Cabras, S., Castellanos, M. E., & Ratmann, O. (2021). Goodness of fit for models with intractable likelihood.*TEST*, 30 (3), pp. 713-736.

DOI: 10.1007/s11749-020-00747-7

Goodness of fit for models with intractable likelihood

Stefano Cabras[1] · **María Eugenia Castellanos**[2,3] · **Oliver Ratmann**[4]

Abstract

Routine goodness-of-fit analyses of complex models with intractable likelihoods are hampered by a lack of computationally tractable diagnostic measures with well-understood frequency properties, that is, with a known sampling distribution. This frustrates the ability to assess the extremity of the data relative to fitted simulation models in terms of pre-specified test statistics, an essential requirement for model improvement. Given an Approximate Bayesian Computation setting for a posited model with an intractable likelihood for which it is possible to simulate from them, we present a general and computationally inexpensive Monte Carlo framework for obtaining p-valuesthat are asymptotically uniformly distributed in [0, 1] under the posited model when assumptions about the asymptotic equivalence between the conditional statistic and the maximum likelihood estimator hold. The proposed framework follows almost directly from the conditional predictive p-value proposed in the Bayesian literature. Numerical investigations demonstrate favorable power properties in detecting actual model discrepancies relative to other diagnostic approaches. We illustrate the technique on analytically tractable examples and on a complex tuberculosis transmission model.

Authors have been founded by MINECO-Spain projects PID2019-104790GB-I00 (M.E. Castellanos and S. Cabras) and Wellcome Trust fellowship WR092311MF (O. Ratmann).

✉ Stefano Cabras
stefano.cabras@uc3m.es

María Eugenia Castellanos
maria.castellanos@urjc.es

Oliver Ratmann
oliver.ratmann@imperial.ac.uk

1 Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain

2 Department of Informatics and Statistics, Universidad Rey Juan Carlos, Madrid, Spain

3 Department of Economics, Università degli Studi di Cagliari, Cagliari, Italy

4 Imperial College London, London, UK

Keywords Approximate Bayesian computation · Model adequacy · Model checking · Simulation-based modeling

Mathematics Subject Classification 62F15

1 Introduction

Complex simulation models are increasingly used in all areas of applied sciences for experimental design (Barnes et al. 2011), understanding the behavior of complex systems (Norris et al. 2016), or forecasting (Gneiting and Raftery 2005; Granich et al. 2009). These models are fitted to available data with computationally expensive, customized statistical techniques such as Latin hypercube sampling (Stein 1987), indirect inference (Gouriéroux et al. 1993), kriging (Cressie 2015), or with the technique of interest here named Approximate Bayesian Computations (ABC) (see for instance, Lintusaari et al. 2017; Sisson et al. 2017). Multiple versions of these models typically need to be formulated through several iterations, before suitable model behavior is achieved in terms of pre-specified diagnostic measures (Box 1976). Indeed, this iterative process is essential because the behavior of complex simulation models is a priori poorly understood.

For complex models, those in which the complexity is such that the corresponding likelihood cannot be treated analytically but it is possible to simulate from, a large range of diagnostic measures are easily calculated from simulation output (Becquet and Przeworski 2007; Hickerson and Meyer 2008; Liepe et al. 2012; Fearnhead and Prangle 2012; Poon 2015). However, the frequency properties of such diagnostic measures usually remain elusive (Meng 1994; Gelman et al. 1996; Ratmann et al. 2009; Lemaire et al. 2016). This hinders the interpretation of current goodness-of-fit (GOF)-type analyses, and therefore the process of model building.

Of particular interest is the distribution of diagnostic measures under the posited model (Meng 1994). If it were known, then the extremity of observed diagnostic values, calculated for the current model version, could be appropriately visualized and quantified. Efficient statistical procedures for obtaining such distributions would facilitate routine GOF testing of complex simulation models, as is standard practice for statistical models in other tractable settings (Huber-Carol et al. 2012).

The tail area probability of the observed diagnostic measure in the distribution under the posited model, h, i. e., the p-value, is the single most widely used tool to quantify discrepancies (or surprise) between the data and the currently entertained model. For complex models, a Bayesian formulation is indispensable because h remains dependent on unknown parameters (Rubin et al. 1984), which are nuisance in GOF and whose uncertainty is rationally accounted via integration, a natural operation in the Bayesian formulation.

Consider an arbitrary complex model $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}^K$ and assume that it is identifiable and estimable through observed data $x = x_{\text{obs}}$. Consider a collection of GOF test statistics $T = \{T_1, \ldots, T_L\}$ and denote the observed value of T_l by $t_{l,\text{obs}} = T_l(x_{\text{obs}})$, where the subscript l is omitted unless necessarily. Our focus is on separate p-values for each GOF test statistic T_l, defined, as usual, by

$$p_l = Pr^{h_l}(T_l \geq t_{l,\text{obs}}) = \int_{t_{l,\text{obs}}}^{\infty} h_l(t)\mathrm{d}t, \quad l = 1, \ldots, L \tag{1}$$

where it is assumed that large values of T_l indicate model incompatibility with x_{obs}, and h_l generally denotes the sampling distribution of T_l used to calculate p_l. As done for the observed GOF statistic and when it is not necessary, the subscript l in p_l or h_l will be suppressed. Here θ is usually fixed in the frequency setting or integrated out through a suitable distribution later specified. In regular models with analytic likelihoods, there are well-studied GOF statistics T (D'Agostino 1986), while in ABC choices of T are most of the time intuitive in the sense that it is recognized some agreement between the GOF feature of interest and the meaning of the observed values of T.

We here describe a new and general Bayesian approach for defining and sampling from a distribution h that leads to p-values that are asymptotically Uniform(0, 1) under the posited model, which we refer to as *asymptotically calibrated*. These have interpretable frequency properties (i.e., the same interpretation as in classical statistics) and are thus desirable for GOF.

Our results rest on how the dependence of the predictive distribution of diagnostic statistics T on unknown parameters θ is accounted for when integrating out θ to obtain the distribution h. A proper account of such dependency is necessary to obtain calibrated p-values, as θ is regarded as nuisance parameters in GOF. It can be integrated out in different ways and we focus on the proposal in Bayarri and Berger (1997) in order to obtain h that leads to calibrated p-values when very general assumptions described later hold for the posited complex model. In the current literature, integration of θ could be done either with respect to the prior density $\pi(\theta)$ or with respect to the posterior density

$$\pi(\theta \mid x_{\text{obs}}) = \frac{f(x_{\text{obs}} \mid \theta)\, \pi(\theta)}{m(x_{\text{obs}})}, \tag{2}$$

where $m(x_{\text{obs}}) = \int f(x_{\text{obs}} \mid \theta)\, \pi(\theta)\, d\theta$, leading to the prior and posterior predictive density of t

$$h^{\text{prior}}(t) = \int f(t \mid \theta)\, \pi(\theta)\, d\theta \tag{3a}$$

$$h^{\text{post}}(t \mid x_{\text{obs}}) = \int f(t \mid \theta)\, \pi(\theta \mid x_{\text{obs}})\, d\theta. \tag{3b}$$

(Box 1980; Guttman 1967; Rubin et al. 1984; Bayarri and Berger 2000), where $f(t \mid \theta)$ denotes the density associated with the sampling distribution of T. (In general, we follow the usual practice that the argument of the density indicates of which random variable f is a density). The prior predictive approach in (3a) is sensitive to prior choice, while the posterior predictive approach in (3b) is complicated through the double use of the data in (1) and (3b) which does not lead to asymptotically calibrated p-values (Bayarri and Berger 2000; Bayarri and Castellanos 2007). In fact, when we use (3b) in conjunction with (1) to calculate p_l the data x_{obs} are used twice: to compute

the observed value of the statistic T_l, $t_{l,\text{obs}}$, and to train the vague or even improper prior, $\pi(\theta)$, into the proper posterior $\pi(\theta \mid x_{\text{obs}})$.

Under approaches (3a) and (3b) p-values are not calibrated in small samples and are typically intractable to calculate for complex models. For instance, (3b) requires additional model simulations which we precisely want to avoid.

In order to account for double use of the data in marginalizing with respect to θ, it is necessary then to separate the information from data used for testing and that used for fitting by approximating the following conditional predictive density of the test statistics,

$$h^{\text{cond}}(t \mid s_{\text{obs}}) = \int f(t \mid s_{\text{obs}}, \theta)\pi(\theta \mid s_{\text{obs}})\, d\theta, \tag{4}$$

where $s_{\text{obs}} = (S_1(x_{\text{obs}}), \ldots, S_K(x_{\text{obs}})) \in \mathbb{R}^K$ denotes the observed value of a conditioning statistic $\mathbf{S}$. Especially in the ABC context, S is used for parameter estimation and we assume, without loss of generality, that the dimension of S is the same as that of θ.

The conditional predictive density (4) is of particular interest here because it is less sensitive to the prior through the information in $\pi(\theta \mid s_{\text{obs}})$ provided by s_{obs} (Bayarri and Berger 1997). In addition, since we are conditioning on s_{obs} in $f(t \mid s_{\text{obs}}, \theta)$ then (4) does not suffer from the same double use of the data as (3b) (Bayarri and Berger 1997, 2000). The novelty of this paper consists in implementing (4) and providing its properties in the ABC context.

