel Markov chains that were started at overdispersed values. Markov chains were highly autocorrelated. Figure 1 displays representative Markov chain trajectories, and two dimensional estimates of the seven dimensional ABC error density for three different models.

Inspection of the multi-dimensional ABC error density enabled us to diagnose specific deficiencies in models of network evolution. In our approach to ABC, each error \( \varepsilon_k \) corresponds directly to one summary, thereby enabling targeted model refinement. For example, Figure 1A illustrates that the fitted link rewiring model with link subsampling (DD+LNR+PA-L) fails to match local connectivity patterns (CC) in the observed Treponema pallidum network [Titz et al. 2008] while it reproduces global topological patterns (ND, WR). This prompted us to remove the link rewiring component and to consider DD+PA-L, which results in some improvement (Figure 1B). We also replaced link subsampling with bias-prey sampling. The fitted DD+PA-BP model is consistent with the seven considered aspects of the data in that the point \( \varepsilon_{1,K} = 0 \) is well in the support of the ABC error density (Figure 1C).

5 Application of sisABCµ to dynamic systems

The behavior of complex dynamic systems is often described by sets of non-linear ordinary or stochastic differential equations. Typically, these systems are only partially observed in aggregated form, and associated models are not analytically solvable and may fail to reproduce important aspects of the system’s dynamics. Toni et al. [2008] proposed sequential ABC algorithms (in Table 3A) for parameter inference in this setting. In this section, we demonstrate with the following example that the algorithm in Table 3B can expose deficiencies of partially observed, nonlinear dynamic models.

The disease dynamics of influenza infecting humans in temperate regions are characterized by explosive, seasonal epidemics during the winter months and marked, irregular variation across consecutive seasons. To understand the epidemiology of seasonal influenza, various complex non-linear dynamic models that track the number of susceptible (S), infected (I) and recovered/immune (R) individuals in a population have been considered [Earn et al. 2002]. Much of this work is motivated by the fact that simple models that assume gradual loss of immunity to reinfection do not fully describe observed disease patterns. We consider a stochastic SIRS model defined by the transmission rates

\[
\begin{align*}
\frac{dS}{dt} &= \mu(N - S) - \beta t \frac{S}{N} I + \gamma(N - S - I), \\
\frac{dI}{dt} &= \beta t \frac{S}{N} I - (\mu + \nu) I,
\end{align*}
\]

where \( \mu \) is the birth/death rate, \( N \) is the finite population size, \( \nu \) is the recovery rate, \( \gamma \) the rate by which immunity is lost and \( \beta \) the transmission rate that is further assumed to vary seasonally as \( \beta t = \beta (1 + s \sin(2\pi t)) \).

In practice, we account for long-term demographic trends (thus, \( N \) and \( \mu \) are fixed), and reparameterize \( \beta, \nu, \gamma \) in terms of the basic reproductive number \( R_0 = \beta(1/(\nu + \mu)) \), the average duration of infection per day, \( D = 1/\nu \), and the average duration of immunity per year, \( \Gamma = (365\gamma)^{-1} \). See the Appendix for further details. We fit and contrast \( 5 \) to weekly Dutch influenza-like illness data \( \{x_t\} \) of influenza A (H3N2) that was collected between 1968–99 (http://www.nivel.nl/griep/). Influenza-like illness data is subject to fluctuating reporting biases. We use a Poisson observation model that accounts for unknown reporting biases \( \rho \) and known seasonal fluctuations in reporting fidelity \( f_t \) (see Appendix),

\[
\frac{dx}{dt} = \rho f_t^{-1} I,
\]

yielding five unknown parameters \( \theta = (R_0, D, \Gamma, s, \rho) \) in total. The prior densities are \( R_0 \sim U(1, 20), \Gamma \sim U(1, 160), s \sim U(0.075, 0.6), \rho \sim U(0.04, 0.4) \) (uninformative), and \( D \sim U(2.2, 2.8) \) (informative) [Gupta et al. 1998]. We use the standard Euler multinomial scheme [Tian and Burage 2004] to simulate from \( \theta \).

