Considerate Approaches to Achieving Sufficiency for ABC model selection

Chris Barnes[1,*], Sarah Filippi[1,*], Michael P.H. Stumpf[1,*,#], Thomas Thorne[1,*]

[1]Centre for Integrative Systems Biology and Bioinformatics,
Imperial College London, London SW7 2AZ, UK.

July 4, 2011

Abstract

For nearly any challenging scientific problem evaluation of the likelihood is problematic if not impossible. Approximate Bayesian computation (ABC) allows us to employ the whole Bayesian formalism to problems where we can use simulations from a model, but cannot evaluate the likelihood directly. When summary statistics of real and simulated data are compared — rather than the data directly — information is lost, unless the summary statistics are sufficient. Here we employ an information-theoretical framework that can be used to construct (approximately) sufficient statistics by combining different statistics until the loss of information is minimized. Such sufficient sets of statistics are constructed for both parameter estimation and model selection problems. We apply our approach to a range of illustrative and real-world model selection problems.

* All authors contributed equally.
# To Whom Correspondence Should be Addressed: m.stumpf@imperial.ac.uk

1 Introduction

Mathematical models are widely used to describe and analyze complex systems and processes across the natural, engineering and social sciences. Formulating a model to describe, e.g. a predator-prey system, geophysical process, communication system, or social network requires us to condense our assumptions and knowledge into a single coherent framework (May, 2004). In constructing these models we have to state our assumptions about their constituent parts and their interactions explicitly. Mathematical analysis or computer simulations of these models then allows us to compare model predictions with experimental observations in order to test and ultimately improve the models. Even previously largely observational sciences such as biology, geology and meteorology are now heavily influenced by computer simulations which are employed for explanatory as well as predictive purposes.

Because many of the mathematical models in these disciplines are too complicated to be analyzed in closed form, computer simulations have become the primary tool in the theoretical analysis of very large or complex models. Modelling such systems is often (relatively) straightforward if the mathematical structure of the model and reliable estimates for the model parameters are known. Unfortunately, it is considerably harder to infer the structure of mathematical models and estimate their respective parameters based on experimental data, in particular if we seek to infer a model that can describe the data. Whenever probabilistic models exist we can employ standard model selection approaches of either a frequentist, Bayesian, or information theoretic nature (Cox & Hinkley, 1974; Mackay, 2003; Burnham & Anderson, 2002). But if suitable probability models do not exist, or if the evaluation of the likelihood is computationally intractable, then we have to base our assessment on the level of agreement between simulated and observed data. This is particularly challenging when the parameters of simulation models are not known but must be inferred from observed data as well.

For such cases — cases where conventional statistical approaches fail because of the enormous computational burden incurred in evaluating the likelihood — so-called approximate Bayesian computation (ABC) schemes have recently come to the fore (Pritchard *et al.*, 1999; Beaumont *et al.*, 2002; Tanaka *et al.*, 2006; Secrier *et al.*,

2009). These forgo the explicit evaluation of the likelihood by a principled comparison between the observed and simulated data. In many cases inferences are furthermore based not on the data themselves, but on summary statistics of the data. Such statistics serve as data compression tools(Cover & Thomas, 2006) and, if used sensibly, enable computationally efficient inference from data sets, where the complexity of the data would stymie conventional likelihood-based methods (Pritchard *et al.*, 1999; Ratmann *et al.*, 2007).

ABC schemes have become increasingly popular, because of their flexibility and their deceptive conceptual simplicity. While especially some computationally demanding areas have fuelled the development of powerful ABC approaches, notably population genetics (Fagundes *et al.*, 2007), evolutionary biology (Wilkinson *et al.*, 2010), systems biology (Liepe *et al.*, 2010), dynamical systems theory (Toni *et al.*, 2009), and epidemiology (Blum & Tran, 2008), a worrying increase in naive (and plainly incorrect, see e.g. Walker *et al.* (2010)) applications are beginning to emerge. Such problems, as recent results by Didelot *et al.* (2010) and Robert *et al.* (2011) suggest, are more imminent in model selection applications. In the present context, all problems stem back to the issue of *sufficiency* of statistics, and its role in model selection. The present paper sets out to develop remedies for such problems. Below we will begin with an outline of the basic ideas underlying ABC, before discussing the particular challenges raised in particular by Robert and colleagues. We will then list the cases where ABC-based model selection is possible; in essence it is the ill-judged use of summary statistics and failure to ensure sufficiency which lies at the heart of the problem identified by Robert *et al.* (2011), before setting out methods that allow us to remedy these problems. We illustrate the use of these methods in a number of applications before concluding with some more general remarks on the conceptual and mathematical foundations of ABC approaches.

2 Approximate Bayesian Computation

2.1 Sufficient Statistics

Bayesian inference centres around the *posterior distribution*,

$$p(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{p(x)} \tag{1}$$

where x are the data, which are drawn from some sample space, $x \in \Omega \subseteq \mathbb{R}^D$, $f(x|\theta)$ is the *likelihood*, $\pi(\theta)$ the *prior distribution*, and θ an unknown parameter (Robert, 2007); $p(x)$ is often called the *evidence* (Mackay, 2003), but in many applications or discussions dismissed as a normalization constant. The *Likelihood principle* states that all the information about parameter θ is contained in the likelihood function $f(x|\theta)$, i.e. once we have the form of the likelihood, we do not have to retain any of the data. This principle is complemented by the *sufficiency principle*. Here a summary statistic of the general form

$$\mathcal{S} : \mathbb{R}^d \longrightarrow \mathbb{R}^w, \quad \mathcal{S}(x) = s \tag{2}$$

with $w \ll d$ typically, is called sufficient if the likelihood is independent of the parameter conditional on the value of the summary statistic. We denote by $f(x|\theta, s)$ the likelihood conditionally on the value of the summary statistic $\mathcal{S}(x) = s$ and $g(x|s)$ the density probability of the data given the summary statistic. The statistic is sufficient if and only if:

$$f(x|s, \theta) = g(x|s).$$

The likelihood can then generally be written in the Neyman-Fisher factorized form

$$f(x|\theta) = g(x|s)\tilde{f}(s|\theta), \tag{3}$$

where $s = \mathcal{S}(x)$ and $\tilde{f}(s|\theta)$ is the likelihood of the sufficient statistic (Cox, 2006). The function $g(\cdot)$ is independent of the parameter θ. Thus $\tilde{f}(s|\theta)$ carries all the information about the parameter.

This factorization is, however, not unique, as it depends on the sufficient statistic, which is generally not unique. For example, any statistic containing additional information in addition to a sufficient statistic is also sufficient. Therefore we typically seek to determine the minimally sufficient statistic, which is generally unique. As a consequence the functional forms (and values) of $g(\cdot)$ and $\tilde{f}(\cdot|\theta)$ depend on the choice of sufficient statistics.

In order to understand the terms in the factorization theorem, Eqn. (3), we focus on the case where X is a discrete random variable, we then have

$$g(x|s) = \Pr(X = x|\mathcal{S}(X) = s)$$

and

$$\tilde{f}(s|\theta) = \Pr(\mathcal{S}(X) = s|\theta).$$

Thus $g(x|s)$ is really the conditional probability of X given an observed value for the summary statistic, $\mathcal{S}(X) = s$ and it is therefore linked to the compression of the data achieved by the summary statistic. Here it is worth remembering that the complete data also form a valid summary statistic, and for this choice of statistic we have trivially, for all s, $g(x|s) = 1, \quad \forall x \in \Omega$.

2.2 ABC for parameter inference

In practical applications we are interested in evaluating the posterior distribution for model parameters, θ, defined Eqn. (1). When the likelihood is hard to evaluate it is still often possible to simulate from the model according to $f(\cdot|\theta)$. It is easy to show that for simulated data y we have

$$p(\theta|x) = \frac{\mathbb{1}(x = y)f(y|\theta)\pi(\theta)}{p(x)}, \tag{4}$$

which in many practical applications can be approximated using suitable distance functions, $\Delta(x, y)$, whence, after marginalization over simulated data we get,

$$p(\theta|x) \approx \int_{\Omega} \frac{\mathbb{1}(\Delta(x, y) \leq \epsilon)f(y|\theta)\pi(\theta)}{p(x)}dy. \tag{5}$$

This is obviously correct, as $\epsilon \longrightarrow 0$.

Based on the fact that sufficient statistics contain all the information about the θ that is contained in the data, we may be tempted to replace the data by the corresponding summary statistics. We thus replace the comparison of the data in Eqn. (5) by a comparison of the values of their respective summary statistics, using a distance function which, by abusing the notation, is denoted by $\Delta(\mathcal{S}(x), \mathcal{S}(y))$,

$$\begin{aligned} p(\theta|x) &\approx \int_{\Omega} \frac{\mathbb{1}(\Delta(\mathcal{S}(x), \mathcal{S}(y)) \leq \epsilon)f(y|\theta)\pi(\theta)}{p(x)}dy \\ &= \frac{\int_{\mathbb{R}^D} \mathbb{1}(\Delta(\mathcal{S}(x), s) \leq \epsilon)\tilde{f}(s|\theta)\pi(\theta)ds}{\int_{\Theta} \int_{\mathbb{R}^w} \mathbb{1}(\Delta(\mathcal{S}(x), s) \leq \epsilon)\tilde{f}(s|\theta)\pi(\theta)dsd\theta}, \end{aligned} \tag{6}$$

where we have made it explicit in the second line that once we use summary statistics we are only considering the second term on the right-hand side of Eqn. (3); any dependence on $g(\cdot)$ is lost, and so is therefore the effect of the data-compression in the summary statistic. Something similar is also implicit in conventional Bayesian inference. Cox (2006) reinforces this point by stating that "*Any Bayesian inference uses the data only via the minimal sufficient statistic. This is because the calculation of the posterior distribution involves multiplying the likelihood by the prior and normalizing. Any factor of the likelihood that is a function of y alone will disappear after normalization.*"

Quite generally, the choice of the summary statistic is important: without sufficiency the whole inference will only map the parameter regimes that will lead to model behaviour which embodies the constraints implied by specified summary statistic. Only if the summary statistic is sufficient, however, will we be able to infer the model parameters (Fearnhead & Prangle, 2010a). For some non-sufficient statistics, however, some aspects of the true posterior can be elucidated as shown in Figure 1. We will return to a discussion of non-sufficient statistics later.