To illustrate the context of ABC, suppose it is possible to generate $\mathbf{s} \sim f(\mathbf{s}|\theta)$, by first generating from $f(x \mid \theta)$ and then calculating $\mathbf{s}$ from x. Then, for a given distance metric $\rho(\mathbf{s}, s_{\text{obs}})$ and a tolerance parameter $\epsilon > 0$, ABC operates (for instance through Algorithm 1 later exposed) the following approximation to the *true* and *unavailable* posterior (2), that is, for ϵ small enough,

$$\pi(\theta \mid s_{\text{obs}}) \approx \pi(\theta \mid x_{\text{obs}}),$$

where

$$\pi(\theta \mid s_{\text{obs}}) = \frac{f(s_{\text{obs}} \mid \theta)\, \pi(\theta)}{m(s_{\text{obs}})}, \tag{5}$$

and $m(s_{\text{obs}}) = \int f(s_{\text{obs}} \mid \theta)\, \pi(\theta)\, d\theta$ (Doksum and Lo 1990).

Incidentally, if $\mathbf{S}$ were sufficient for θ, then as $\epsilon \to 0$ we would have $\pi(\theta \mid s_{\text{obs}}) \equiv \pi(\theta \mid x_{\text{obs}})$. However, checking sufficiency is not possible under the assumed setup, and usual practice in ABC is to choose it manually or automatically to optimize parameter estimation (Csilléry et al. 2010; Barnes et al. 2012a; Fearnhead and Prangle 2012; Prangle 2015).

We argue, however, that the model being fitted here is not that of $\mathbf{X} \mid \theta, \pi(\theta)$, but rather that of $\mathbf{S} \mid \theta, \pi(\theta)$. That is, the model building process ended up in a choice of $\mathbf{S}$, ρ, and ϵ which are here considered as given. A key observation is that having accepted that this is the model under fitting, we have that $\mathbf{S}$ is trivially a sufficient statistic for

θ; the data $\mathbf{X}$ only enter the ABC setting through the conditioning summary statistic $\mathbf{S}$. We assume that $\mathbf{S}$ is not ancillary for θ, which is a requirement implicit in order to successfully estimate θ. For the trivial reason that GOF refers to model feature not governed by θ, we only require that T is not a deterministic function of $\mathbf{S}$. Otherwise, conditioning the distribution of T on the observed value of $\mathbf{S}$ would be immaterial, and (4) would be equivalent to (3b). However, T and $\mathbf{S}$ can also be dependent as such a dependency is accounted when integrating out θ in (4).

We show in Sect. 2 that p-values calculated over (4) can be easily obtained from ABC routines with very limited additional computational cost, and that such p-values are asymptotically calibrated. This, in the end, leads to a calibrated GOF procedure for complex models with intractable likelihood, which is new in ABC literature.

Although the level of exposition of the ABC routines is intentionally maintained at a very basic level, enough for our purposes, an extensive introduction to ABC techniques is not possible due to space constraints, and the reader not familiar with ABC could refer to Sisson et al. (2018).

Section 3 illustrates the performance of the GOF technique, indicating broad applicability. Discussions are left to Sect. 4.

It is important to state here for the sequel of the exposition that the following objects are given for the procedure we propose: (1) posited complex model $f(x \mid \theta)$; (2) prior $\pi(\theta)$; (3) conditioning statistics $\mathbf{S}$ used for ABC approximation; (4) the specific ABC algorithm including all relevant parameters as, for instance, the tolerance ϵ; and finally (5) a GOF statistic T which reflects model features that are of interest in assessing its fit. Objects (1)–(4) are related to the inferential problem on θ, while object (5) to the GOF analysis. The aim of this paper is to propose and discuss a GOF procedure given (1)–(5).

2 Asymptotically calibrated *p*-values for simulation models with intractable likelihoods

We begin by detailing the needed additional calculation step to the standard ABC for the proposed GOF test, which does not alter parameter estimation.

ABC circumvents evaluations of intractable likelihoods through repeated simulations of data x for proposed model parameters θ', and then records those θ' for which summaries of model simulations, $\mathbf{s} = (S_1(x), \ldots, S_K(x)) \in \mathbb{R}^K$, are close enough to summaries of the observed data, $s_{\text{obs}} = (S_1(x_{\text{obs}}), \ldots, S_K(x_{\text{obs}}))$. Closeness of the observed and simulated summaries is measured in terms of a distance function $\rho(\mathbf{s}, s_{\text{obs}})$ between $S_k(x)$ and $S_k(x_{\text{obs}})$ which can be the absolute, the euclidean or any other distance.

We extend the standard algorithm by recording the value test statistics $\boldsymbol{T} = \{T_1, \ldots, T_L\}$, at each accepted iteration as detailed in Algorithm 1.

Algorithm 1 (Calibrated goodness-of-fit via ABC)

1: Set the tolerance parameter $\epsilon > 0$ and the number of steps M. If the output of the distance metric is multivariate, multiple tolerance parameters are set;
2: **for** $m = 1$ to M **do**

3: **repeat**
4: sample $\theta^* \sim \pi(\theta)$;
5: simulate $x_{(m)} \sim f(\cdot \mid \theta^*)$;
6: calculate the summaries $\mathbf{s} = \big(S_1(x_{(m)}), \ldots, S_K(x_{(m)})\big)$;
7: calculate the error distance $\rho(\mathbf{s}, s_{\text{obs}})$;
8: **until** $\rho(\mathbf{s}, s_{\text{obs}}) < \epsilon$
9: Set $\theta_{(m)} = \theta^*$;
10: calculate $\boldsymbol{t}_{(m)} = \big(T_1(x_{(m)}), \ldots, T_L(x_{(m)})\big)$.
11: **end for**
12: **return** $\theta_{(1)}, \ldots, \theta_{(M)}$ and $\boldsymbol{t}_{(1)}, \ldots, \boldsymbol{t}_{(M)}$.

The new step 10 does not involve any further computational cost other than calculating the diagnostics $\boldsymbol{T}$ and can be integrated into more computationally advanced ABC Monte Carlo samplers (Marjoram et al. 2003; Beaumont et al. 2010; Wegmann et al. 2009; Jasra et al. 2012). For these reasons, the proposed extension does not compromise the general versatility of the standard ABC approach and all problems related to ABC estimation and tuning parameters (e.g., tolerance ϵ) remain unchanged. The sampling distribution of $\boldsymbol{\theta}$ converges, as $\epsilon \to 0$, to (5). However, in every ABC approximation, $\epsilon > 0$, therefore, the superscript ϵ indicates in the sequel an ABC approximation.

To be more specific, what we are actually approximating with ABC and $\epsilon > 0$ is

$$\pi(\theta \mid B_\epsilon),$$

where

$$B_\epsilon = \Big\{ x \ : \ \rho\big(\mathbf{s}(x), s_{\text{obs}}\big) < \epsilon \Big\}. \tag{6}$$

and thus we can define

$$\begin{aligned} h^{\text{cond-}\epsilon}(t) &\propto \int_\theta \int_x \mathbb{I}\{ \rho(\mathbf{s}(x), s_{\text{obs}}) < \epsilon \} f(t, \mathbf{s}(x) \mid \theta) \pi(\theta) \, dx \, d\theta, \\ &= \int \frac{f(t, B_\epsilon \mid \theta)}{f(B_\epsilon \mid \theta)} f(B_\epsilon \mid \theta) \pi(\theta) \, d\theta \\ &= \int f(t \mid B_\epsilon, \theta) f(B_\epsilon \mid \theta) \pi(\theta) \, d\theta \\ &\propto \int f(t \mid B_\epsilon, \theta) \pi(\theta \mid B_\epsilon) \, d\theta, \end{aligned} \tag{7}$$

and the last integral is approximated by Algorithm 1. For $\epsilon > 0$, $h^{\text{cond-}\epsilon}(t)$ is not exactly conditioned to s_{obs}. However, for $\epsilon \to 0$, $B_\epsilon \to \{x : \mathbf{s}(x) = s_{\text{obs}}\}$ and thus

$$h^{\text{cond-}\epsilon}(t) \approx h^{\text{cond}}(t \mid s_{\text{obs}}). \tag{8}$$

Again, as in practice $\epsilon > 0$ then $h^{\text{cond-}\epsilon}(t)$ is the actual available distribution of T used to approximate the proposed notion of p-value

$$\begin{aligned} p_l^{\text{cond-}\epsilon} &= Pr^{h_l^{\text{cond-}\epsilon}}(\, t_l \geq T_l(x_{\text{obs}})\,) \\ &= \frac{1}{C_\epsilon} \int_{T_l(x_{\text{obs}})}^{\infty} \int_\theta \int_x \mathbb{I}\{\, \rho(\mathbf{s}(x), s_{\text{obs}}) < \epsilon \,\} h_l(t, \mathbf{s}(x) \mid \theta) \pi(\theta)\, dx\, d\theta\, \mathrm{d}t_l, \end{aligned} \tag{9}$$

where C_ϵ is the unknown normalizing constant of (7). $p_l^{\text{cond-}\epsilon}$ is approximated with an ABC algorithm, either with the usual acceptance/rejection (Algorithm 1) or other variants including post-processing approaches as discussed in Sect. 4.

We refer to the $p_1^{\text{cond-}\epsilon}, \ldots, p_L^{\text{cond-}\epsilon}$ as ϵ-conditional p-values. The ϵ-conditional p-values have the same scale, which facilitates cross-comparison and prioritization of model components that may need to be updated (Fisher 1925). In addition, it will often be informative to reconstruct the multivariate distribution $h^{\text{cond-}\epsilon}$ and then locate the observed diagnostic statistics t_{obs} in $h^{\text{cond-}\epsilon}$ for model checking. This will be illustrated in an example of a better assessment of the GOF. In any case, ϵ-conditional p-values are always obtained by marginalizing the multivariate version of $h^{\text{cond-}\epsilon}$.