To compare the observed influenza-like illness counts \( x_0 \) to simulated data \( x \), we use five summary statistics that reflect the characteristic dynamic features of influenza A (H3N2). Interannual seasonal variation is primarily reflected by the cumulative distribution and autocorrelation in peak differences (CDF-ΔPK, ACF-ΔPK). The explosiveness of seasonal epidemics is reflected by the average duration of an epidemic at or above half its peak size (M-EXPL), and we query overall magnitude with the cumulative distribution of peak incidence (CDF-PK) and the average annual attack rate (M-ATTR). Here, annual attack rates are computed directly as the ratio of cumulative influenza-like illness counts in a winter season against average population size in that season; see the Appendix for further details.
The sequential algorithm sisABCμ (Table 3) readily detects several shortcomings of the nonlinear stochastic SIRS model, while fitting it simultaneously to the data. Here, we actually seeded sisABCμ with a few samples from algorithm mhABCμ taken after burn-in, rather than seeding from the prior density. As we discuss later, this hybrid approach (hybridABCμ) often improves overall efficiency. Our sequential sampler was then run for two iterations at the final ABC thresholds with 1000 particles. We find that model CL.EMERGE produces disease dynamics with too regular and too strong interannual variation. In Figure 2 we display a representative particle sampled at the last ABC iteration and two-dimensional projections of the estimated five-dimensional ABC error density.

We also compared the numerical efficiency of algorithms mhABCμ, sisABCμ, as well as the hybrid approach in terms of simulation effort per effectively independent sample from the target density. Here, we present results on the SIRS model (8). Algorithm mhABCμ was run with annealing on the ABC threshold and the variance of the proposal density. Four chains were generated in parallel. For comparison, burn-in of mhABCμ is here the number of iterations to anneal...
Table 4 Case study to analyze the performance of algorithms mhABCμ, sisABCμ, and the sequential importance sampler that is seeded with mhABCμ (hybridABCμ) on the example in Section 5. The first column gives the burn-in of the algorithms under investigation. The second column reports effective sample size per 1000 samples from (6) after burn-in. The third column gives the number of total simulations, including burn-in, per effective sample.

| Algorithm    | Burn-in (ESS/1000) | #sim/ESS |
|--------------|--------------------|----------|
| mhABCμ      | 4639 [3963, 5178]  | 60 [49, 75] 859 [705, 1186] |
| sisABCμ     | 24286 [21117, 26481] | 125 [34, 184] 557 [374, 1834] |
| hybridABCμ  | 19263 [18884, 19605] | 117 [41, 200] 363 [194, 968] |

Fig. 3 The two evolutionary scenarios considered in Section 6. A coalescent process evolves under time over each branch of the two scenarios. Left: Model Div23, where Population 3 diverged from Population 2 at time $t = 30$. Right: Model Div13, where Population 3 diverged from Population 1 at time $t = 30$.

6 Application of ABCμ to population genetics

The ABC errors make possible to contrast a model against the real data in absolute terms. While this enables to successively refine any given model, there is, currently, no general approach to compare two models based on these errors. We illustrate the interplay of model checking and model comparison on a simple population genetics example where the true Bayes factor can be numerically estimated.

Population genetics seek to infer aspects of the evolutionary history of a population of related individuals from genetic data. We assume that the different populations evolve from a common ancestral population and that a tree topology encodes this history. The dynamics of changing genotype frequencies in populations that comply with the tree topology can often be represented by the Coalescent process, and methods have been developed to infer the associated process parameters from a small sample of genetic data of these populations (Crow and Kimura 1970; Kingman 1982).