2.3 ABC for model selection

One of the perhaps most useful (and aesthetically pleasing) aspects of the Bayesian inferential frameworks is that model selection is natural and intrinsic, especially compared to frequentist frameworks (Robert, 2007). From the earliest days ABC approaches for model selection have also been promoted, see e.g. Beaumont *et al.* (2002) and Fagundes *et al.* (2007). Recently, however, Robert *et al.* (2011) have issued a note of caution. This is based on the observation that a statistic, or set of statistics, which is sufficient for model parameters in different models, may still not be sufficient across models (Didelot *et al.*, 2010; Robert *et al.*, 2011).
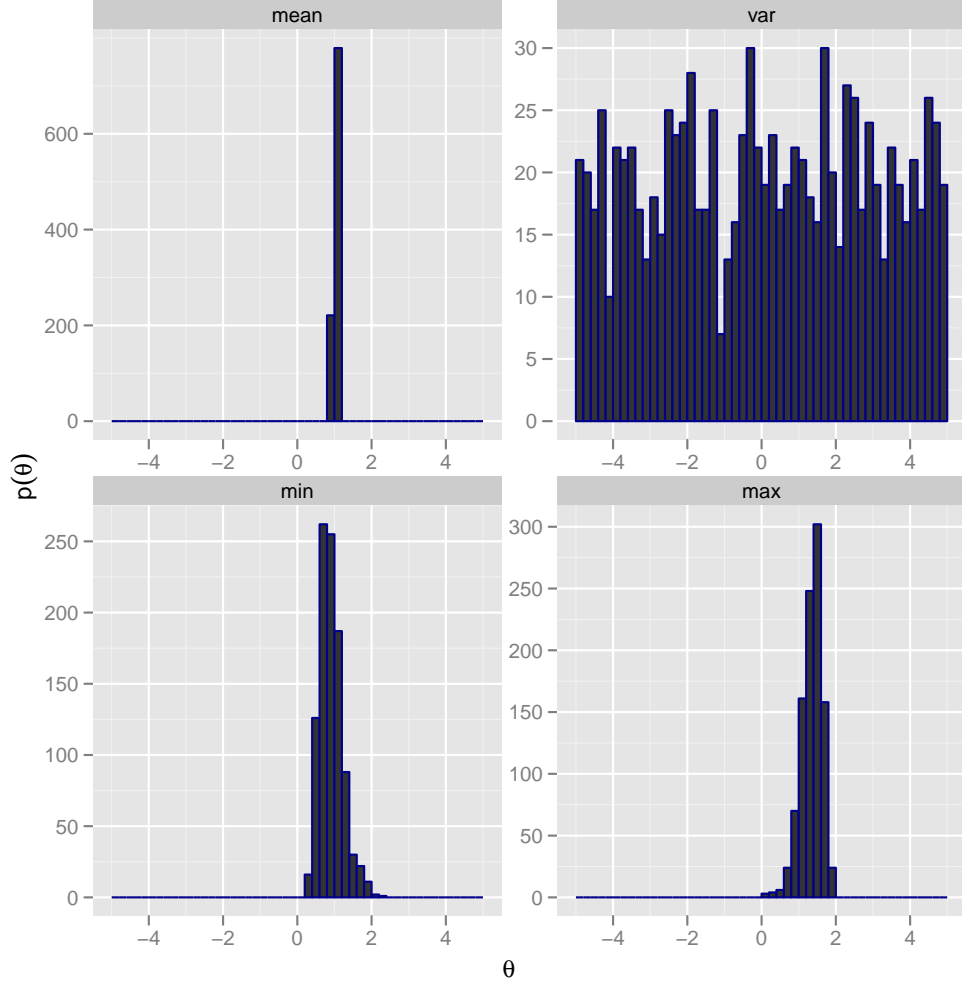
Figure 1: Parameter inference for the mean of a normal model with known standard deviation, $\sigma^2 = 1$, using the mean, variance, maximum and minimum as a statistic. We find that only the truly sufficient statistic ($\theta = \mu = 1$) yields the correct posterior distribution despite the fact that we generated 1,000 acceptances of samples of size 10,000 with $\epsilon = 0.001$.

To illustrate this point we now consider a finite set of models, $\mathcal{M} = \{M_1, \ldots, M_q\}$, each of which has an associated parameter vector $\theta_m \in \Theta_m$, $1 \leq m \leq q$. We aim to perform inference on the *joint space* over models and parameters, (m, θ_m). Robert *et al.* (2011) have focused on the Bayes Factors, but, of course, similar problems arise also for the marginal model likelihoods,

$$p(m|x) = \frac{\int_{\Theta_m} f(x|\theta_m)\pi(\theta_m)d\theta_m \pi(m)}{\sum_{i=1}^q \int_{\Theta_i} f(x|\theta_i)\pi(\theta_i)d\theta_i \pi(i)}. \tag{7}$$

Again, we can apply ABC by replacing evaluation of the likelihood in favour of comparing simulated and real data for different parameters drawn from the posterior, whence we obtain

$$p(m|x) \approx \frac{\int_{\Theta_m} \int_{\Omega} \mathbb{1}(\Delta(x,y) \leq \epsilon)f(y|\theta_m)\pi(\theta_m)d\theta_m dy \ \pi(m)}{\sum_{i=1}^q \int_{\Theta_i} \int_{\Omega} \mathbb{1}(\Delta(x,y) \leq \epsilon)f(y|\theta_i)\pi(\theta_i)d\theta_i dy \ \pi(i)}, \tag{8}$$

which is of course always exact once $\epsilon \longrightarrow 0$. The same is no longer true, however, once the complete data have been replaced by summary statistics. So in general

$$p(m|x) \neq \frac{\int_{\Theta_m} \int_{\mathbb{R}^w} \mathbb{1}(\Delta(\mathcal{S}_m(x), s_m \leq \epsilon)\tilde{f}(s_m|\theta_m)\pi(\theta_m)d\theta_m ds_m \ \pi(m)}{\sum_{i=1}^q \int_{\Theta_i} \int_{\mathbb{R}^w} \mathbb{1}(\Delta(\mathcal{S}_i(x), s_i) \leq \epsilon)\tilde{f}(s_i|\theta_i)\pi(\theta_i)d\theta_i ds_i \ \pi(i)}; , \tag{9}$$

where $\mathcal{S}_i$, $1 \leq i \leq q$ are the summary statistics for each model. An equality can only hold if the factors $g_i(x)$, $1 \leq i \leq q$ are all identical. Otherwise the different levels of data-compression are lost and unbiased model selection is no longer possible.

2.4 Resuscitating ABC Model Selection

As shown by Robert *et al.* (2011) and Didelot *et al.* (2010) even if we choose a set of statistics that is sufficient for parameter estimation across models, this does not guarantee that the same set of statistics are sufficient for model selection. R Robert *et al.* (2011) argue that therefore model selection, though not parameter estimation is fraught with problems in an ABC framework. Here we will argue that this is not the case. While we do agree that sufficiency and problems when using inadequate (or *insufficient*) statistics for model selection, we maintain that

- this mirrors problems that can also be observed in the parameter estimation context (see Figure 1),

- for many important, and arguably the most important applications of ABC, this problem can in principle be avoided by using the whole data rather than summary statistics,

- in cases where summary statistics are required, we argue that we can construct approximately sufficient statistics in a disciplined manner,

- when all else fails, a change in perspective, allows us to nevertheless make use of the flexibility of the ABC framework (see Discussion).

The bulk of this article will deal with the derivation and application of a method that constructs summary statistics appropriate for the twin use of parameter estimation and model selection. After this we will briefly return to an outline of alternative approaches and map the applicability of ABC-based model selection more generally.

3   Some Basic Concepts from Information Theory

Our method is most easily framed in the terms of information theory (Mackay, 2003; Cover & Thomas, 2006; Mézard & Montanari, 2009), and in order to keep this article self-contained we briefly review some of the basic concepts. We let X denote a discrete random variable with potential states $\mathcal{X}$ and probability mass function $p_X(x) = \Pr(X = x)$, $x \in \mathcal{X}$. In this section, for sake of clarity, we explicitly denote the random variables as subscript of the corresponding probability density.

3.1 Entropy and Mutual Information

The *entropy* of X, denoted by H, measures the uncertainty of X and is defined as follows,

$$H(X) = -\sum_x p_X(x) \log p_X(x) = -E_X \left[\log p_X(X) \right] = E_X \left[\frac{1}{\log p_X(X)} \right] \geq 0,$$

where E_X denotes the expectation under the probability mass function p_X. Let (X, Y) be a pair of discrete random variables with joint distribution $p_{X,Y}$. The *conditional entropy* $H(Y|X)$ is defined as

$$H(Y|X) = -E_{X,Y} \left[\log p_{Y|X}(Y|X) \right].$$

The *mutual information* $I(X; Y)$ between two discrete random variables X and Y measures the amount of information that Y contains about X. It can be seen as the reduction of the uncertainty about X due to the knowledge of Y,

$$I(X; Y) = H(X) - H(X|Y) = \sum_{x,y \in \mathcal{X}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} = KL(p_{X,Y} || p_X p_Y) \geq 0,$$

where $KL(P||Q)$ refers to the *Kullback-Leibler* (KL) divergence between probabilities P and Q. The mutual information $I(X; Y)$ is equal to 0 if and only if the random variables X and Y are independent.