The calibration properties of ϵ-conditional p-values are derived by recalling that of p^{cond} derived from (4), out of the ABC context, when the likelihood is available analytically. In order to have a calibrated p^{cond}, Bayarri and Berger (2000) propose as conditioning statistics in (5) the conditional maximum likelihood estimate (MLE) of θ. Further, Robert and Rousseau (2002) and Fraser and Rousseau (2008) considered the much easier MLE $\hat{\theta} = \hat{\theta}(x_{\text{obs}})$ to the true but unknown θ_0 as the conditioning statistic in (5). Robert and Rousseau (2002) and Fraser and Rousseau (2008) showed that in a large class of parametric models, which encompasses the exponential family and satisfying certain moment conditions illustrated in Fraser and Rousseau (2008), $\hat{\theta}$-conditional type p-values

$$p_l^{\text{cond-}\hat{\theta}} = Pr^{h^{\text{cond-}\hat{\theta}}}(\, T_l \geq t_l(x_{\text{obs}})\,) \tag{10}$$

with $h^{\text{cond-}\hat{\theta}}(t) = \int f(t \mid \hat{\theta}, \theta)\pi(\theta \mid \hat{\theta}) d\theta$ are asymptotically Uniform(0, 1) distributed as $n \to \infty$. That is, for any fixed θ, if we repeatedly sample from $f(x \mid \theta)$ and compute (10), then for large n we obtain a sample from a distribution that is approximately Uniform(0, 1).

This result is not directly applicable in the ABC context as the complex model at hand cannot be easily assigned to the above mentioned class of models. However, the key assumption in Fraser and Rousseau (2008) is that the conditioning statistics $\mathbf{S}$ are actually a one-to-one transformation g of the unavailable $\hat{\theta}$ and

$$f(t \mid s_{\text{obs}}, \theta) \to f(t \mid g(\hat{\theta}), \theta), \tag{11}$$

as $n \to \infty$ in distribution.

Property (11) is not easily verifiable for the model $f(x \mid \theta)$. However, as argued at the end of Sect. 1 this is not the model actually fitted. The non-ancillary statistic $\mathbf{S}$

is sufficient for the actual fitted model, and hence any estimator $\hat{\theta}$ of $f(t \mid s_{\text{obs}}, \theta)$ is a one-to-one function of $\mathbf{S}$, and thus $\mathbf{S} = g(\hat{\theta})$ for all n and $\epsilon \to 0$.

Property (11) is understood to be valid for model $f(\mathbf{s} \mid \theta)$ when this model approximates adequately $f(x \mid \theta)$, but whether or not this occurs is related to estimation of θ rather than the GOF of $f(x \mid \theta)$. The example in Sect. 3.3 supports this under three situations: (a) $f(\mathbf{s} \mid \theta)$ and $f(x \mid \theta)$ are identical, being $\mathbf{S}$ sufficient, (b) only some information in $f(\mathbf{s} \mid \theta)$ is relevant for estimating $f(x \mid \theta)$ up to the situation (c) in which $\mathbf{S}$ is ancillary with respect to θ. Only in situation (c), property (11) cannot be assumed as valid for $f(\mathbf{s} \mid \theta)$ and still holds under the more plausible situation (b). Moreover, $\mathbf{S}$ ancillary for θ is useless in ABC.

We therefore claim that ϵ-conditional p-values under the posited model, for ϵ small enough, are calibrated even for small n as actually, the self-sufficiency argument of s_{obs} for $\pi(\theta \mid s_{\text{obs}})$ applies for all n. A part of the above general characterization, the claimed calibration can be assessed also in specific situations by simulating ϵ-conditional p-values under the null model in a neighborhood of the most probable a posteriori values of θ. Such a null distribution is understood to be that induced by the posited model under criticism. If uniformity of the resulting simulated ϵ-conditional p-values cannot be rejected, then this may suggest that also the observed ϵ-conditional p-value is calibrated and thus interpretable. Such an approach will be illustrated for an example below.

3 Examples and applications

We now further characterize the ϵ-conditional p-values in inferential settings in which the ground truth properties of the involved objects are known, with a focus on the impact of the tolerance parameter ϵ, the conditioning statistics, and power. This is done in order to make the conceptual points of the proposed approach. Section 3.1 summarizes alternative Bayesian p-values for ease of exposition. In Sect. 3.2, we illustrate our goodness-of-fit approach on one of the rare examples where the conditional predictive density h^{cond} is known analytically and thus compare with a ground truth $p_l^{\text{cond-}\hat{\theta}}$

In Sect. 3.3, we compare the behavior of ϵ-conditional p-values as the conditioning statistics $\mathbf{S}$ contain less information on the model parameters and it is in the limit ancillary to θ. This is done for a model in which we know $\hat{\theta}$.

In Sect. 3.4, we compare the power and computational efficiency of ϵ-conditional p-values in identifying model discrepancies to those of alternative diagnostic approaches.

Throughout, we will choose tolerance parameters along with standard recommendations for ABC parameter inference as this choice is related to the required precision in estimating the posterior distribution with ABC. This is done to illustrate that choices of ϵ that lead to satisfactory approximations to the posterior density (2) also lead to calibrated ϵ-conditional p-values.

We finally illustrate GOF for a more complex model than those illustrated above. Section 3.5 shows the use of ϵ-conditional p-values to assess a complex tuberculosis transmission model.

In order to make the exposition clear, details on simulations have been postponed in the supplementary text. In all examples, except in the tuberculosis transmission model, ρ is the absolute distance. Again all settings concerned to parameter estimation are assumed as given, meaning that all concerns on ABC for estimating posteriors distributions also apply here and we focus mainly on providing a reliable measure for model assessment.

Numerical examples may contain additional specific mathematical deductions and properties of the involved objects that are not of general interest for the proposed approach and are reported case by case.

3.1 Related *p*-value approaches

For comparison to the ϵ-conditional p-values in (9), we denote the prior predictive, posterior predictive and conditional predictive p-values by

$$p_l^{\text{prior}} = Pr^{h^{\text{prior}}}(t_l \geq T_l(x_{\text{obs}})) \tag{12a}$$

$$p_l^{\text{post}} = Pr^{h^{\text{post}}}(t_l \geq T_l(x_{\text{obs}})) \tag{12b}$$

$$p_l^{\text{cond}} = Pr^{h^{\text{cond}}}(t_l \geq T_l(x_{\text{obs}})), \tag{12c}$$

with densities defined in (3a), (3b) and (4), respectively. In case of models with intractable likelihoods, a part of the proposed $p_l^{\text{cond-}\epsilon}$ the corresponding ϵ-approximation to (12b) is

$$p_l^{\text{post-}\epsilon} = Pr^{h^{\text{post-}\epsilon}}(t_l \geq T_l(x_{\text{obs}})) \tag{13}$$

with density

$$h^{\text{post-}\epsilon}(t) = \int f(t \mid \theta)\, \pi(\theta \mid B_\epsilon) d\theta. \tag{14}$$

Computationally, $p_l^{\text{post-}\epsilon}$ are obtained by calculating t in step 10 in Algorithm 1 as $t = \boldsymbol{T}(\boldsymbol{x}')$, where $\boldsymbol{x}'$ are re-simulated from the accepted θ. Straightforwardly, a Monte Carlo sum on M summands is employed to obtain all above-mentioned ϵ-approximated p-values.

3.2 Analytically tractable toy example

We illustrate the basic behavior of ϵ-conditional p-values on a well-known example that is analytically tractable (Bayarri and Castellanos 2001) and Bertolino and Racugno (1997). We consider the data $x_{\text{obs}} = (0.7, 1, 1, 1, 1, 1, 2, 3, 4, 5)$ and aim at testing if x_{obs} were generated from an exponential model with rate parameter θ. The usual default prior is $\pi(\theta) \propto 1/\theta$. As summary and also conditioning statistic, we use the (known) sufficient statistic $S(x) = \sum_{i=1}^{n} x_i$. For the model diagnostic $T(x) = \min_{i=1,\ldots,n} x_i$, the density $h(t \mid s_{\text{obs}}, \theta)$ can be derived analytically

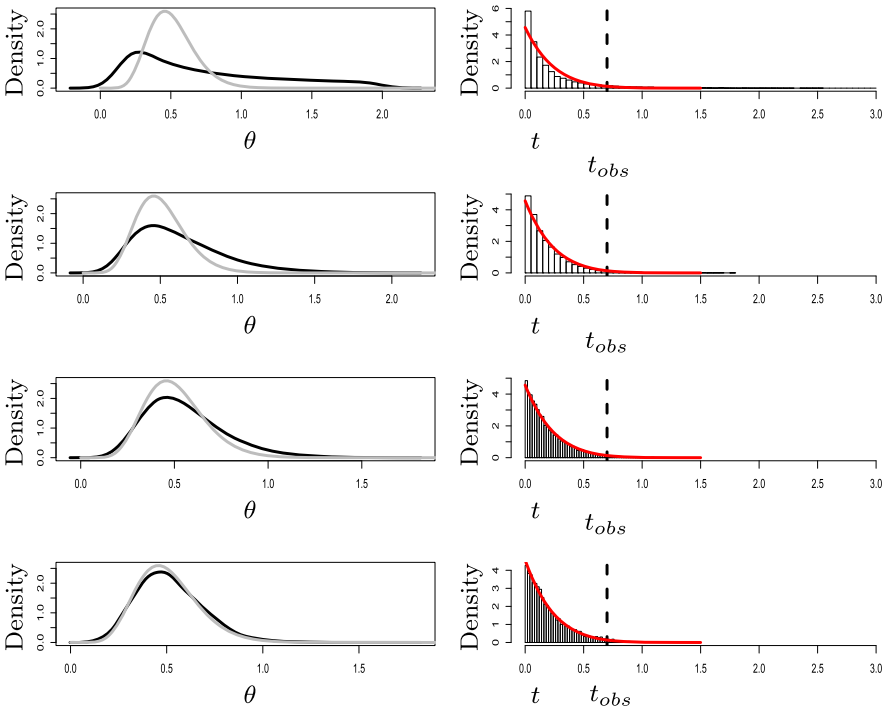

Fig. 1 Left column: Exact $\pi(\theta \mid x_{\text{obs}})$ in gray and ABC approximation in black as ϵ decreases from top to bottom, $\epsilon = 40, 10, 6.8, 3.7$, (respective MC acceptance rates were 90%, 35%, 20%, 10%). Right column: $h^{\text{cond-}\epsilon}(t)$ approximation (histogram) and exact $h(t \mid s_{\text{obs}})$ in red, t_{obs} is the vertical dashed line

$$h(t \mid s_{\text{obs}}) = \frac{(n-1)n}{s_{\text{obs}} - n\,t}\left(\frac{s_{\text{obs}} - n\,t}{s_{\text{obs}}}\right)^{n-1}.$$

Plugging this into (4) gives the conditional predictive p-value

$$p^{\text{cond}} = \left(1 - \frac{n\,t_{\text{obs}}}{s_{\text{obs}}}\right)^{n-1}.$$

For the above data, the conditional predictive p-value is $p^{\text{cond}} = 0.019$, indicating certain model incompatibility. The posterior predictive p-value is more conservative, $p^{post} = 0.048$, as expected.