We consider here the case of three populations having evolved according to one of the two models described in Figure 3. Under both models, some individuals were genotyped from each population at time $t = 0$. Populations 1 and 2 diverged at a known date ($t = 60$), and the deepest divergence event in the tree occurred at $t = 30$. Each sample from one of the three populations is composed of 15 diploid individuals, which have been genotyped at 5 independent microsatellite loci (different genotypes differ in the size of the repeated region). Evolution of these loci over time follows the above model, i.e., when a mutation occurs, the length of the sequence is increased or decreased by the length of the short repeated motif (Ohta and Kimura 1973). The per capita mutation rate was set to 0.005 for all five loci. We only infer the effective population size $N_e$, assuming homogeneity across branches of the different scenarios. The prior density of $N_e$ was set to $U(2, 150)$. This setup is quite simple in order to allow computation of the true posterior density with importance sampling techniques (combining techniques based on Stephens and Donnelly 2000).

To perform ABC analysis, the data are summarized through 24 statistics. Some of these numerical sum-
To compare different methods for ABC model choice in a controlled setting, we simulated two data sets under model Div23 with $N_c = 75$ and applied algorithm rejABC$_H$; see the Appendix. Both models are similar enough so that the evidence for model Div13 can be larger than the evidence for Div23. Figure 3A reports importance sampling estimates of the true Bayes factor as well as ABC approximations thereof. The former are computed by importance sampling approximations to both marginal likelihoods, while the later are computed as ratios of frequencies of acceptances of simulations from both models, see Robert et al. (2011) for details. Robert et al. (2011) demonstrated that the ABC approximation of the Bayes factor does not converge to the (numerically estimated) Bayes factor with increasing sample size and/or increasing computation runtimes. Considering the first data set (D1), the evidence for model Div23 is larger when compared to the evidence for Div13. For the second data set (D2), the situation is reversed, reflecting chance events with which the simulated data sets were generated. The ABC errors in Figure 4B reveal that, only, Div13 matches data set D2 reasonably well. Model Div13 shows larger discrepancies on this data set, which is in agreement with the Bayes factor. Turning to data set D1, the ABC errors show that none of the fitted models are consistent with the data, although the errors are slightly smaller under model Div23, again in agreement with the Bayes factor.

### 7 Discussion

We showed how existing ABC algorithms can be modified to aid targeted model refinement on the fly, and
illustrated these algorithms on examples from evolutionary biology and dynamic systems. Underlying this workhorse machinery is a simple re-interpretation of ABC as a particular data augmentation method on error space, and the presented algorithms are an immediate consequence of this reinterpretation. Previously, we recognized the utility of the ABC errors as (unknown but estimable) compound random variables that may reflect both stochastic fluctuations and systematic discrepancies between a model and the data (Ratmann et al. 2009). There are two elementary points to make. The joint distribution of ABC errors (7) faithfully reflects dependencies among different aspects of the full, intractable data. Indeed, the product ABC kernel that is used here does not confound the relation between multiple, lower-dimensional projections on error space. Thus, each dimension of the ABC error density retains an intrinsic meaning and can be used to diagnose specific model deficiencies. Second, the ABC errors are already computed by existing ABC algorithms (see Tables 1-3) and there is no further computational cost associated for the purpose of model assessment.

The ABC error density (7) provides an exploratory rather decisional tool for targeted, iterative model refinement. One obstacle in using (7) more formally for the purpose of model comparison is that the ABC errors have an intrinsic scale that is model-dependent, and are often not directly comparable. However, as there is no direct connection or convergence of approximations to the Bayes factor to the true Bayes factor between no direct connection or convergence of approximations are often not directly comparable. However, as there is no direct connection or convergence of approximations to the Bayes factor to the true Bayes factor between.

From a theoretical point, there is no possibility of conflict between summary errors whenever the respective summaries are independent of each other.