The *conditional mutual information* of discrete random variables X, Y and Z is defined as

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z);$$

it is the reduction in uncertainty of X due to knowledge of Y when Z is given. This quantity is zero if and only if X and Y are conditionally independent given Z, which means that Z contains all the information about X in Y. The conditional mutual information satisfies the chain rule: for discrete random variables $X_1, X_2, \cdots, X_n$ and Y we have

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^n I(X_i; Y|X_1, X_2, \cdots, X_{i-2}, X_{i-1}).$$

In the following we only consider entropy, rather than the differential entropy, which applies to continuous random variables.

3.2 Data Processing Inequality and Sufficient Statistics

The *data processing inequality* (DPE) states that for random variables X, Y, and Z such that $X \to Y \to Z$, (i.e. Y depends, deterministically or randomly, on X and Z depends on Y)

$$I(X; Y) \geq I(X; Z),$$

with equality only if $X \to Y \to Z$ forms a Markov Chain, which means that the random variables X and Z are conditionally independent given Y: $p_{X,Z|Y} = p_{X|Y}p_{Z|Y}$.

Now consider a family of distributions $\{f(\cdot|\theta)\}_{\theta \in \Theta}$ and let X be a sample from a distribution in this family. Let $\mathcal{S}$ be a deterministic statistic and denote by S the random variable such that $S = \mathcal{S}(X)$. Therefore $\theta \to X \to S$. By the DPE

$$I(\theta; S) \leq I(\theta; X).$$

Analogously to the discussion above, a statistic $\mathcal{S}$ is said *sufficient for parameter* θ if and only if S contains all the information in X about θ, i.e.

$$I(\theta; S) = I(\theta; X) \quad \text{where } S = \mathcal{S}(X)$$

Equivalently we may write:

Result 1. *$\mathcal{S}$ is a sufficient statistic for parameter θ if and only if*

$$I(\theta; X|S) = 0.$$

In that case,

$$E_{\theta,X} \left[\log \frac{p(\theta|X)}{p(\theta|S)} \right] = 0.$$

From now on, we resume the use of typically bayesian notation (see section 2) where the random variables are no longer given as subscripts but are unambiguously inferred from context. For example, the density of probability $p(\theta|S)$ involved in the theorem designates both the posterior probability of the parameter θ conditional on S and its value when the parameter is actually equal to θ.

Proof. By definition, $\mathcal{S}$ is a sufficient statistic if and only if $I(\theta; X) = I(\theta; S)$. $\mathcal{S}$ being a deterministic fonction of X, the previous equation is equivalent to

$$I(\theta; X, S) - I(\theta; S) = 0$$
$$\Leftrightarrow \quad H(\theta) - H(\theta|X,S) - H(\theta) + H(\theta|S) = 0$$
$$\Leftrightarrow \quad H(\theta|S) - H(\theta|X,S) = 0$$
$$\Leftrightarrow \quad I(\theta; X|S) = 0$$

If $\mathcal{S}$ is a sufficient statistic for parameter θ then

$$I(\theta; X) = I(\theta; S)$$
$$\Leftrightarrow \quad \sum_{\theta, x} p(\theta, x) \log \frac{p(\theta, x)}{p(\theta)p(x)} = \sum_{\theta, s} p(\theta, s) \log \frac{p(\theta, s)}{p(\theta)p(s)}$$
$$\Leftrightarrow \quad \sum_{\theta, x} p(\theta, x) \log \frac{p(\theta, x)}{p(\theta)p(x)} = \sum_{\theta, x} p(\theta, x) \log \frac{p(\theta, \mathcal{S}(x))}{p(\theta)p(\mathcal{S}(x))}$$
$$\Leftrightarrow \quad \sum_{\theta, x} p(\theta, x) \log \frac{p(\theta, x)p(\theta)p(\mathcal{S}(x))}{p(\theta)p(x)p(\theta, \mathcal{S}(x))} = 0$$
$$\Leftrightarrow \quad \sum_{\theta, x} p(\theta, x) \log \frac{p(\theta|x)}{p(\theta|\mathcal{S}(x))} = 0$$

$\square$

4 Constructing Sufficient Statistics

We consider the following situation: suppose that we have a finite set of summary statistics $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_w\}$ and assume that $\mathcal{S}$ is a sufficient statistic. We aim to identify a subset $\mathcal{U}$ of $\mathcal{S}$ which is sufficient for θ. The following result characterizes such a subset.

Result 2. *Let $\mathcal{S}$ be a finite set of summary statistics of X, and assume that $\mathcal{S}$ is a sufficient statistic. Denote by $\mathcal{U}$ a subset of $\mathcal{S}$. For a random variable X distributed according to a distribution parametrized by θ, let $U = \mathcal{U}(X)$ and $S = \mathcal{S}(X)$. The following statements hold*

$$\mathcal{U} \text{ is a sufficient statistic}$$
$$\Leftrightarrow \quad I(\theta; S|U) = 0$$
$$\Leftrightarrow \quad E_X\left[KL(p(\theta|S)||p(\theta|U))\right] = 0 \ .$$

Proof. By definition of the conditional mutual information

$$I(\theta; S|U) = H(\theta|U) - H(\theta|S,U) = H(\theta) - H(\theta|S,U) - H(\theta) + H(\theta|U)$$
$$= I(\theta; U, S) - I(\theta; U) = I(\theta; S) - I(\theta; U)$$

since U is a vector composed of elements of S. The statistic $\mathcal{S}$ being sufficient, $I(\theta; S) = I(\theta; X)$, and therefore $I(\theta; S|U) = 0$ if and only if $\mathcal{U}$ is a sufficient statistic. Denote by h the function such that for all x, $u = \mathcal{U}(x) = $

$h(S(x))$,

$$I(\theta; S|U) = H(\theta|U) - H(\theta|S, U) = \sum_{\theta, s, u} p(\theta, s, u) \log\left(\frac{p(\theta|u, s)}{p(\theta|u)}\right)$$

$$= \sum_{\theta, s} p(\theta, s) \log\left(\frac{p(\theta|h(s), s)}{p(\theta|h(s))}\right) = \sum_{\theta, s} p(\theta, s) \log\left(\frac{p(\theta|s)}{p(\theta|h(s))}\right)$$

$$= \sum_{s} p(s) \sum_{\theta} p(\theta|s) \log\left(\frac{p(\theta|s)}{p(\theta|h(s))}\right) = \sum_{s} p(s) KL\left(p(\theta|s)||p(\theta|h(s))\right)$$

$$= \sum_{x} p(x) KL\left(p(\theta|\mathcal{S}(x))||p(\theta|\mathcal{U}(x))\right) = E_X\left[KL\left(p(\theta|S)||p(\theta|U)\right)\right] .$$

$\square$

According to this result identifying a sufficient statistic with minimum cardinality from a sufficient family $\mathcal{S}$ of summary statistics boils down to identifying the smallest subset $\mathcal{U}$ of $\mathcal{S}$ such that $I(\theta; S|U) = 0$ where $U = \mathcal{U}(X)$ and $S = \mathcal{S}(X)$ or, equivalently, $E_X[KL(p(\theta|S)||p(\theta|U))] = 0$. Here we aim to determine a sufficient statistic for Approximate Bayesian Computation (ABC) methods for parameter inference and then for model selection. We focus in this section on the parameter inference task. In the ABC framework, the expectation over the data of the Kullback-Leibler divergence (Cover & Thomas, 2006) between the two posterior distributions $p(\theta|S)$ and $p(\theta|U)$ cannot be exactly computed since we only have at our disposal a dataset x and the value of the statistics for this dataset $s^* = (s_1^*, \ldots, s_w^*) = \mathcal{S}(x)$. Thus, we approximate it by the expectation with respect to the empirical measure of the data. The method is summarized in Algorithm 1. We denote by $|U|$ the cardinality of a set U.

Algorithm 1 Minimization of the mutual information

1: **input:** a sufficient set of statistics whose values on the dataset is $s^* = \{s_1^*, \ldots, s_w^*\}$
2: **output:** a subset U^* of s^*
3: **for** all $u^* \subset s^*$ **do**
4:     perform ABC to obtain $\hat{p}(\theta|u^*)$
5: **end for**
6: let $T^* = \{u^* \subset s^*$ such that $KL(\hat{p}(\theta|s^*)||\hat{p}(\theta|u^*)) = 0\}$
7: **return** $U^* = \text{argmin}_{u^* \in T^*} |u^*|$

This methodology is computationally prohibitive since it enumerates all possible subsets u^* of s^* and perform the ABC algorithm for all of these. Moreover, it is challenging to obtain a precise estimate of the posterior $\hat{p}(\theta|u^*)$ of θ given a value of the statistics $U = u^*$ when the cardinality of u^* is large. In particular, it is often impossible to obtain an estimate of $p(\theta|s^*)$. It is then necessary to design an algorithm which does not need the computation of this probability. The following result provides a first step into this direction.