We can use this example to compare the accuracy with which p^{cond} is approximated by $p^{\text{cond-}\epsilon}$. Since the prior is improper in this example, we used a Markov Chain Monte Carlo (MCMC) version of Algorithm 1 (see supplementary text 5.1).

Figure 1 compares the posterior density and the conditional predictive density to the corresponding ABC posterior densities and ϵ-conditional predictive densities $h^{\text{cond-}\epsilon}$ as the tolerance parameter ϵ decreases from top to bottom. Left column of Figure 1 shows exact posterior density of the rate parameter θ (grey) and ABC approximated posterior distribution (black) as the tolerance parameter ϵ decreases from top to bottom,

$\epsilon = 40, 10, 6.8, 3.7$ (the respective Monte Carlo (MC) acceptance rates were 90%, 35%, 20%, 10%). In comparison (the right column of Fig. 1), the ϵ-approximation to the conditional predictive density (histogram) is considerably less sensitive to increasing tolerance parameters when compared with the exact conditional predictive density (red). Dashed lines indicate the observed test statistic t_{obs}. For larger tolerances, ABC posterior densities are an increasingly poor approximation to the posterior density. However, this is not the case for $h^{\text{cond-}\epsilon}$, which remains an excellent approximation to the conditional predictive density h^{cond} up to very large tolerances that would not be used in typical ABC applications. The estimated $p^{\text{cond-}\epsilon}$ was 0.020 for tolerances of $\epsilon \geq 10$ and it was 0.06 for $\epsilon = 40$. Acceptance rates are meant as a proxy of the computational costs for either parameter estimation and model assessment: a lower acceptance rate implies a larger number of model simulations necessary to obtain the M samples for approximate the posterior of θ and the p-value.

3.3 Conditioning statistics with decreasing information on model parameters

Here are an illustration of how the ϵ-conditional p-values depend on reasonable and *unreasonable* choices of $\mathbf{S}$ and how much their sampling distribution is robust with respect to such a choice. We consider normally distributed data and a range of summary statistics which are, at different degrees, more or less informative for both the unknown mean θ_1 and variance θ_2. Let $x = (x_1, \ldots, x_n)$ with $x_i \sim \mathcal{N}(\theta_1, \theta_2)$ for $i = 1, \ldots, n = 100$. The posited model is the standard normal (e. g. $\theta_1 = 0, \theta_2 = 1$), and we considered a Uniform prior with $\theta_1 \in [-2, 2]$ and $\theta_2 \in [0.001, 2]$. In order of decreasing information, the conditioning statistics are $\mathbf{S}_1 = \{\bar{x}, s_x\}$, $\mathbf{S}_2 = \{x_{q1}, x_{q3}\}$, $\mathbf{S}_3 = \{x_{q1}, s_x\}$, and $\mathbf{S}_4 = \{\bar{x}\}$, where $\bar{x}$ is the sample mean, s_x the sample standard deviation, and x_{q1}, x_{q3} the first and third sample quartiles, respectively. The summaries $\mathbf{S}_1$ are sufficient for θ, $\mathbf{S}_2$ contain some information on θ, $\mathbf{S}_3$ contains less information about θ_1 and $\mathbf{S}_4$ is ancillary with respect to θ_2. As diagnostic statistic, we used the observed maximum of all data points, $T(x) = \max_{1,\ldots,n}(x_i)$.

For each set of conditioning statistics $\mathbf{S}_1$ - $\mathbf{S}_4$, we generated 1000 pseudo datasets and used Algorithm 1 under the true, data-generating model to estimate 1000 replicate p-values $p^{\text{cond-}\epsilon}$ (see supplementary text 5.2). In each case, and following standard practice (Csilléry et al. 2010), tolerances were set to obtain pre-specified Monte Carlo acceptance rates: we used 10% (larger than in typical ABC applications), 1% (representing a typical choice), and 0.1% (strict).

Figure 2 characterizes the sampling distribution of the ϵ-conditional p-values against quantiles of the Uniform(0, 1) distribution. We evaluated the extent to which the ϵ-conditional p-values in (9) are no longer Uniform(0, 1) distributed as conditioning statistics have decreasing information on model parameters (Fig. 2, panels top left to bottom right) and as tolerance parameters are relaxed (colors). Tolerances correspond to 10%, 1% and 0.1% acceptance probabilities of Algorithm 1. The ϵ-conditional p-values were approximately Uniform(0, 1) distributed for the conditioning statistics $\mathbf{S}_1$–$\mathbf{S}_3$ when tolerances corresponded to acceptance rates of 1% or less. The ϵ-conditional p-values were not Uniform(0, 1) distributed for $\mathbf{S}_4$, indicating that conditioning statistics need to contain information on all model parameters in

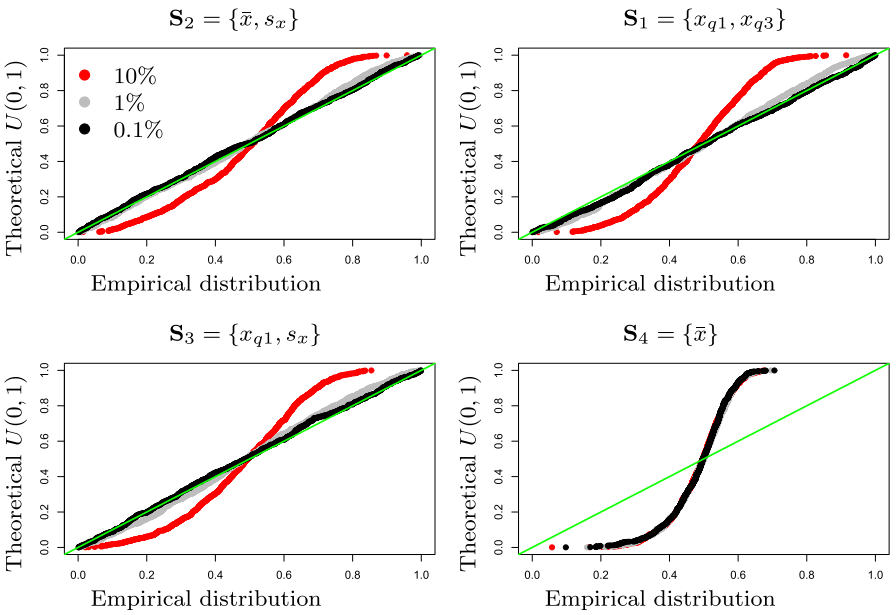

Fig. 2 Impact on ϵ-conditional p-values of conditioning statistics with decreasing information on model parameters

order to obtain approximately calibrated ϵ-conditional p-values, although conditioning statistics not necessarily need to be sufficient.

An interactive web tool to explore the behavior of ϵ-conditional p-values, in this normal setting under other configurations, is available at https://stefano-cabras.shinyapps.io/abc-pcpred-tutorial-app/.

3.4 Computational costs and ability to detect model discrepancies

Calibrated p-values although interpretable would be of limited utility if they fail to detect model discrepancies. For models with intractable likelihood, GOF has been poorly treated in the literature and as far as we know, at the moment no work has been published in the statistical literature, except Lemaire et al. (2016) which is still a preprint version. The work of Lemaire et al. (2016) presented p-values with good empirical calibration properties (no theory behind their calibration) in the sense that on several case studies, their p-values were below a nominal α-level of 5% of cases when the data were generated from the posited model and, in contrast to our investigations, the true θ_0 was not held fixed but re-simulated from the prior of the posited model. Notably, their proposed p-values had highly variable power properties. This prompted us to check in this section the power behavior of the ϵ-conditional p-values and compare with the p-values implemented in Lemaire et al. (2016).

We followed the Normal/Laplace case study by Lemaire et al. (2016) (see details in Sect. 5.3), generated normally distributed data, and studied to what extent model

mismatch of a posited Laplace model is identified with ϵ-conditional p-values, the two p-values proposed in Lemaire et al. (2016), and the ϵ-posterior predictive p-values in (13). We also evaluated the power in an opposite setting in which mismatch of a posited Normal model is to be identified on data generated under a Laplace model. This study was not meant to clarify under what alternatives the distribution of ϵ-conditional p-value would be stochastically smaller than the $U(0,\ 1)$, a condition that would justify rejecting the posited model for small realized p-values. Instead, this study illustrates how the choice of diagnostic statistics $\mathbf{T}$ is relevant to detect model discrepancy even under well-calibrated p-values. A general guideline for choosing $\mathbf{T}$ does not exist as it is model specific and in this paper is assumed to be given by the analyst who understands the involved model predictive features (as done here for the illustrated examples).