Based on our experience with the algorithms presented in Tables 2-3, we further recommend to combine mhABCμ and sisABCμ to sample from (6). Intuitively, mhABCμ updates only a single particle in relation to its previous value, and may lead to rapid convergence to the target density. Techniques aiding rapid convergence are discussed in Ratmann et al. (2007). As the Metropolis-Hastings sampler might become stuck Sisson et al. (2007), we run multiple chains in parallel until shortly after burn-in. Our acceptance probabilities of mhABCμ are typically below 5%, hence the generated Markov chains are highly autocorrelated. In our case study presented in Table 4, we find that sisABCμ, seeded with samples from mhABCμ, may quickly replenish effective sample sizes because it also proposes ancestor indices. We present a small case study in Table 4. There are perhaps two main caveats with this hybrid approach. First, all Markov chains generated by
mhABCµ might get stuck. However, this is a generic feature with all MCMC implementations that can be generically attenuated by the annealing nature of ABCµ (via the choice of the τnk’s). Second, the acceptance probabilities in the rejection control step of sisABCµ can be significantly lower than those of mhABCµ because the latter are not computed in relation to previous error magnitudes.

8 Conclusion

We presented methods and algorithms for exploratory model assessment when the likelihood is computationally intractable. To us, the advantages of this approach towards ABC model choice are that sufficient summaries are often not required to detect existing discrepancies between a model and the data, and that the algorithms incur no or little extra computational cost as compared to standard ABC methods. Perhaps, the main shortcomings of this approach may be the scale dependency of the ABC errors and their sensitivity to the ABC threshold. We believe that, in addition to investigating the effects of the ABC approximation to standard Bayesian machinery such as the Bayes factor or the deviance information criterion, exploiting the particular properties of that approximation may provide complementary, useful tools for ABC model choice.

A Proofs of sections 2 and 7

We first derive the marginal density of \( \theta \) in \( \theta \). By construction,

\[
\pi_{ABC}(\theta|X_0) \propto \pi(\theta) \int \kappa(\varepsilon_{1,K}|\tau_0) \xi_{S(0)}(dx_1 ... dx_K) f(dx|\theta).
\]

Consequently, the marginal target densities \( \pi_\tau(\theta|x_0) \) of algorithms rejABC and rejABCµ coincide if

\[
\kappa(\rho(S(x),S_0(x));\tau) \propto \prod_k \kappa(\rho_k(S_k(x),S_k(x_0));\tau_k).
\]

(We conjecture this property only happens for the normal kernel \( \kappa \) and the Euclidean distance \( \rho \), as well as for the indicator kernel and the Manhattan distance.) To establish 4, consider the following error measure on the (same) associated kernel and the Manhattan distance.) To establish (7), kernel (We conjecture this property only happens for the normal \( \kappa \).

\[\pi_{\pi}(\varepsilon_{1,K}|X_0) \propto \prod_k \kappa(\varepsilon_{\theta_k}|\tau_k) \pi_{\pi}(\varepsilon_{1,K}).\]

We previously derived the same relations under a Riemann interpretation of the integrals involved (Ratmann et al. 2009). The Lebesgue approach presented here is much less convoluted.

Finally, we establish the intuitive result that conflict cannot emerge between independent summary errors. If the summary errors in ABC are independent of each other,

\[
\pi_\pi(\varepsilon_{1,K}|X_0) = \prod_{k=1}^K \pi_{\pi_k}(\varepsilon_{\theta_k};\tau_k) / \int \pi_{\pi_k}(\varepsilon_{\theta_k};\tau_k) d\varepsilon_1 ... d\varepsilon_K
\]

factorizes componentwise. Likewise, the prior predictive \( \kappa \) of \( \varepsilon_{1,K} \) admits the factorization \( \pi_{\pi}(\varepsilon_{1,K}) = \prod_{k=1}^K \pi_{\pi_k}(\varepsilon_{\theta_k}) \). Since the ABC kernel is here assumed to factorize as well, we obtain

\[
\pi_\pi(\varepsilon_{1,K}|X_0) = \prod_{k=1}^K \pi_{\pi_k}(\varepsilon_{\theta_k}) / \int \pi_{\pi_k}(\varepsilon_{\theta_k}) d\varepsilon_1 ... d\varepsilon_K
\]