Result 3. *Let X be a random variable generated according to $f(\cdot|\theta)$. Let $\mathcal{S}$ be a sufficient statistic and $\mathcal{U}$ and $\mathcal{T}$ two subsets of $\mathcal{S}$ such that $U = \mathcal{U}(X)$, $T = \mathcal{T}(X)$ and $S = \mathcal{S}(X)$ satisfy $U \subset T \subset S$. We have*

$$I(\theta; S|T) = I(\theta; S|U) - I(\theta; T|U) .$$

Proof. For all subset T of S,

$$I(\theta; S|T) = H(\theta|T) - H(\theta|S, T) = H(\theta|T) - H(\theta|S) .$$

Then, for U and T subsets of S,

$$I(\theta; S|T) - I(\theta; S|U) = H(\theta|T) - H(\theta|U) .$$

If U in included in T, then $H(\theta|T) = H(\theta|T, U)$ and

$$I(\theta; S|T) - I(\theta; S|U) = -I(\theta; T|U)$$

which proves the result. $\square$

It follows that the information contained in S on θ given T is smaller than the information contained in S on θ given $U \subset T$. In order to construct a subset $\mathcal{T}$ of $\mathcal{S}$ such that $I(\theta; S|T) = 0$, where $S = \mathcal{S}(X)$ and $T = \mathcal{T}(X)$, it is thus sufficient to add one by one statistics of $\mathcal{S}$ until the condition holds. Indeed, if we denote the statistic added at time step $k \leq w$ by $\mathcal{S}_{(k)}$ and $S_{(k)} = \mathcal{S}_{(k)}(X)$, then

$$I(\theta; S|S_{(1)}, \ldots, S_{(k)}) \leq I(\theta; S|S_{(1)}, \ldots, S_{(k+1)}) .$$

And there exists an integer $k \leq w$ such that $I(\theta; S|S_{(1)}, \ldots, S_{(k)}) = 0$. According to result 3, at each time k,

$$I(\theta; S|S_{(1)}, \ldots, S_{(k)}) = I(\theta; S|S_{(1)}, \ldots, S_{(k-1)}) - I(\theta; S_{(1)}, \ldots, S_{(k)}|S_{(1)}, \ldots, S_{(k-1)}) .$$

The mutual information is a non negative function, then, in order to decrease as much as possible $I(\theta; S|S_{(1)}, \ldots, S_{(k)})$, the added statistic at time $k > 1$ should be such that

$$S_{(k)} = \mathrm{argmax}_{V \in S \backslash \{S_{(1)}, \ldots, S_{(k-1)}\}} I(\theta; S_{(1)}, \ldots, S_{(k-1)}, V|S_{(1)}, \ldots, S_{(k-1)})$$
$$= \mathrm{argmax}_{V \in S \backslash \{S_{(1)}, \ldots, S_{(k-1)}\}} E_X \left[KL \left(p(\theta|S_{(1)}, \ldots, S_{(k-1)}, V)||p(\theta|S_{(1)}, \ldots, S_{(k-1)}) \right) \right] .$$

As previously mentioned, in practice, this expectation can not be computed. We hence replace it by the expectation with respect to the empirical measure of the data, leading to the approximation, for all $1 \leq k \leq n$,

$$\mathrm{argmax}_{V \in S \backslash \{S_{(1)}, \ldots, S_{(k-1)}\}} E_{p(X)} \left[KL \left(p(\theta|S_{(1)}, \ldots, S_{(k-1)}, V)||p(\theta|S_{(1)}, \ldots, S_{(k-1)}) \right) \right]$$
$$\approx \mathrm{argmax}_{v^* \in s^* \backslash \{s^*_{(1)}, \ldots, s^*_{(k-1)}\}} KL \left(p(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)}, v^*)||p(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)}) \right) ,$$

where s^* is the value of the statistic $\mathcal{S}$ on the dataset. In addition, the first statistic $\mathcal{S}_{(1)}$ should contain the maximum information about θ, in the sense that

$$S_{(1)} = \mathrm{argmax}_{V \in S} I(\theta; V) = \mathrm{argmin}_{V \in S} H(\theta|V) = \mathrm{argmax}_{V \in S} E_{p(\theta, V)} \left[\log p(\theta|V) \right] .$$

This results in algorithm 2.

Algorithm 2 Greedy minimization of the mutual information

1: **input:** a sufficient set of deterministic statistics whose values on the dataset is $s^* = \{s^*_1, \ldots, s^*_w\}$
2: **output:** a subset U^* of s^*
3: for all $u^* \in s^*$, perform ABC to obtain $\hat{p}(\theta|u^*)$
4: let $s^*_{(1)} = \mathrm{argmax}_{u^* \in s^*} \log \hat{p}(\theta|u^*)$
5: **for** $k \in \{2, \ldots, w\}$ **do**
6: for all $u^* \in s^* \backslash \left\{ s^*_{(1)}, \ldots, s^*_{(k-1)} \right\}$, perform ABC to obtain $\hat{p}(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)}, u^*)$
7: let

$$s^*_{(k)} = \mathrm{argmax}_{u^*} KL(\hat{p}(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)}, u^*)||\hat{p}(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)})) \qquad (10)$$

8: **if** $KL(\hat{p}(\theta|s^*_{(1)}, \ldots, s^*_{(k)})||\hat{p}(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)})) \leq \epsilon$ **then**
9:   **return** $U^* = (s^*_{(1)}, \ldots, s^*_{(k-1)}))$
10: **end if**
11: **end for**
12: **return** $U^* = s^*$

This algorithm is computationally expensive if the cardinality of the set of statistics $|\mathcal{S}| = w$ is large. Indeed at each iteration k, one performs the ABC algorithm $w - k + 1$ times. A simplification of it consists in replacing the maximization step (see equation (10)) by testing randomly various statistics and choosing a statistic $\mathcal{S}_{(k)}$ such that $I(\theta; S_{(1)}, \ldots, S_{(k)}|S_{(1)}, \ldots, S_{(k-1)})$ is large. Different criteria may be used to determine if the mutual information is large or not, and then decide if the statistic should be included or not. Most of these criteria consist of determining if the posterior probability of θ given $S_{(1)}, \ldots, S_{(k-1)}$ and the posterior probability of θ given $S_{(1)}, \ldots, S_{(k)}$ are significantly different. If so, adding the statistic $\mathcal{S}_{(k)}$ is justified, otherwise we do not add it and instead turn to a different statistic. In the algorithm 3, we denote by $C \left[p(\theta|s^*_{(1)}, \ldots, s^*_{(k)}), p(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)}) \right]$ the criterion which is equal to 1 if the statistic $\mathcal{S}_{(k)}$ should be added and 0 otherwise. In practice, we can use, for instance, one of the following measures:

- we add the statistic $\mathcal{S}_{(k)}$ if $KL(p(\theta|s^*_{(1)}, \ldots, s^*_{(k)})||p(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)})) \geq \delta_k$ where δ_k is a threshold which may be computed by bootstrapping the data to estimate $E_{p(X)}\left[p(\theta|S_{(1)}, \ldots, S_{(k-1)})\right]$:

$$\delta_k = KL\left(p(\theta|s^*_{(1)}, \ldots, s^*_{(k)})||E_{p(X)}\left[p(\theta|S_{(1)}, \ldots, S_{(k-1)})\right]\right),$$

- the Kolmogorov-Smirnov test enables us to compare $p(\theta|s^*_{(1)}, \ldots, s^*_{(k-1)})$ and $p(\theta|s^*_{(1)}, \ldots, s^*_{(k)})$ and so, the statistic $\mathcal{S}_{(k)}$ is added if the test has a p-value smaller than a certain threshold, say 0.01.

Algorithm 3 Stochastic minimization of the mutual information

1: **input:** a sufficient set of deterministic statistics whose values on the dataset is $s^* = \left\{s^*_1, \ldots, s^*_w\right\}$
2: **output:** a subset V^* of s^*
3: choose randomly u^* in s^*
4: $T^* \leftarrow s^* \backslash \{u^*\}$
5: $V^* \leftarrow u^*$
6: **repeat**
7:     **repeat**
8:         **if** $T^* = \varnothing$ **then return** V^*
9:         **end if**
10:         choose randomly u^* in T^*
11:         $T^* \leftarrow T^* \backslash u^*$
12:         perform ABC to obtain $\hat{p}(\theta|V^*, u^*)$
13:     **until** $\mathcal{C}\left[\hat{p}(\theta|V^*, u^*), \hat{p}(\theta|V^*)\right] = 1$
14:     $T^* \leftarrow s^* \backslash \{V^*, u^*\}$
15:     *optionally:* $V^* \leftarrow OrderDependency(V^*, u^*)$ and $T^* \leftarrow s^* \backslash V^*$
16:     $V^* \leftarrow V^* \cup u^*$
17: **until** $T^* = \varnothing$
18: **return** V^*

The order in which the statistics are added matters (Joyce & Marjoram, 2008; Nunes & Balding, 2010) and the only way to avoid this inconvenience is to use the computationally expensive algorithm 2. Nevertheless, we suggest to test for order dependency before deciding on whether to add a statistic. This consists of determining if, given the recently added statistic u^*, the previously added statistics in V^* still bring relevant information or are not necessary anymore and hence may be released from the set under construction. More precisely, after line 15 of the algorithm 3, we add the function described in algorithm 4.

Algorithm 4 Order Dependency

1: **Input:** A set of accepted statistics $V^* = \left\{s^*_{(1)}, \ldots, s^*_{(k-1)}\right\}$ and the last accepted statistic u^*
2: **Output:** A subset U^* of $\{V^* \cup u^*\}$
3: $U^* \leftarrow u^*$
4: **for** $i \in \{1, \ldots, k-1\}$ **do**
5:     **if** $\mathcal{C}(p(\theta|U^*, s^*_{(i)}), p(\theta|U^*)) = 1$ **then**
6:         $U^* \leftarrow U^* \cup s^*_{(i)}$
7:     **end if**
8: **end for**
9: **return** U^*

---

5   Relation to previous work

The presented information theoretic framework builds upon two previous methods for summary statistic selection. Our main contributions are the generalisation of the notion of approximate sufficiency, rigorous derivations

of algorithms using information theory and the application of summary statistic selection for the joint space.