Datasets $x = (x_1, \ldots, x_n)$ of two sample sizes were considered, $n = 50$ and $n = 100$. Throughout, the actual and posited models were parametrized in terms of location and variance parameters $\theta = (\theta_1, \theta_2)$. Data were generated from $\theta_0 = (2, 4)$ in both cases and prior distributions are $\theta_1 \sim U(-10, 10), \theta_2 \sim \text{Inv-}\chi^2$ with 3 degrees of freedom, for both cases. To obtain a Monte Carlo estimate of the ϵ-conditional p-values and the ϵ-posterior predictive p-values, we used the sample mean and variance as conditioning statistics $\mathbf{S}_1$, and the skewness and kurtosis as diagnostic statistics $\mathbf{T}$. We then applied Algorithm 1 and a similar algorithm to obtain estimates of $p_l^{\text{post-}\epsilon}$ (see Sect. 3.1 and supplementary text 5.3).

The first p-value by Lemaire et al. (2016) is based on the average error between the data and simulations in terms of some distance of all summaries $\boldsymbol{Q}$ (sample mean, variance, skewness, kurtosis), when the simulations are from the ABC posterior density:

$$D^{\text{Lemaire1}-\epsilon}(\boldsymbol{q}_{\text{obs}}) = \frac{1}{M}\sum_{i=1}^{M} \rho(\boldsymbol{q}_i, \boldsymbol{q}_{\text{obs}}), \tag{15a}$$

$$\boldsymbol{q}_i \sim h_{B_\epsilon}(\boldsymbol{q}) \propto \int \mathbb{I}\{\rho(\boldsymbol{q}(x), \boldsymbol{q}_{\text{obs}}) < \epsilon\}\ f(x|\theta)\,\pi(\theta)\,d\theta. \tag{15b}$$

In practice, the $\boldsymbol{q}_i, i = 1, \ldots, M$, are accepted values of all summaries from a standard ABC routine. The first p-value is then given by

$$p^{\text{Lemaire1}-\epsilon} = Pr^{h_{\text{prior}}}\left(D^{\text{Lemaire1}-\epsilon}(\tilde{\boldsymbol{q}}) \geq D^{\text{Lemaire1}-\epsilon}(\boldsymbol{q}_{\text{obs}}) \right), \tag{16}$$

where the pseudo-data $\tilde{\boldsymbol{q}}$ are generated from the prior predictive distribution of all summaries under the posited model. The second p-value by Lemaire et al. (2016) is obtained similarly, but the simulations in step (15b) are resampled from the ABC posterior density:

$$D^{\text{Lemaire2}-\epsilon}(\boldsymbol{q}_{\text{obs}}) = \frac{1}{M}\sum_{i=1}^{M} \rho(\boldsymbol{q}_i, \boldsymbol{q}_{\text{obs}}), \tag{17a}$$

$$\boldsymbol{q}_i \sim h(\boldsymbol{q}|B_\epsilon) \propto \int f(\boldsymbol{q}|\theta)\left[\int \mathbb{I}\{\rho(\boldsymbol{q}(x), \boldsymbol{q}_{\text{obs}}) < \epsilon\}\, f(x|\theta)\, \pi(\theta)\, dx\right] d\theta, \quad (17b)$$

and

$$p^{\text{Lemaire2}-\epsilon} = Pr^{h_{\text{prior}}}\left(D^{\text{Lemaire2}-\epsilon}(\tilde{\boldsymbol{q}}) \geq D^{\text{Lemaire2}-\epsilon}(\boldsymbol{q}_{\text{obs}}) \right). \quad (18)$$

The second p-value by Lemaire et al. (2016) is more costly to calculate than ϵ-conditional p-values and ϵ-posterior predictive p-values (see supplementary text 5.3).

Table 1 compares power properties for the four p-values. Model discrepancies were detected in about the same number of replicate runs for the ϵ-posterior predictive p-values and the ϵ-conditional p-values. In this setting, these are very similar as the sample size is large for estimating the mean and variance of a normal distribution (although the likelihood of the posited model is not used explicitly). Moreover, in such a setting and because of the large sample size, even a subsequent calibration of the ϵ-posterior predictive p-values (with a nonnegligible increase in computations) would have produced the same performance in terms of power.

This was expected, because both diagnostics statistics are ancillary with respect to the location and scale parameters, and in this case, both p-values have similar power properties (Robins et al. 2000). Of note, the power behavior of $p^{\text{post-}\epsilon}$ and $p^{\text{cond-}\epsilon}$ was much less variable across scenarios when compared to that of the two p-values by Lemaire et al. (2016). The poor power behavior of the two p-values by Lemaire et al. (2016) can be attributed to very broad distributions of the diagnostics that can arise when assuming vague priors or heavier tails in the posited model. By contrast, the ϵ-conditional p-values are less sensitive to vague priors, which may explain why the power behavior of ϵ-conditional p-values were less variable in our evaluations.

Table 1 also compares computational costs for the four p-values. These are expressed as the number of simulations from the posited model that are required because simulation runtime in ABC is essentially driven by model simulations. The ϵ-conditional p-values were computationally cheapest, and this is an advantage especially in situations where simulations from the posited model are costly.

3.5 Goodness-of-fit of tuberculosis transmission model

Tanaka et al. (2006) estimated, using a stochastic simulation model, aspects of tuberculosis transmission dynamics by clustering patterns of genetic isolates belonging to the *Mycobacterium tuberculosis* pathogen at the population level.

Briefly, the model simulates the spread and diversification of such genetic isolates in a population through competing for stochastic events that correspond to the birth of an identical genotype (transmission), death of a genotype (recovery of the infected individual), and mutation of an existing genotype into a new type (diversification). The model generates a genotype profile of the infected population at any time point, counting how many individuals are infected with a distinct variant of the pathogen per time step. Simulated profiles can be compared to those observed in a sample of patients, in an ABC setting.

Table 1 Power and computational cost associated with ϵ-conditional p-values

True model (from which x_{obs} generated)	Posited model (used in inference)	Sample size	Power using kurtosis diagnostic			
			$\boldsymbol{p}^{\text{cond-}\boldsymbol{\epsilon}}$	$\boldsymbol{p}^{\text{post-}\boldsymbol{\epsilon}}$	$\boldsymbol{p}^{\text{Lemaire1}-\boldsymbol{\epsilon}}$	$\boldsymbol{p}^{\text{Lemaire2}-\boldsymbol{\epsilon}}$
			(proportion of repeated runs with p-value <5%)			
Normal	Laplace	50	56%	56%	7%	5%
Normal	Laplace	100	83%	85%	8%	8%
Laplace	Normal	50	63%	60%	10%	52%
Laplace	Normal	100	87%	86%	11%	78%
			Proper prior required			
			No	No	Yes	Yes
			Avg no. model simulations required*			
			$\frac{M}{AR}$	$M(1+\frac{1}{AR})$	$\frac{M}{AR}$	$n'M(1+\frac{1}{AR})$

*Computational cost is reported as a function of model simulations, where: M is the number of requested ABC samples, AR is the acceptance rate of the ABC algorithm, n' is the number of requested pseudo-data replicates

Through this fitting process, the simulation model allows us to infer the following features: net transmission rate, doubling time, and reproductive value of the pathogen. Various aspects of the fitting process have been discussed previously (Stadler 2011; Aandahl et al. 2014). However, the adequacy of the simulation model in capturing the genotype profile data has not been formally assessed although yet questioned in Aandahl et al. (2014). We endeavor here to assess the GOF of this model through the use of ϵ-conditional p-values, for which we also assess their calibration with respect to the Uniform(0, 1) distribution.

Our re-analysis of the tuberculosis transmission model consists of two parts. In the first, we verify on simulations that ϵ-conditional p-values can indeed be applied in this setting, in the sense that the $p^{\text{cond-}\epsilon}$ are approximately Uniform(0, 1) distributed under the posited model for previously published choices of conditioning statistics and tolerances (Tanaka et al. 2006). In the second part, we applied the ϵ-conditional p-values to assess the GOF of the simulation model against the genotype clustering data reported by Small et al. (1994).

The Birth–Death–Mutation (BDM) model was previously fitted to 473 cross-sectional patient samples from San Francisco, that could be grouped into the genotype clustering profile

$$x_{\text{obs}} = \{30^1,\ 23^1,\ 15^1,\ 10^1,\ 8^1,\ 5^2,\ 4^4,\ 3^{13},\ 2^{20},\ 1^{282}\}, \tag{19}$$

where $n_i^{k_i}$ indicates that there are k_i genotype clusters of n_i sampled cases (Small et al. 1994). Tanaka et al. (2006) used two summary statistics for fitting the BDM model, the number of distinct genotypes in the sample $S_1(x) = \sum k_i$ and the gene diversity $S_2(x) = 1 - \sum (n_i/n)^2$, where n is the total of sampled patients. As usually in ABC, it is not clear if the two summaries are sufficient, but ABC analyses using S_1, S_2 were shown to produce similar parameter estimates as exact likelihood methods (Aandahl et al. 2014). We here consider two diagnostic statistics, the maximum cluster size $T_1(x) = \max n_i$ and the number of unique genotypes $T_2(x) = \sum_{i:\, n_i=1} k_i$. Throughout, we adopt the same prior densities of the model parameters and the same conditioning statistics, distance functions and tolerance parameters as in the original study; see Tanaka et al. (2006) for details. We finally considered the prior specification in Aandahl et al. (2014) for a prior sensitivity analysis. A MCMC version of Algorithm 1 was used to fit the model and to obtain the ϵ-conditional p-values (see supplementary text 5.4).