The 4th component of the vector-valued mean ABC error \( E_{\pi}(\varepsilon_{1,K}|X_0) \) collapses to

\[
\int \varepsilon_k \prod_{k=1}^K \pi_{\pi_k}(\varepsilon_{\theta_k}|X_0) d\varepsilon_1 ... d\varepsilon_K
\]

\[
= \int \varepsilon_k \pi_{\pi_k}(\varepsilon_{\theta_k}|X_0) \left[ \int \cdots \int \pi_{\pi_j}(\varepsilon_j|x_0) d\varepsilon_{-j} \right] d\varepsilon_k
\]

\[
= \int \varepsilon_k \pi_{\pi_k}(\varepsilon_k|x_0) d\varepsilon_k
\]

\[
= \int \rho_k(S_k(x),S_k(x_0)) \pi_{\pi_k}(\varepsilon_k|\tau_k) \pi(dx).
\]

B Simulation from the network models

Each of the considered network evolution models defines a discrete-state discrete time Markov chain of growing networks. For network data of some organism, we grow a network according to model-specific transition probabilities until the number of nodes in the network equals the number of genes in the genome of that organism. Transition probabilities are implicitly defined by the following probabilistic rules. The preferential attachment (PA) model adds a new node to an existing node with probability that is proportional to the node degree of the existing node. In the duplication-divergence model
(DD), a parent node is randomly chosen and its edges are duplicated. For each parental edge, the parental and duplicated one are then lost with probability $2p_v$, each, but not both; moreover, at least one link is retained to any node. The parent node may be attached to its child with probability $d_A$. The mixture model DD+PA either performs PA with probability $\alpha$, or DD with probability $1-\alpha$. Model DD+LNK+PA is a mixture of PA, DD with $d_A = 0$, link addition and deletion. Link addition (deletion) proceeds by choosing a node randomly, and attaching it preferentially to another node (deleting it preferentially from its interaction partners). Unnormalized mixture weights are calculated as follows. For duplication-divergence, the rate $\lambda_{dup}$ is multiplied by the order of the current network; for link addition, the rate $\lambda_{Add}$ is multiplied by $\left(\frac{\text{Order}}{\text{Size}}\right)$; for link deletion, the unnormalized weight of link addition is multiplied by $\lambda_{Del}$. Preferential attachment occurs at a constant frequency $\alpha$, and the weights of duplication, link addition and link deletion are normalized so that their sum equals $1-\alpha$. Each of the components is chosen according to the normalized weights.

We consider here two alternative observation models to account for missing data. Denote the number of proteins in the observed data set $x_0$ with $n(x_0)$, the number of bait proteins with $n_{bait}(x_0) \leq n(x_0)$, the number of prey proteins with $n_{prey}(x_0) \leq n(x_0)$, the number of links with $m(x_0)$ and the fully simulated network with $x$. Link subsampling (L): Set $x$ to be empty and repeat until the number of links in $x$ is equal or larger than $m(x_0)$, or no more links can be added: pick a link $(u,v)$ at random and without replacement from $x$ and add the nodes $u,v$ and the link $(u,v)$ to $x$. Bait-prey subsampling (BP): Create two random lists of bait and prey proteins. Set $x$ to be empty and repeat until the number of links $n_{bait}(x) = n_{bait}(x_0)$ and the number of prey $n_{prey}(x)$ is equal or larger than $n_{prey}(x_0)$, or no more links can be added: pick a link $(u,v)$ at random and without replacement from $x$ such that $u$ is in the bait list and $v$ is in the prey list, mark $u$ as a bait and $v$ as a prey protein and add the nodes $u,v$ and the link $(u,v)$ to $x$.