Joyce & Marjoram (2008) developed a notion of approximate sufficiency for parameter inference and presented a sequential algorithm to score statistics according to whether their inclusion would improve the inference. Their sequential algorithm resembles (and indeed inspired) Algorithm 3 although they do not retain statistics once they have failed to be added which we feel is required since whether a statistic is added depends strongly on the statistics already accepted. Their rule for adding statistics is essentially an approximate test for independence on the posterior distribution under the addition of a new statistic but can only be used for single parameter models. We have shown that the true stopping criterion should be the change in KL divergence which can be used for multivariate posteriors, although tests for independence (KS, χ^2) can be used as an approximation in single parameter models.

Nunes & Balding (2010) proposed a heuristic algorithm to minimize the entropy of the posterior with respect to sets of summary statistics . Additionally they proposed a refinement step where the set of statistics that minimised the posterior mean root sum of squared errors (MRSSE) was selected. The minimum entropy approach is related to Algorithm 1 since, when $H(\theta)$ is constant, minimising the entropy maximises the mutual information. However, assuming there exists a sufficient statistic, choosing the set of statistics that minimises the entropy is guaranteed to give sufficiency but not *minimal* sufficiency since adding a statistic to a sufficient set can reduce the entropy by chance (a manifestation of "conditioning always reduces entropy").

6   Automated selection of summary statistics for model selection

As pointed out recently sufficiency across models is still not sufficient to perform reliably model choice in the ABC framework. Here we show how a natural extension of the methodology introduced above can also be employed in order to construct sets of statistics that are sufficient for ABC-based model selection, when it is impractical to use the raw data (Toni & Stumpf, 2010).

Consider q models, each with an associated set of parameters $\Theta_i, i \in \{1, ..., q\}$. We aim to identify a set of sufficient statistic for model selection. Let M being a random variable taking value in $\{1, ..., q\}$. A statistic is sufficient for model selection if and only if it is sufficient for the joint space $\{M, \{\theta_i\}_{1 \le i \le q}\}$. According to result 1, this means that a summary statistic $\mathcal{S}$ is sufficient for model selection if and only if $I(M, \theta_1, ..., \theta_q; X|S) = 0$ where $S = \mathcal{S}(X)$ and X is a sample from a distribution in the family $\{f(\cdot|\theta_i)\}_{\theta_i \in \Theta_i, 1 \le i \le q}$. The following result enables us to link this condition with the sufficiency for parameter inference for each model.

Result 4. *For all deterministic statistic $\mathcal{S}$,*

$$I(M, \theta_1, ..., \theta_q; X|S) = I(M; X|\theta_1, ..., \theta_q, S) + \sum_i I(\theta_i; X|S) \,,$$

where $S = \mathcal{S}(X)$ and X is a sample from a distribution in the family $\{f(\cdot|\theta_i)\}_{\theta_i \in \Theta_i, 1 \le i \le q}$.

Proof. From the chain rule of mutual information we have that

$$I(M, \theta_1, ..., \theta_q; X|S) = I(M; X|\theta_1, ..., \theta_q, S) + \sum_i I(\theta_i; X|\theta_1, \cdots, \theta_{i-1}, S) \,.$$

The result then follows from the fact that, for all i, $I(\theta_i; X|\theta_1, \cdots, \theta_{i-1}, S) = I(\theta_i; X|S)$. □

The mutual information being always non negative, this shows that a statistic $\mathcal{S}$ is sufficient for model selection if and only if $I(M; X|\theta_1, ..., \theta_q, S) = 0$ and $I(\theta_i; X|S) = 0$ for all $i \in \{1, \cdots, q\}$. Therefore, a sufficient statistic for model selection is sufficient for parameter inference in each model and, given the parameter values for every models, the statistic is sufficient for inferring the model. Thus, in order to identify a sufficient statistic for model selection, one should determine a set of minimal sufficient statistics $\mathcal{S}^{m_i}$ for each model $1 \le i \le q$ and then identify among all statistics $\mathcal{T}$ containing $\cup_{i \in \{1...q\}} \mathcal{S}^{m_i}$ the one with the smallest cardinality such that $I(M; X|\theta_1, ..., \theta_q, T) = 0$. The method, summarized in algorithm 5, consists in running one of the previous algorithms, for example algorithm 3 and then add statistics which bring new information about the models in the sense that the posterior probability of the model conditional on the statistics varies significantly if we add a this new statistic.

Similarly to algorithm 3, it is possible to test for order dependency before deciding to add a statistic. To do so, we apply the algorithm 4 in which at step 3, the set U^* is initialized by $M^* \cup u^*$ such that we always keep the set M^* and condition line 5 is replaced by $\mathcal{C}\left[\hat{p}(M|\theta_1, ..., \theta_q, U^*, s^*_{(i)}), \hat{p}(M|\theta_1, ..., \theta_q, U^*)\right] = 1$.

Algorithm 5 Stochastic minimization of the mutual information for model selection
---
1: **input:** a sufficient set of deterministic statistics whose values on the dataset is $s^* = \{s_1^*, \ldots, s_w^*\}$
2: **output:** a subset V^* of s^* which is sufficient for model selection
3: **for** $i \in \{1, \ldots, q\}$ **do**
4:     determine a sufficient statistic whose value on the dataset is $S^{m_i} \subset s^*$ using algorithm 3
5: **end for**
6: Let $M^* = \cup_{1 \leq i \leq q} S^{m_i}$
7: Let $W^* \leftarrow s^* \backslash M^*$
8: choose randomly u^* in W^*
9: $T^* \leftarrow W^* \backslash \{u^*\}$
10: $V^* \leftarrow u^*$
11: **repeat**
12:     **repeat**
13:         **if** $T^* = \emptyset$ **then return** V^*
14:         **end if**
15:         choose randomly u^* in T^*
16:         $T^* \leftarrow T^* \backslash u^*$
17:         perform ABC to obtain $\hat{p}(M|\theta_1, \ldots, \theta_q, M^*, V^*, u^*)$
18:     **until** $\mathcal{C}\left[\hat{p}(M|\theta_1, \ldots, \theta_q, M^*, V^*, u^*), \hat{p}(M|\theta_1, \ldots, \theta_q, M^*, V^*)\right] = 1$
19:     $T^* \leftarrow W^* \backslash \{V^*, u^*\}$
20:     *optionally:* $V^* \leftarrow OrderDependenceModelSelec\,(V^*, M^* \cup u^*)$ and $T^* \leftarrow W^* \backslash V^*$
21:     $V^* \leftarrow V^* \cup u^*$
22: **until** $T^* = \emptyset$
23: **return** V^*
---

7 Applications

We illustrate the framework developed above in three different contexts. First we consider a simple model selection problem involving two normal distributions. We then consider a typical population genetics example on three demographic scenarios for simulated data, before finally turning to a problem where we consider alternative random walk models; this last example should be typical for applications where likelihood-based inferences are out of the question due to the complexity of the models and the data.

7.1 Normal example

The developed framework was illustrated on a simple example with two models

$$y_1, \ldots y_d \sim \mathcal{N}(\mu, \sigma_1^2) \quad \text{and} \quad y_1, \ldots y_d \sim \mathcal{N}(\mu, \sigma_2^2),$$

where the variances of the normal distributions, σ_1 and σ_2, are fixed. Under a conjugate prior, $\mu \sim \mathcal{N}(0, a^2)$, the true Bayes factor is given by

$$BF(\mathbf{y}) = \frac{\sigma_1^{-d} \exp\{-S^2/2\sigma_1^2\} \exp\{-\bar{y}^2/2(a^2 + \sigma_1^2/d)\}\sqrt{a^{-2} + d\sigma_2^{-2}}}{\sigma_2^{-d} \exp\{-S^2/2\sigma_2^2\} \exp\{-\bar{y}^2/2(a^2 + \sigma_2^2/d)\}\sqrt{a^{-2} + d\sigma_1^{-2}}},$$

where

$$\bar{y} = d^{-1} \sum_{i=1}^{d} y_i \quad \text{and} \quad S^2 = \sum_{i=1}^{d} (y_i - \bar{y})^2.$$

In this case $\bar{y}$ is sufficient for parameter inference but the pair $\{\bar{y}, S^2\}$ is sufficient for the joint space.

To test the automated choice of approximate summary statistics, 100 data sets were sampled under model 1 and the algorithm run to select statistics for parameter inference and for the joint space from a pool of 5

statistics including $\bar{y}$, S^2, range, maximum and a non informative statistic $u \sim \mathcal{U}(0,2)$. The values of the parameters were chosen to be $\sigma_1 = 0.3$, $\sigma_2 = 0.6$, $n = 15$ and $a = 2$. Since there are difficulties that arise from the possibly very different scales of the summary statistics the distance was defined as

$$\Delta(\mathcal{S}(x), \mathcal{S}(y)) = \sum_i [\log \mathcal{S}_i(x) - \log \mathcal{S}_i(y)]^2,$$

which accounts for the relative difference between data and simulation and avoids the need to known the scales of the statistics *a priori*. Here, $\mathcal{S}_i(x)$ denotes the i-th component of $\mathcal{S}(x) \in \mathbb{R}^w$ The stopping criteria were defined through tests for independence ($p < 1 \times 10^{-5}$) between the posterior distributions under different summary statistics; a Kolmogorov-Smirnov test in the case of the continuous parameter posterior and a Pearson test in the case of the discrete model posterior. The ABC was run with 500 particles with $\epsilon = 0.1$.