To evaluate if the ϵ-conditional p-values of the diagnostics T_1 and T_2 are calibrated under the previously used conditioning statistics and tolerances, we generated 200 pseudo-datasets with the BDM model, using previous maximum a posteriori point estimates as true value θ_0 (Tanaka et al. 2006). For each dataset, we obtained two ϵ-conditional p-values for the two diagnostics, respectively, and then evaluated the frequency properties of the two ϵ-conditional p-values over the 200 pseudo-datasets. To ease the computational burden associated with this analysis, we simulated only until a population size of 1000 infected individuals (rather than 10,000 for the analysis of the San Francisco data) and calculated summaries and diagnostics on a sample of 50 individuals (rather than 426 for the analysis of the San Francisco data). In addition to

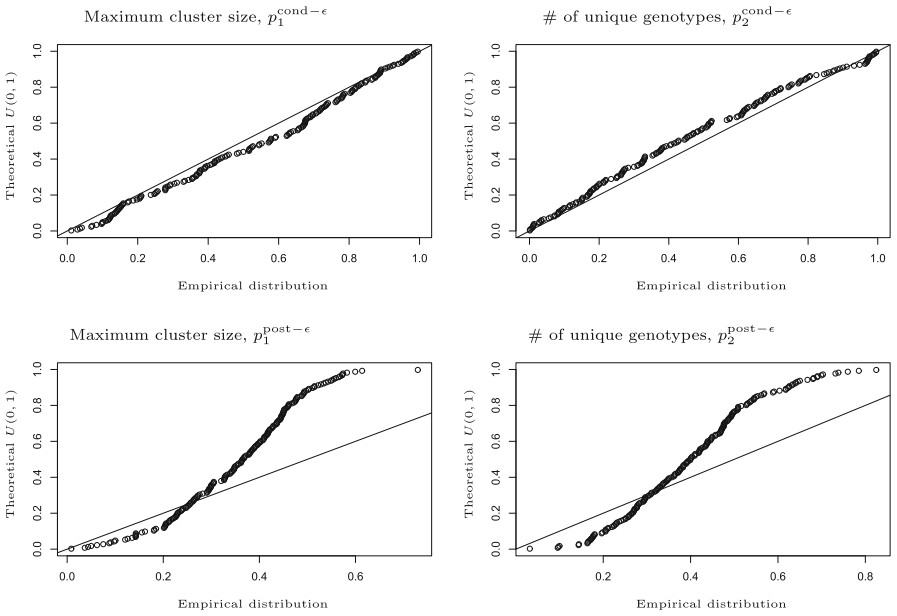

Fig. 3 Calibration properties of ϵ-conditional p-values and ϵ-posterior p-values on simulations from the tuberculosis transmission model

the ϵ-conditional p-values in (9), we also calculated for comparison the corresponding ϵ-posterior p-values in (13).

Figure 3 presents quantile-quantile plots of the distribution of the ϵ-conditional p-values for the two diagnostics T_1, T_2 against quantiles from the Uniform$(0, 1)$ distribution, and similar plots for the corresponding ϵ-posterior p-values. While the ϵ-posterior p-values are far from uniformity, the proposed ϵ-conditional p-values are approximately Uniform(0, 1) distributed.

This validation process may serve as a heuristic argument to validate, for this ABC setup, the assumptions in Fraser and Rousseau (2008) that otherwise cannot be assessed because of the involved intractable likelihood.

After this validation process, we applied the same GOF routine to check the adequacy of the BDM model in two specific directions: T_1, the maximum cluster size, and T_2, the number of unique genotypes. Posterior estimates of the model parameters coincided with those presented in Tanaka et al. (2006).

Figure 4 illustrates the estimated two-dimensional distribution $h^{\text{cond-}\epsilon}(t)$ in (7) and maps the location of the sample diagnostic statistics. The observed values of the two model diagnostics (marked by "x") are overlaid over the estimated ϵ-conditional distribution in (7) (colors).

The observed number of unique genotypes is significantly larger than expected under the fitted BDM model, while the observed maximum cluster size is supported by the fitted model. The marginal ϵ-conditional p-values were $p_1^{\text{cond-}\epsilon} = 0.258$ and $p_2^{\text{cond-}\epsilon} = 0.011$, marginal ϵ-conditional distributions appear in Fig. 5. The posterior

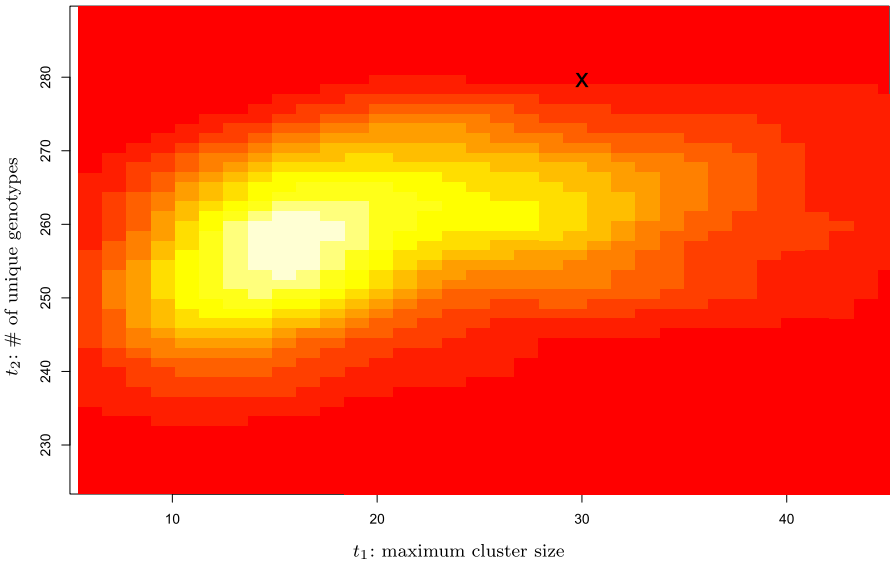

Fig. 4 Position of observed GOF statistics in the tuberculosis transmission model relative to the bivariate version of the ϵ-conditional distribution (7)

predictive densities in Fig. 5 are broader and lead to more conservative tail area probabilities of the diagnostics that are not calibrated. $p_1^{\text{post-}\epsilon} = 0.336$, $p_2^{\text{post-}\epsilon} = 0.103$.

Similar conclusions are reached for alternative prior specifications as shown in Fig. 6. This is a prior sensitivity analysis of results in Fig. 5 as the prior specification in Aandahl et al. (2014) was used. The MCMC variant of Algorithm 1 generated samples from the ϵ-conditional distribution in (7) (Fig. 6 top) and the ϵ-posterior distribution in (14) (Fig. 6 bottom). The p-values were $p_1^{\text{cond-}\epsilon} = 0.3$, $p_2^{\text{cond-}\epsilon} = 0.01$ and $p_1^{\text{post-}\epsilon} = 0.14$, $p_2^{\text{post-}\epsilon} = 0.048$. With the prior specification in Aandahl et al. (2014), T_2 reflects stronger incompatibility of the posited model against the genotype data, suggesting that the informative prior is less compatible with the observed number of unique genotypes. In contrast, the corresponding ϵ-posterior p-values did not convey the full extent of model mismatch.

These GOF analyses indicate that the fitted tuberculosis transmission model fails to adequately represent the number of unique genotypes. Unmodeled heterogeneity in disease transmission across individuals or time, importation of viral lineages into the local population, or uneven patient sampling could explain the larger number of observed unique genotypes, as also suggested in Aandahl et al. (2014).

4 Discussion

We here present a new and general computational procedure for assessing the GOF of complex simulation models through p-values that have interpretable frequency properties. The ϵ-conditional p-values, described in (9), can be calculated as a by-product

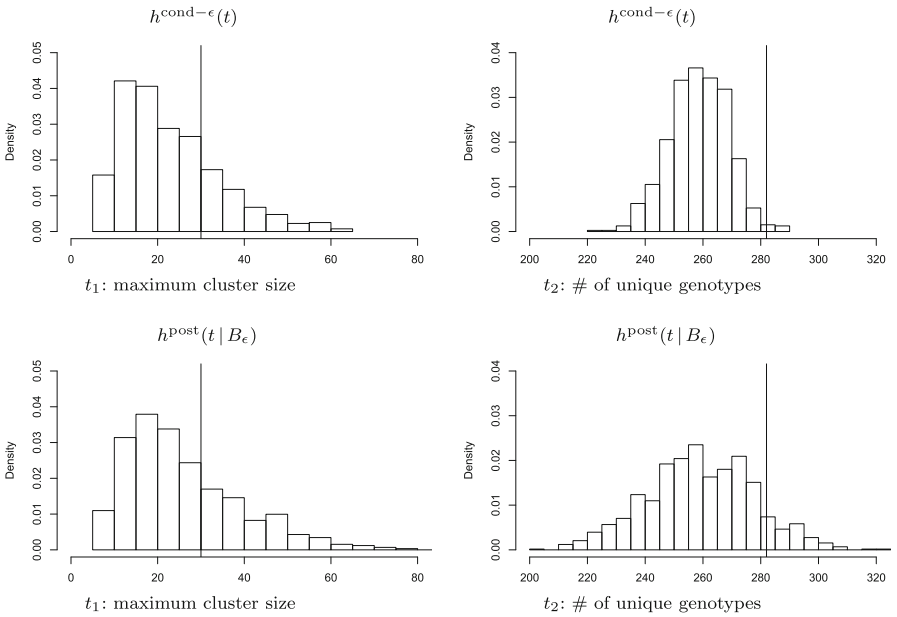

Fig. 5 ϵ-Conditional p-values and ϵ-posterior p-values for the tuberculosis transmission model on San Francisco genotype data