C Simulation from the stochastic SIRS model

Assuming that the infinitesimal probability of either simultaneous or multiple transitions between compartments is negligible, the transmission rates $\beta_1$ define a finite-state continuous-time Markov chain that accounts for demographic stochasticity. To adjust for long-term demographic trends in the Netherlands, we fix $N$ to historical population data obtained from [http://statline.cbs.nl/statweb/], and set $1/\mu = 80$. To avoid stochastic extinction, we add a small, constant number of infected visitors $I_0$, which can be interpreted as the average number of infected travelers that are visiting the Netherlands at any day. The first equation in (5) is thus

$$\frac{dS}{dt} = \mu(N - S) - \beta_1 \frac{S}{N} (I + I_v) + \gamma (N - S - I).$$

The value $I_0$ is set to the expected number of infected travelers for a given model parameterization. More precisely, we estimate the average total number of international travelers in the Netherlands at 5m/yr from available tourist information [http://statline.cbs.nl/statweb/] and set $I_0$, the proportion of international travelers that would be infected at any day at endemic equilibrium,

$$I_v = \frac{\mu + \gamma}{\mu + \gamma + \nu} (1 - 1/R_0) \times 5 \times 10^6/365.$$
We sum up the whole simulated data set with the 24 usual summary statistics given in Robert et al. [2011]. In the simple situation considered here, it is possible to apply the rejection algorithm, and in fact, we here re-evaluated the same pseudo data sets that were also used in Robert et al. [2011]. The acceptance function \( \kappa \), see Table 1, is a product of 24 one-dimensional indicator functions centered around 0. And the tolerances \( \tau \) are tuned so that we keep in Model Div23 about 1000 simulated data sets from the 106 simulated data sets of the reference table. Our conclusions on model assessment are correlated with the distances in genetic variation between populations.

Acknowledgements We would like to thank Jean-Marie Cornuet for the simulations with the DIYABC software that were used in Section 6, and Christophe Andrieu for stimulating discussions.