Figure 2 shows the results of the summary statistic selection over the 100 runs. Figure 2 (left) shows the results for parameter inference across the two models. The mean was selected in every replicate, the maximum in 14 replicates and S^2 once. Figure 2 (right) shows the additional statistics selected for the joint space. S^2 was selected in 84 cases, the range in 19 cases, the maximum in 9 cases and the noise statistic once. Figure 3 shows the Bayes Factor obtained via ABC versus the analytical prediction. As expected the Bayes factor calculated using only the statistics selected for parameter inference is uncorrelated with the true Bayes factor (figure on the left). When the statistics sufficient for the joint space are included the ABC Bayes factor correlates well with the analytical prediction.

7.2   Population genetics example

To further demonstrate the efficacy of our methodology we applied the summary statistic selection procedure to a real world ABC problem, that of model selection in population genetics. Data were generated using coalescent simulations (Hudson, 1991) from three competing models, producing 100 data sets from each for fixed parameters using a modified version of the MS software package downloadable from `http://home.uchicago.edu/rhudson1/source/mksamples.html`).

The models considered were:

Model 1 Constant population size (with population mutation rate $\theta = 20$, corresponding to a 20,000bp stretch of DNA in a population of size $N = 1,000,000$.

Model 2 Exponential growth model with exponential growth rate $\gamma = 0.4$ and all other parameters as above.

Model 3 Two island model with scaled migration rate $m = 10$ and all other parameters as above.

To perform ABC we generated $5,000,000$ samples of datasets comprising 100 chromosomes for each model; in each case the population mutation rate θ (Ewens, 2004) was drawn from the prior $\mathcal{U}(5,30)$; the real value for which the data were generated was $\theta = 20$ (measured in units of total population size). The summary statistics summarised below were then used in our ABC summary statistic selection framework to derive sufficient sets of summary statistics for model selection on the observed data. The summary statistics calculated were:

S1 Number of Segregating Sites, N_S.

S2 Number of Distinct Haplotypes, N_H.

S3 Homozygosity, h_H, where h is the probability that two haplotypes are identical,

$$h_H = \sum_{h=1}^{N_H} \nu_h^2.$$

S4 Average SNP Homozygosity,

$$\bar{h}_S = \sum_{i=1}^{N_S} (\nu_0(i)^2 + \nu_1(i)^2).$$

S5 Number of occurences of most common haplotype, f_H.

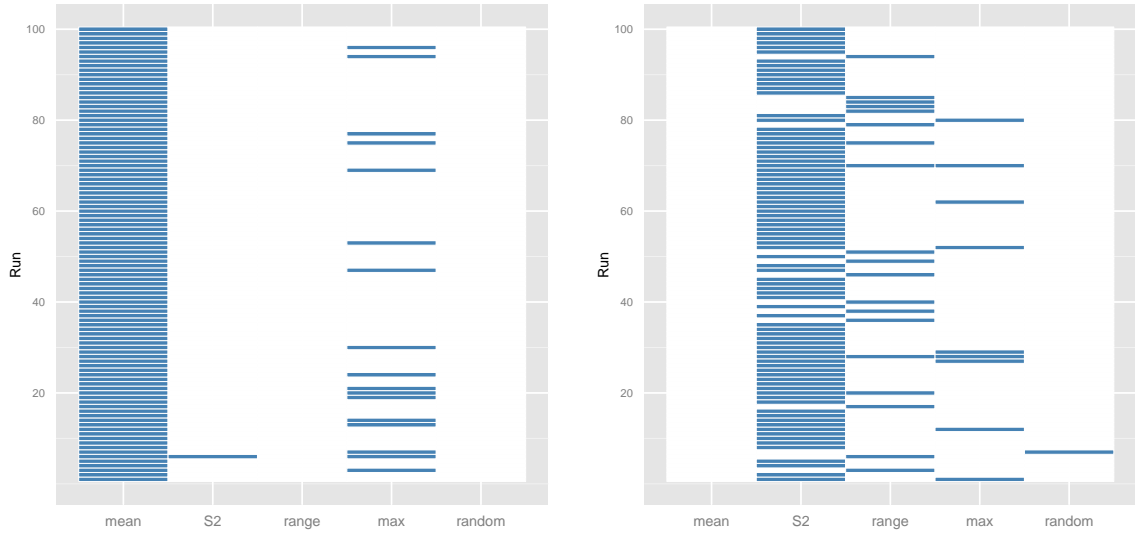**S6** Mean number of pair-wise differences between haplotypes, T.

Figure 2: Summary statistics selected in 100 runs of the automated summary statistic selection. Left: Statistics selected for parameter inference (the union of statistics found under model 1 and model 2). Right: Additional statistics chosen for the joint space.
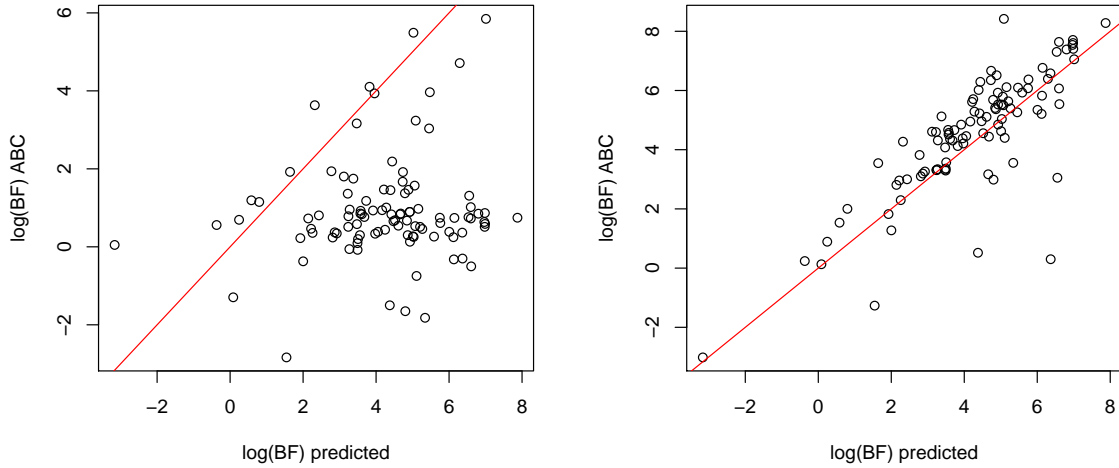

Figure 3: Predicted vs approximated log Bayes Factor for the normal toy model. Left: The case for sufficient statistics selected for parameter inference. Right: The case for sufficient statistics selected for the joint space. In both cases the red line represents the line $y = x$.

S7 Number of Singleton Haplotypes, f_{sH}

S8 Number of Singleton SNPs, f_{sS}.

S9 Linkage Disequilibrium measured by

$$\overline{r^2} = \frac{2}{N_S(N_S - 1)} \sum_{i=1}^{N_S-1} \sum_{j=i+1}^{N_S} \frac{(\nu_{00}(i,j) - \nu_0(i)\nu_0(j))^2}{\nu_0(i)\nu_1(i)\nu_0(j)\nu_1(j)}$$

S10 Fraction of pairs of loci which violate the four-gamete test, i.e. for which the two-locus haplotypes 00, 01, 10 and 11 exist.

S11 Random variable, $\rho \sim \mathcal{U}_{[0,1]}$.

where N is the number of sequences in the data, and for each SNP locus, i, let $\nu_0(i)$ $\nu_1(i)$ denote the frequencies of the ancestral and derived alleles; further for any haplotype h, ν_h is the corresponding frequency.

The results of the selection process when performed over 100 different simulated data sets from each of the three models considered are shown in figure 4. It is apparent that the chosen statistics vary between data sets generated quite considerably — this is to be expected as the statistics required for sufficiency will vary depending on the data. For data generated by all of the models it is apparent that **S4**, the average SNP homozygosity, is selected often, whilst as expected the uninformative random statistic **S11** is rarely chosen. For data generated under the exponential growth model, **S4**, as well as **S3** (homozygosity), **S6** (mean number of pair-wise differences between haplotypes) and **S9** (linkage disequilibrium), appear to be favoured by the model selection approach. Data generated from this model apparently requires more statistics than data generated from the Null model to achieve sufficiency. The method applied to data generated from the two island model selects statistics **S4** and **S9** often, interestingly seeming to require fewer statistics than the exponential growth model to perform model selection.

This is a new and initially perhaps surpising finding: the summaries chosen by our model selection approach depend subtly on the true data-generating model. This is, by hindsight, however, not unexpected: we are trying to achieve sufficiency for model parameters first, and then pool the statistics required to do just that for all models, before refining this set of statistics in order to obtain sufficiency for model selection. As some models will generate data that is more difficult to obtain under other models than is the case vice versa, such relative biases will affect the set of statistics chosen. In light of population genetics theory, therefore, our observations are completely in line with our understanding of coalescent processes (see e.g. Hein *et al.* (2005)).

7.3 Random walk models

We also apply our framework to the problem of model selection on random walks, using a number of summary statistics. The models (Rudnick & Gaspari, 2010) under consideration were:

Model 1 Brownian motion.

Model 2 Persistent random walk (where the walk is more likely to continue in the same direction over successive steps but does not have a particular favoured orientation).

Model 3 Biased random walk (where one direction is favoured).

We used five summary statistics, that are individually not sufficient in more than one dimension for any of the random walk models:

S1 Mean square displacement.

S2 Mean x and y displacement.

S3 Mean square x and y displacement.

S4 Straightness index.