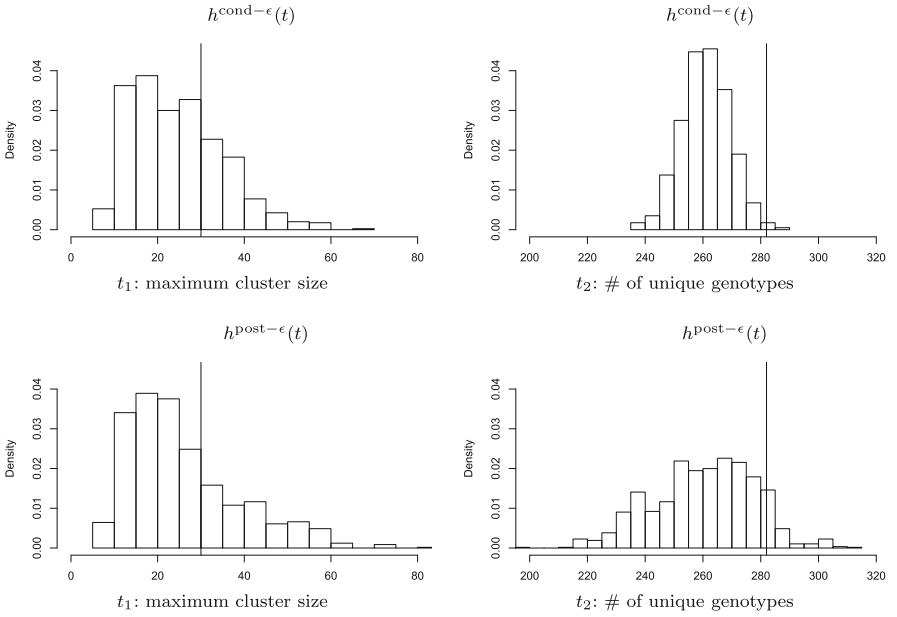

Fig. 6 ϵ-Conditional p-values and ϵ-posterior p-values for the tuberculosis transmission model on San Francisco genotype data under the prior specification in Aandahl et al. (2014)

of ABC routines at negligible additional computational cost. In fact, no additional simulations beyond standard ABC routines are required. ABC has been applied to a very broad class of complex simulation models, and for this reason, the proposed numerical approach for obtaining asymptotically calibrated p-valuesis also broadly applicable.

Among the different ABC approaches, it is worth discussing post-processing approaches as, for instance, the regression adjustment approach which re-weights or adjusts the distribution of parameter values obtained (for instance) by means of a rejection sampling in order to account for the imperfect match between simulations and observations Beaumont et al. (2002). Post-processing ends up in re-weighting the simulated values of θ and such a new set of weights can be also used to re-weight the simulated values of $\boldsymbol{T}$ in the proposed Algorithm 1. However, there are actually criticisms on the post-processing approach under model misspecification Frazier et al. (2017) which is exactly the main question addressed in this paper. For such a reason, we did not cast the message of using such post-processing approaches during model building (i.e., before a successful result on model GOF).

We showed first that ABC routines can be readily adapted to sample diagnostic statistics whose distribution approximates the conditional predictive distribution (4). A requirement for this approach is that the GOF statistics are not directly used for fitting as the conditioning of T to $\mathbf{S}$ would vanish in (4). This requirement is stated also because, in GOF literature, diagnostic statistics $\boldsymbol{T}$ should measure some type of departure from observed data and the assumed model, while conditioning statistics $\mathbf{S}$ should contain information about parameters, in order to infer on them. As an example, when checking the normality of data, the use of the mean or standard deviation does not measure any departure from data and normal model, while skewness or kurtosis are appropriate departure statistics.

Other Bayesian GOF tests refer to the existence of pivotal quantities as in Johnson (2004, 2007). Basically, the approach relies on choosing $\mathbf{S}$ and T such that $f(x|\theta) = f(\mathbf{s}, t|\theta)$ and thus obtain two types of factorization: $f(\mathbf{s}|t, \theta)h(t|\theta)$ or $h(t|\mathbf{s}, \theta)f(\mathbf{s}|\theta)$ and analyze when conditional or marginal distribution is independent of the nuisance parameters. Unfortunately, in ABC setting such factorizations are not analytically available, and furthermore, their approximation would require a double simulation as also the variability of T must be accounted in the joint $f(\mathbf{s}, t|\theta)$. This again complicates the GOF procedure making it unfeasible.

Bayesian p-valueshave been seriously criticized, including as a measure of discrepancy (Berger and Delampady 1987; Berger and Sellke 1987). We here take the view that, for complex models with intractable likelihoods, it is advantageous to consider p-valueswith acceptable frequency properties, rather than using ones with unknown frequency properties. Diagnostic measures without such properties, such as the ϵ-posterior p-values in (13), need to be calibrated by resampling from the model as proposed by Hjort et al. (2006). This incurs substantial computational costs especially when simulation models are complex and, in addition, tend to reduce power in detecting actual model discrepancies due to stochastic variation in repeated simulations.

With routine applications of ϵ-conditional p-values in mind, we emphasize that asymptotic calibration properties are guaranteed under the unverifiable regularity conditions in Fraser and Rousseau (2008) for the true and inaccessible posited model

$(\mathbf{X} \mid \theta, \pi(\theta))$, and the verifiable assumption (11) for the ABC posited model $(\mathbf{S} \mid \theta, \pi(\theta))$ which is the one under fitting. For example, if the conditioning statistics $\mathbf{S}$ are ancillary for θ, i. e., the conditioning statistics contain no information on θ, then $h^{\text{cond-}\epsilon}(t) = \int_\theta h(t \mid B_\epsilon, \theta)\pi(\theta)d\theta$ for any ϵ (as long as the prior density is proper). In this case, the ϵ-conditional p-values are no longer asymptotically Uniform(0, 1) distributed. The remark of this argument here is important because it extends more generally when the tolerances ϵ are very large because in this case, any conditioning statistic contains no information on θ, and thus, for instance, two models with similar GOF can differ just because different ϵ is chosen.

Sophisticated techniques now exist for choosing or constructing informative summary statistics (Csilléry et al. 2010; Fearnhead and Prangle 2012; Barnes et al. 2012b; Silk et al. 2013; Jiang et al. 2015; Sisson et al. 2017), making concerns around ancillary of $\mathbf{S}$ practically not relevant. In line with these expectations, we found in our investigations in Sects. 3.3 and 3.5 that ϵ-conditional p-values were approximately Uniform(0, 1) distributed also when the conditioning statistics were not sufficient. Nonetheless, we recommend verifying the frequentist properties of ϵ-conditional p-values on simulations prior to model assessment. For example, our validation analysis of the tuberculosis transmission model on simulated data in Sect. 3.5 indicated that the ϵ-conditional p-values are indeed asymptotically Uniform(0, 1) distributed for the considered configuration of conditioning statistics and tolerance parameters.

In summary, the proposed approach for obtaining asymptotically calibrated p-values is computationally cheap, easy to implement as part of ABC routines, and robust within typical choices of the conditioning statistics $\mathbf{S}$ and the tolerance parameter ϵ.

5 Supplementary text

5.1 Simulation details for Sect. 3.2

A Markov Chain Monte Carlo (MCMC) version of Algorithm 1 was implemented to obtain the ϵ-conditional p-value defined in (9) for different choices of the ABC tolerance parameter ϵ (Marjoram et al. 2003). The MCMC routine was run for 100, 000 MCMC steps after a burn-in of 10,000 steps, with proposal density $q(\theta) = U(0.01, 2)$ and ABC distance function $\rho = |s - s_{\text{obs}}|$. MCMC output was trimmed to every fifth iteration, yielding $M = 20{,}000$ samples from the approximate distribution $h^{\text{cond-}\epsilon}(t)$ of the diagnostic statistic. The ϵ-conditional p-value was obtained through the usual Monte Carlo proportions used to estimate of (9), that is the proportion of $t_{(1)}, \ldots, t_{(M)}$ obtained from Algorithm 1, that are greater than t_{obs}.

5.2 Simulation details for Sect. 3.3

Algorithm 1 was run for 10,000 simulations. The distance function was set to $\rho(s, s_{\text{obs}}) = |s - s_{\text{obs}}|$ for both components and rejections occurs if either $\rho(s_1, s_{1obs}) > \epsilon_1$ or $\rho(s_2, s_{2obs}) > \epsilon_2$.

5.3 Simulation details for Sect. 3.4

To obtain the p-values in Sect. 3.4, we used throughout the prior density $\pi(\theta_1, \theta_2) = U(-10, 10)(\theta_1) \times \text{Inv-}\chi^2(3)(\theta_2)$, where Inv-$\chi^2(d)$ denotes the density of the inverse χ^2 distribution with $d = 3$ degrees of freedom. Algorithm 1 was used to generate a Monte Carlo estimate of ϵ-conditional p-values. To generate a Monte Carlo estimate of the ϵ-posterior predictive p-values, we added to Algorithm 1 the following simulation step consisting of $x_i^{\text{rep}} \sim f(x|\theta_i)$ for each accepted θ_i and $t_i^{\text{rep}} = T(x_i^{\text{rep}})$. To obtain the ABC approximation of the first p-value by Lemaire et al. (2016), it is possible to use the same output from the ABC routine run to evaluate $D^{\text{Lemaire1}-\epsilon}(\boldsymbol{q}_{\text{obs}})$, and so calculation of $p^{\text{Lemaire1}-\epsilon}$ does not increase computational cost. However, to obtain the second p-value by Lemaire et al. (2016), it is necessary to make a double simulation. Pseudo-data is simulated from the prior predictive n' times (in order to mitigate computational cost, $n' << M$). Hence, for $i = 1, \ldots, n'$, $\theta_i \sim \pi(\theta)$ and for each θ_i, $x_i \sim f(x \mid \theta_i)$, $\tilde{\boldsymbol{q}}_{\text{obs},i} = \boldsymbol{Q}(x_i)$ and thus $D^{\text{Lemaire2}-\epsilon}(\tilde{\boldsymbol{q}}_{\text{obs},i})$ is obtained. This implies $n' \times (M/AR + M)$ simulations from the original model, where again M denotes the number of final simulations and AR the acceptation rate. Last term M is added because in order to calculate each $D^{\text{Lemaire2}-\epsilon}(\tilde{\boldsymbol{q}}_{\text{obs},i})$, it is necessary to obtain replicates $\boldsymbol{q}_i^{\text{rep}} \sim m(\boldsymbol{q} \mid \boldsymbol{q}_i)$, what in turn implies M more evaluations of the model.