References

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Ann Rev Ecol Ev and Syst 41(1):379–406
Beaumont MA, Cornuet JM, Marin JM, Robert CP (2010) Adaptivity for ABC algorithms: the ABC-PMC scheme. Biometrika 96(4):983–990
Blum M (2009) Approximate Bayesian Computation: a non-parametric perspective. Tech. Rep. 0904.0635v4, arXiv
Bortot P, Coles S, Sisson S (2007) Inference for stereological extremes. J Am Stat Ass 102(9):84–92
Box GEP (1980) Sampling and Bayes’ inference in scientific modelling and robustness. J Roy Stat Soc A (General) 143(4):383–430
Chakraborty R, Jin L (1993) A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. EXS 67:153–175
Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ, Guillemaud T, Estoup A (2008) Inference population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. Bioinformatics 24(23):2713–2719
Crow JF and Kimura M (1970) An introduction to population genetics theory. Harper & Row, Publishers
Dean TA, Singh A S Sand Jasra, Peters G (2011) Parameter estimation for hidden markov models with intractable likelihoods, arXiv:11035399
Del Moral P, Doucet A, Jasra A (2008) An adaptive Sequential Monte Carlo method for Approximate Bayesian Computation. Tech. rep., University of Bordeaux, France, URL http://stats.ma.ic.ac.uk/a/a12/public.html/papers/delmoral_doucet_jasra_empabc.pdf
Drovandi CC, Pettitt AN, Faddy MJ (2011) Approximate Bayesian computation using indirect inference. J Roy Stat Soc (C) 60(3):317–337
Earn DDD, Dushoff J, Levin SA (2002) Ecology and evolution of the flu. Trends Ecol Evol 17(7):334–340
Fearnhead P, Prangle D (2010) Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic ABC. ArXiv e-prints 1004.1112
Francois O, Laval G (2011) Deviance Information Criteria for Model Selection in Approximate Bayesian Computation. ArXiv e-prints 1105.0269
Grelaud A, Robert CP, Marin JM, Rodolphe F, Taly JF (2009) Likelihood-free methods for model choice in Gibbs random fields. Bayesian Analysis 4(2):317–336
Gupta S, Ferguson N, Anderson R (1998) Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. Science 280(5385):912–915
Heggland K, Fregiess A (2004) Estimating functions in indirect inference. J Roy Stat Soc B 66:447–462
Jaakkola TS, Jordan MI (2000) Bayesian parameter estimation via variational methods. Statistics and Computing 10:25–37
Kingman JFC (1982) The Coalescent. Stoch. Proc. and Their Applications 13: 235–248
Liu J, Chen R (1998) Sequential Monte Carlo Methods for Dynamic Systems. J Am Stat 93:1032–1044
MacKay DJC (2002) Information Theory, Inference & Learning Algorithms. Cambridge University Press
Marin JM, Pudlo P, Robert CP, Ryder R (2011) Approximate Bayesian Computational methods. ArXiv e-prints 11010955
Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov Chain Monte Carlo without likelihoods. Proc Natl Acad Sci USA 100(26):15,324–15,328
Nunes MA, Balding DJ (2010) On optimal selection of summary statistics for approximate Bayesian computation. Stat App Gen Mol Biol 9(1)
Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetics Research 22:201–204
Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16:1791–1798
Rannala B, Mountain J (1997) Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci USA 94:9197–9201
Ratmann O, Jorgensen O, Hinkley T, Stumpf MP, Richardson S, Wifuf C (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H.pylori and P.falciparum. PLoS Comp Biol 3(2007):e230
Ratmann O, Andrieu C, Wifuf C, Richardson S (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. Proc Natl Acad Sci USA 106(26):10,576–10,581
Ratmann O, Andrieu C, Wifuf C, Richardson S (2010) Reply to Robert et al.: Model criticism informs model choice and model comparison. Proc Natl Acad Sci USA 107(3):E6–E7
Ranneh B, Mountain J (1997) Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci USA 104(26):10,576–10,581
Ratmann O, Andrieu C, Wifuf C, Richardson S (2010) Reply to Robert et al.: Model criticism informs model choice and model comparison. Proc Natl Acad Sci USA 107(3):E6–E7
Robert CP, Mengsen K, Chen C (2009) Letter: Model choice versus model criticism. Proc Natl Acad Sci USA 106(26):10,576–10,581
Robert CP, Mengersen KL, Chen C (2009) Letter: Model choice versus model criticism. Proc Natl Acad Sci USA 106(26):10,576–10,581
Ranneh B, Mountain J (1997) Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci USA 104(26):10,576–10,581
Sissons SA, Fang Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. Proc Natl Acad Sci USA 104(26):10,576–10,581
Sokal A (1989) Monte Carlo methods in statistical mechanics: foundations and new algorithms. Tech. rep., Department of Physics, New York University
Stephens D and Donnelly P (2000) Inference in population genetics (with discussion). J. Royal Statist. Soc. 62:602–655.
Stumpf MPH, Kelly W, Thorne T, Wifuf C (2007) Evolution at the system level: the natural history of protein interaction networks. Trends Ecol Evol 22:366–373
Tian T, Burrage K (2004) Binomial leap methods for simulating stochastic chemical kinetics. J Chem Phys 121(21):10,356–10,364, DOI 10.1063/1.1810475
Titz B, Rajagopala SV, Goll J, Häuser R, McKevitt MT, Palzkill T, Uetz P (2008) The binary protein interactome of Treponema pallidum – the Syphilis spirochete. PLoS One 3(5):e2292
Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2008) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J Roy Soc Interface
Weir BS, Cockerham CC (1984) Estimating f-statistics for the analysis of population structure. Evolution 38:1358–1370