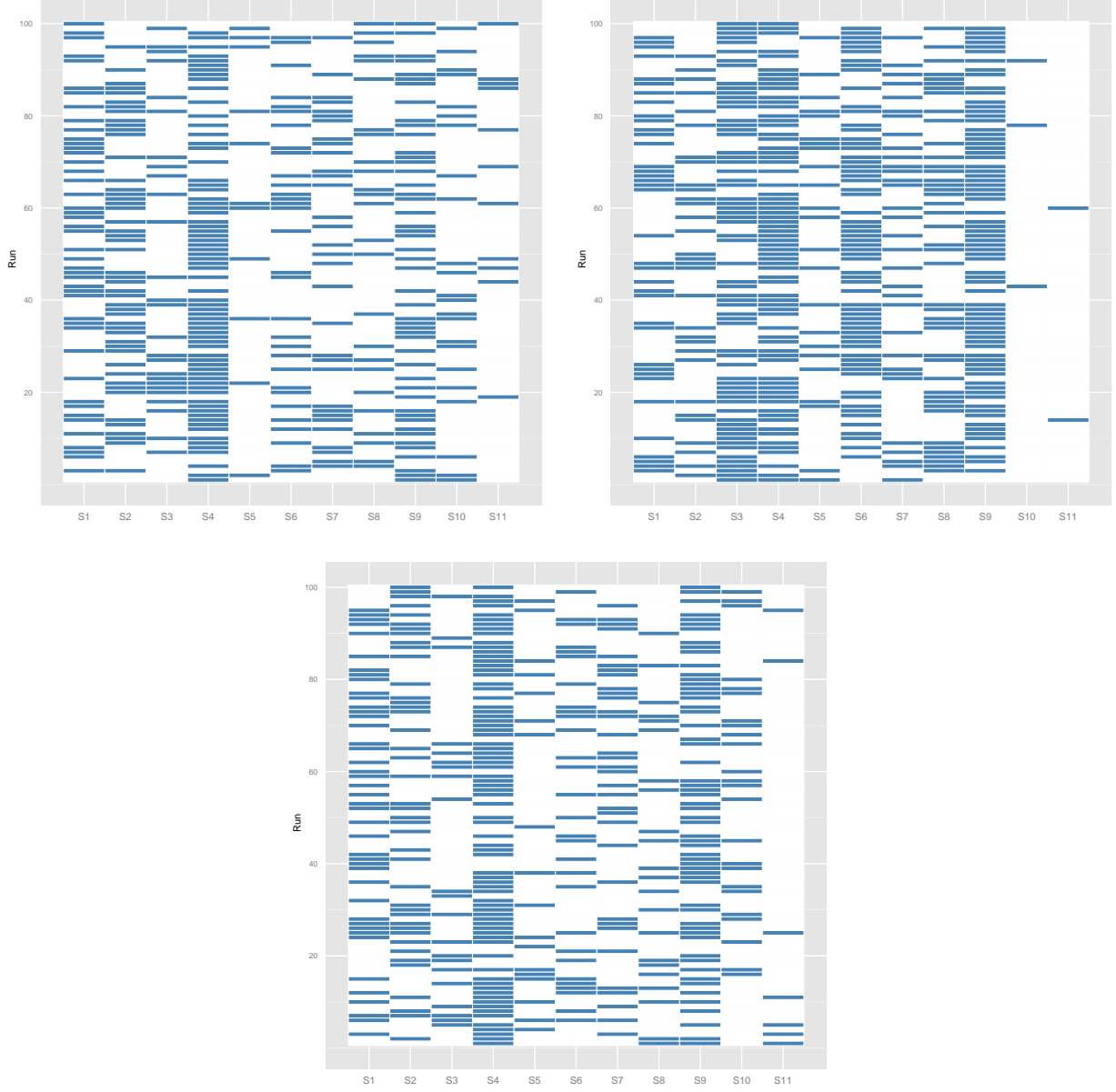**S5** Eigenvalues of gyration tensor (reference random walks book).

Figure 4: Summary statistics selected in 100 runs of the automated summary statistic selection procedure on simulated data sets from our three population genetics models. (each run is performed on a different simulated observed data point). A) Statistics chosen for model selection with data generated from model 1. B) Statistics chosen for model selection with data generated from model 2. C) Statistics chosen for model selection with data generated from model 3.

Since the models have multiple parameters, we can no longer apply a χ^2 test to select sufficient statistics, and so instead we approximate the KL-divergence using the posterior, applying the formula described in (Boltz *et al.*, 2007),

$$KL(p_X||p_Y) \approx \log \frac{N_V}{N_U - 1} + d\mathrm{E}_U[\log \rho_k(\cdot, V)] - d\mathrm{E}_U[\log \rho_k(\cdot, U)], \tag{11}$$

where U and V are the sets of posterior particles drawn from distributions p_X and p_Y respectively, d is the number of parameters and $\mathrm{E}_U[\log \rho_k(\cdot, V)]$ is the expectation of the distance to the kth nearest neighbour in the set of particles V, $\rho_k(u, V)$ of each particle $u \in U$.

Applying formula (11) in our summary statistic selection framework to data simulated from the three different models over 100 runs, the statistics shown in figure 5 are chosen. Again it is apparent that there are some differences in the selected statistics for different data sets generated. Looking at the summary statistics selected by our method, statistic **S5**, the eigenvalues of the gyration tensor, a measure of the anisotropy of the random walk, appears to be chosen often for data generated by all three models. There also appears to be a slight preference for statistic **S2**, the mean x and y displacement, which can be understood given that this statistic is necessary for sufficiency for parameter inference on the biased random walk model. We need to stress that the structure of the data here is complex and summary statistics are expected to be hugely variable.

8 Discussion

Sufficient statistics are rare; in convenient form — i.e. where the number of statistics is equal to the number of parameters to be estimated — they are restricted to problems that can be described in terms of models that belong to the exponential family (Lehmann & Casella, 1993; Didelot *et al.*, 2010). As previous authors have pointed out it is necessary to develop methods that construct sets of statistics that are (at least approximately (Le Cam, 1964; Kusama, 1976)) sufficient (Joyce & Marjoram, 2008; Nunes & Balding, 2010; Fearnhead & Prangle, 2010a,b). It is either this, or reinterpreting ABC-based inferences not as approximations to the full Bayesian (and thus likelihood-based) apparatus but as inference procedures in their own right (Wilkinson, 2008; Drovandi *et al.*, 2011), potentially systematically biased or for approximate models. A third approach, previously advocated, is to consider model checking rather than model selection as a viable way of ensuring that only appropriate models are calibrated against data. We believe that the latter position fails to acknowledge the role of sufficiency of statistics also in the context of parameter estimation; and we will briefly return to this point below after having addressed the other two points.

All methods aimed at constructing collections of statistics that taken together are (approximately) sufficient will fail, almost trivially, unless an exhaustive set of summary statistics can be envisioned which fulfils the sufficiency criteria as outlined above. If that is not the case, then we might naively expect that all candidate summary statistics from our starting set $\mathcal{S}$ will be included in $\mathcal{U}$. This, however, need not (and we believe generally will not) be the case, as the information theoretical framework will tend to bias against inclusion of statistics that are in some way co-linear to any statistics that are already included in the constructed set. It is, of course, in principle possible to use the KL divergence with respect to the distribution obtained with the full data as an overall benchmark, but in cases where this is indeed possible, it may be best to use the full data (see e.g. Toni *et al.* (2009)) for inference rather than risk the information reduction inherent to most summary statistics.

There has been much interest in trying to interpret ABC not solely as an approximation to the "true" posterior, but as an inferential framework in its own right (Wilkinson, 2008; Drovandi *et al.*, 2011). This is perhaps an attractive option. One way of achieving this shift in perspective is to consider distributions such as