5.4 Simulation details for Sect. 3.5

A MCMC version of Algorithm 1 (Marjoram et al. 2003) was used to approximate the ϵ posterior predictive distribution in (14) and the ϵ-conditional predictive distribution in (7). The MCMC routine was run for 50, 000 MCMC steps after a burn-in of 10,000 steps, with proposal density a truncated multivariate normal to positive values, with mean the previous accepted θ_t and covariance matrix the same used in Tanaka et al. (2006),

$$\Sigma = \begin{bmatrix} 0.5^2 & 0.225 & 0 \\ 0.225 & 0.5^2 & 0 \\ 0 & 0 & 0.015^2 \end{bmatrix}.$$

The ABC distance function used is also the same than in Tanaka et al. (2006), $\rho(\mathbf{s}, \mathbf{s}_{\text{obs}}) = 1/n|s_1(x) - s_{1,\text{obs}}| + |s_2(x) - s_{2,\text{obs}}|$ and $\epsilon = 0.025$. MCMC output was trimmed to every fifth iteration. Both ϵ p-values were obtained through the usual Monte Carlo proportions used to estimate (9) and (13).

References

Aandahl RZ, Stadler T, Sisson SA, Tanaka MM (2014) Exact vs. approximate computation: reconciling different estimates of mycobacterium tuberculosis epidemiological parameters. Genetics 196(4):1227–1230

Barnes CP, Silk D, Sheng X, Stumpf MP (2011) Bayesian design of synthetic biological systems. Proc Nat Acad Sci 108(37):15190–15195

Barnes CP, Filippi S, Stumpf MP, Thorne T (2012a) Considerate approaches to constructing summary statistics for ABC model selection. Stat Comput 22(6):1181–1197
Barnes CP, Filippi S, Stumpf MP, Thorne T (2012b) Considerate approaches to constructing summary statistics for abc model selection. Stat Comput 22(6):1181–1197
Bayarri MJ, Berger JO (1997) Measures of surprise in bayesian analysis. ISDS Discussion Paper, Duke University, Technical report
Bayarri MJ, Berger JO (2000) P values for composite null models. J Am Stat Assoc 95(452):1127–1142. https://doi.org/10.1080/01621459.2000.10474309
Bayarri MJ, Castellanos ME (2001) A comparison between p-values for goodness-of-fit checking. In: George EI (ed) Monographs of official statistics bayesian methods with applications to science. Policy and Official Statistics 1, pp 1–10
Bayarri MJ, Castellanos ME (2007) Bayesian checking of the second levels of hierarchical models. Stat Sci 22(3):322–343
Beaumont MA, Zhang W, Balding DJ (2002) Approximate bayesian computation in population genetics. Genetics 162(4):2025–2035
Beaumont MA, Cornuet JM, Marin JM, Robert CP (2010) Adaptivity for ABC algorithms: the ABC-PMC scheme. Biometrika 96(4):983–990
Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res 17(10):1505–1519
Berger JO, Delampady M (1987) Testing precise hypotheses. Stat Sci 3:317–352
Berger JO, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of p-value and evidence. J Am Stat Assoc 82:112–122
Bertolino F, Racugno W (1997) Is the intrinsic bayes factor intrinsic. Metron 54:5–15
Box GEP (1976) Science and statistics. J Am Stat Ass 71(356):791–799
Box GEP (1980) Sampling and bayes' inference in scientific modelling and robustness. J R Stat Soc Ser A (General) 143(4):383–430
Cressie N (2015) Statistics for spatial data. Wiley, New York
Csilléry K, Blum MG, Gaggiotti OE, François O (2010) Approximate Bayesian computation (ABC) in practice. Trends Ecol Evol 25(7):410–418
D'Agostino RB (1986) Goodness-of-fit-techniques, vol 68. CRC Press, Cambridge
Doksum KA, Lo AY (1990) Consistent and robust Bayes procedures for location based on partial information. Ann Stat 18:443–453
Fearnhead P, Prangle D (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J Roy Stat Soc B (Methodological) 74(3):419–474
Fisher RA (1925) Statistical methods for research workers. Genesis Publishing Pvt Ltd, Delhi
Fraser D, Rousseau J (2008) Studentization and deriving accurate p-values. Biometrika 95(1):1–16
Frazier DT, Robert CP, Rousseau J (2017) Model misspecification in abc: Consequences and diagnostics. arXiv:1708.01974
Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. Stat Sin 6:773–807
Gneiting T, Raftery AE (2005) Weather forecasting with ensemble methods. Science 310(5746):248–249
Gouriéroux C, Monfort A, Renault E (1993) Indirect inference. J Appl. Econom 8:S85–118
Granich RM, Gilks CF, Dye C, De Cock KM, Williams BG (2009) Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. Lancet 373(9657):48–57
Guttman I (1967) The use of the concept of a future observation in goodness-of-fit problems. J R Stat Soc: Ser B (Methodol) 29:83–100
Hickerson MJ, Meyer CP (2008) Testing comparative phylogeographic models of marine vicariance and dispersal using a hierarchical Bayesian approach. BMC Evol Biol 8(1):322
Hjort NL, Dahl FA, Steinbakk GH (2006) Post-processing posterior predictive p values. J Am Stat Assoc 101(475):1157–1174
Huber-Carol C, Balakrishnan N, Nikulin M, Mesbah M (2012) Goodness-of-fit tests and model validity. Springer, Berlin
Jasra A, Singh SS, Martin JS, McCoy E (2012) Filtering via approximate Bayesian computation. Stat Comput 22(6):1223–1237

Jiang B, Wu Ty, Zheng C, Wong WH (2015) Learning summary statistic for approximate bayesian computation via deep neural network. ArXiv e-prints arXiv:1510.02175
Johnson VE (2004) A bayesian χ^2 test for goodness-of-fit. Ann Stat 32(6):2361–2384. https://doi.org/10.1214/009053604000000616
Johnson VE (2007) Bayesian model assessment using pivotal quantities. Bayesian Analysis 2(4):719–733
Lemaire L, Jay F, Lee IH, Csilléry K, Blum MGB (2016) Goodness-of-fit statistics for approximate bayesian computation. Technical report, arXiv:1601.04096
Liepe J, Taylor H, Barnes CP, Huvet M, Bugeon L, Thorne T, Lamb JR, Dallman MJ, Stumpf MP (2012) Calibrating spatio-temporal models of leukocyte dynamics against in vivo live-imaging data using approximate bayesian computation. Integr Biol 4(3):335–345
Lintusaari J, Gutmann MU, Dutta R, Kaski S, Corander J (2017) Fundamentals and recent developments in approximate bayesian computation. Syst Biol 66(1):e66–e82
Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. Proc Nat Acad Sci USA 100:15324–8
Meng XL (1994) Posterior predictive p-values. Ann Stat 22(3):1142–1160
Norris JR, Allen RJ, Evan AT, Zelinka MD, O'Dell CW, Klein SA (2016) Evidence for climate change in the satellite cloud record. Nature. https://doi.org/10.1038/nature18273
Poon AF (2015) Phylodynamic inference with kernel ABC and its application to HIV epidemiology. Mol Biol Evol 32(9):2483–95
Prangle D (2015) Summary statistics in approximate bayesian computation. arXiv preprint arXiv:1512.05633
Ratmann O, Andrieu C, Wiuf 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):10576–10581
Robert C, Rousseau J (2002) A mixture approach to bayesian goodness of fit. Technical Report 9, Cahiers du CEREMADE
Robins JM, van der Vaart A, Ventura V (2000) Asymptotic distribution of p values in composite null models. J Am Stat Assoc 95(452):1143–1156
Rubin DB et al (1984) Bayesianly justifiable and relevant frequency calculations for the applies statistician. Ann Stat 12(4):1151–1172
Silk D, Filippi S, Stumpf MP (2013) Optimizing threshold-schedules for sequential approximate bayesian computation: applications to molecular systems. Stat Appl Genet Mol Biol 12(5):603–618
Sisson SA, Fan Y, Beaumont M (eds) (2017) Handbook of approximate Bayesian computation. Taylor & Francis, New York
Sisson SA, Fan Y, Beaumont M (2018) Handbook of approximate bayesian computation. Chapman and Hall/CRC, New York
Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK (1994) The epidemiology of tuberculosis in San Francisco—a population-based study using conventional and molecular methods. N Engl J Med 330(24):1703–1709
Stadler T (2011) Inferring epidemiological parameters on the basis of allele frequencies. Genetics 188(3):663–672
Stein M (1987) Large sample properties of simulations using Latin hypercube sampling. Technometrics 29(2):143–151
Tanaka MM, Francis AR, Luciani F, Sisson SA (2006) Using approximate bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173:1511–1520
Wegmann D, Leuenberger C, Excoffier L (2009) efficient approximate bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182(4):1207–1218