$$p(M, \theta|S)$$

as distributions which specify the probability of a parameter and model being in concordance with a given summary statistic. If all we care about is that a model and parameter combination have high (or low) probability of producing data with certain mean/maximum/minimum or any other summary statistic value, then this is perfect. It is easy to envisage scenarios where we are only interested in certain aspects of the data (such as maximum water levels). ABC methods can be used to infer parameters (and models) that are more likely to give rise to simulated data that shares some but not all characteristics of the data. Interestingly, this would also allow us to employ ABC as a design tool (Barnes *et al.*, 2011): we specify the data (or system behaviour)
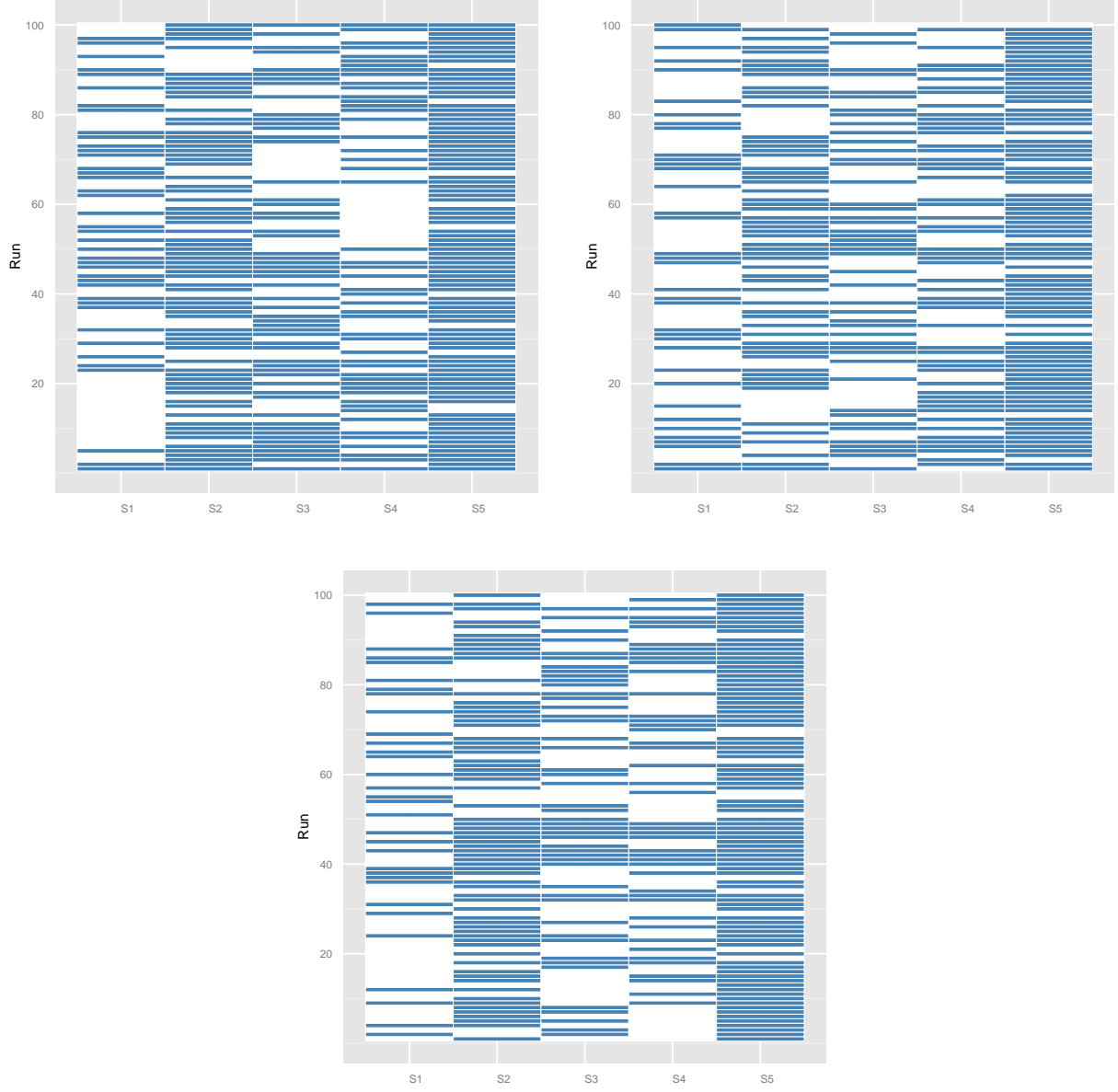
Figure 5: Summary statistics selected in 100 runs of the automated summary statistic selection on different simulated data sets, considering three different random walk models. A) Statistics chosen for model selection with data generated from model 1. B) Statistics chosen for model selection with data generated from model 2. C) Statistics chosen for model selection with data generated from model 3.

that we would like to observe and infer parameters (and models) which have high probability of producing these types of behaviour.

While this may perhaps seem like sophistry it does also have serious implications for model checking: any ABC approach that is based on summary statistics will infer model parameters (or marginal model posteriors) that reflect the behaviour encoded by these statistics. Thus we can no longer use these same statistics for model checking. This reflects the need to use non-sufficient summary statistics for model checking from Bayesian posterior predictive distributions: if we perform inference under a model for which a sufficient statistic exists, then calculating the same statistic for replicate data generated from the posterior predictive distribution will result in test statistics that are in line with the observed data, irrespective of the validity of the model. Hence some authors, in particular Gelman *et al.* (2003), strongly advocate the use of graphical model checking techniques over numerical tests. We feel that the situation in ABC reflects some of the same problems that are also encountered in model checking. Thus in an ABC framework, irrespective of whether the statistics are sufficient or not, the posterior distributions reflect the choice of statistics and the same statistics are therefore ill-suited for model checking.

We conclude by reiterating that ABC approaches employing summary statistics rather than the whole data have to fully engage with the level of information-loss inherent to summary statistics. Notions of simple sufficient statistics probably do not apply for most scientifically interesting and challenging problems and the use of statistics rather than the real data will always result in loss of information. Our approach is based around the assessment of information loss and allows the principled construction of sets of statistics (from a candidate set) that capture as much as possible from the observed data. While not a panacea, it is within the computational reach of ABC practitioners and makes information loss due to inadequate use of statistics apparent, for both the parameter and model selection problems.

References

Barnes, C., Silk, D., Sheng, X. & Stumpf, M. 2011 Bayesian design of synthetic biological systems *ARXIV* **q-bio.MN, stat.AP**, 1103.1046.

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

Blum, M. G. B. & Tran, V. C. 2008 HIV with contact-tracing: a case study in Approximate Bayesian Computation *ARXIV* **stat.AP** published in: Biostatistics 11, 4 (2010) 644-660.

Boltz, S., Debreuve, E. & Barlaud, M. 2007 Eighth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '07) in *Eighth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '07)* 16–16 IEEE.

Burnham, K. & Anderson, D. 2002 Model selection and multimodel inference: A practical information-theoretic approach *Springer Science+Business Media, Inc.* .

Cover, T. & Thomas, J. 2006 *Elements of Information Theory* Wiley-Interscience.

Cox, D. 2006 *Principles of Statistical Inference* Cambridge: Cambridge University Press.

Cox, D. & Hinkley, D. 1974 *Theoretical Statistics* New York: Chapman&Hall/CRC.

Didelot, X., Everitt, R., Johansen, A. & Lawson, D. 2010 Likelihood-free estimation of model evidence *warwick.ac.uk* .

Drovandi, C. C., Pettitt, A. N. & Faddy, M. J. 2011 Approximate Bayesian computation using indirect inference *Journal of the Royal Statistical Society Series C-Applied Statistics* **60**, 317–337.

Ewens, W. 2004 *Mathematical Population Genetics* 2nd edition New York: Springer.

Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S. *et al.* 2007 Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* **104**, 17614–17619.

Fearnhead, P. & Prangle, D. 2010a Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic ABC *arXiv.org* **stat.ME** v2: Revised in response to reviewer comments, adding more examples and a method for inference from multiple data sources.

Fearnhead, P. & Prangle, D. 2010b Semi-automatic Approximate Bayesian Computation *ARXIV* **stat.ME**.

Gelman, A., J.B., C., Stern, H. & Rubin, D. 2003 *Bayesian Data Analysis* 2nd edition Chapman & Hall/CRC.

Hein, J., Schierup, M. & Wiuf, C. 2005 *Gene Genealogies, variation and evolution* Oxford University Press.

Hudson, R. R. 1991 Gene genealogies and the coalescent process.

Joyce, P. & Marjoram, P. 2008 Approximately Sufficient Statistics and Bayesian Computation *Statistical Applications in Genetics and Molecular Biology* Look up nearly sufficient statistics.

Kusama, T. 1976 On approximate sufficiency *Osaka Journal of Mathematics* **13**, 661–669.

Le Cam, L. 1964 Sufficiency and Approximate Sufficiency *The Annals of Mathematical Statistics* **35**, 1419–1455.

Lehmann, E. & Casella, G. 1993 *Theory of point estimation* Springer.

Liepe, J., Barnes, C., Cule, E., Erguler, K. *et al.* 2010 ABC-SysBio–approximate Bayesian computation in Python with GPU support. *Bioinformatics (Oxford, England)* **26**, 1797–1799.

Mackay, D. J. 2003 *Information theory, inference and learning algorithms* Cambridge University Press.

May, R. M. 2004 Uses and abuses of mathematics in biology *Science* **303**, 790–3 doi:10.1126/science.1094442.

Mézard, M. & Montanari, A. 2009 *Information, Physics and Computation* Oxford University Press.

Nunes, M. A. & Balding, D. J. 2010 On Optimal Selection of Summary Statistics for Approximate Bayesian Computation *Statistical Applications in Genetics and Molecular Biology* **9**.

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

Ratmann, O., Jorgensen, O., Hinkley, T., Stumpf, M., Richardson, S. & Wiuf, C. 2007 Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Computational Biology* **3**, e230.

Robert, C. 2007 *The Bayesian Choice* Springer.

Robert, C. P., Cornuet, J.-M., Marin, J.-M. & Pillai, N. 2011 Lack of confidence in ABC model choice **stat.ME** 8 pages, 7 figures, 1 table, submitted to the Proceedings of the National Academy of Sciences, extension of arXiv:1101.5091.

Rudnick, J. & Gaspari, G. 2010 *Elements of the Random Walk* Cambridge University Press.

Secrier, M., Toni, T. & Stumpf, M. P. H. 2009 The ABC of reverse engineering biological signalling systems. *Molecular Biosystems* **5**, 1925–1935.

Tanaka, M. M., Francis, A. R., Luciani, F. & Sisson, S. A. 2006 Using approximate bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics* **173**, 1511–1520 doi:10.1534/genetics.106.055574.

Toni, T. & Stumpf, M. P. H. 2010 Simulation-based model selection for dynamical systems in systems and population biology *Bioinformatics* **26**, 104–110.

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

Walker, D. M., Allingham, D., Lee, H. W. J. & Small, M. 2010 Parameter inference in small world network disease models with approximate Bayesian Computational methods *PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS* **389**, 540–548.

Wilkinson, R. D. 2008 Approximate Bayesian computation (ABC) gives exact results under the assumption of model error *ARXIV* **stat.CO** 13 pages, 1 figure.

Wilkinson, R. D., Steiper, M. E., Soligo, C., Martin, R. D., Yang, Z. & Tavaré, S. 2010 Dating Primate Divergences through an Integrated Analysis of Palaeontological and Molecular Data *Systematic Biology* .