On the Identifiability of Transmission Dynamic Models for Infectious Diseases

Jarno Lintusaari,*,[1] Michael U. Gutmann,*,[†] Samuel Kaski,* and Jukka Corander[†]

*Helsinki Institute for Information Technology (HIIT) and Department of Computer Science, Aalto University, FI-00076 Aalto, Finland, and [†]Helsinki Institute for Information Technology (HIIT) and Department of Mathematics and Statistics, University of Helsinki, FI-00014 Helsinki, Finland

ABSTRACT Understanding the transmission dynamics of infectious diseases is important for both biological research and public health applications. It has been widely demonstrated that statistical modeling provides a firm basis for inferring relevant epidemiological quantities from incidence and molecular data. However, the complexity of transmission dynamic models presents two challenges: (1) the likelihood function of the models is generally not computable, and computationally intensive simulation-based inference methods need to be employed, and (2) the model may not be fully identifiable from the available data. While the first difficulty can be tackled by computational and algorithmic advances, the second obstacle is more fundamental. Identifiability issues may lead to inferences that are driven more by prior assumptions than by the data themselves. We consider a popular and relatively simple yet analytically intractable model for the spread of tuberculosis based on classical IS*6110* fingerprinting data. We report on the identifiability of the model, also presenting some methodological advances regarding the inference. Using likelihood approximations, we show that the reproductive value cannot be identified from the data available and that the posterior distributions obtained in previous work have likely been substantially dominated by the assumed prior distribution. Further, we show that the inferences are influenced by the assumed infectious population size, which generally has been kept fixed in previous work. We demonstrate that the infectious population size can be inferred if the remaining epidemiological parameters are already known with sufficient precision.

KEYWORDS approximate Bayesian computation; identifiability; intractable likelihood; transmission dynamic models; tuberculosis

STATISTICAL models for transmission dynamics are widely employed to answer fundamental questions about the infectivity of bacteria and viruses and to make predictions for intervention policies such as vaccines, decolonization, and case containment. For some types of infectious diseases, the complexity of the transmission process and the corresponding model, combined with the characteristics of the available data, makes the inference an intricate task. A particular difficulty arises from the need to use computationally intensive methods. Examples include the work by Tanaka *et al.* (2006), Sisson *et al.* (2007), Blum (2010), Stadler (2011), Fearnhead and Prangle (2012), Del Moral *et al.* (2012), Baragatti *et al.* (2013), and Albert *et al.* (2015), who considered the transmission dynamics of *Mycobacterium tuberculosis* based on IS*6110* fingerprinting data from tuberculosis (*M. tuberculosis*) cases in San Francisco reported earlier by Small *et al.* (1994). Except for Stadler (2011), who proposed an inference scheme based on likelihood and Markov chain Monte Carlo approximations, the above-mentioned studies employed and improved an approximate inference technique known as *approximate Bayesian computation* (ABC), which was originally introduced by Tavaré *et al.* (1997).

Although estimation of the epidemiological parameters of *M. tuberculosis* with the model of Tanaka *et al.* (2006) has been widely studied, concerns about identifiability have been raised. Originally, Tanaka *et al.* (2006) reported a wide credible interval for the reproductive value R, and later, Blum (2010) suggested that the data of Small *et al.* (1994) are not informative enough for confident estimation of R in the original setting by visually comparing the prior to the inferred posterior distribution. Stadler (2011) further questioned the accuracy of the ABC approach of Tanaka *et al.* (2006) after

doi: 10.1534/genetics.115.180034
Manuscript received July 3, 2015; accepted for publication December 29, 2015; published Early Online January 31, 2016.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.180034/-/DC1.
[1]Corresponding author: Department of Computer Science, P.O. Box 15400, Aalto University, FI-00076 Aalto, Finland. E-mail: jarno.lintusaari@aalto.fi

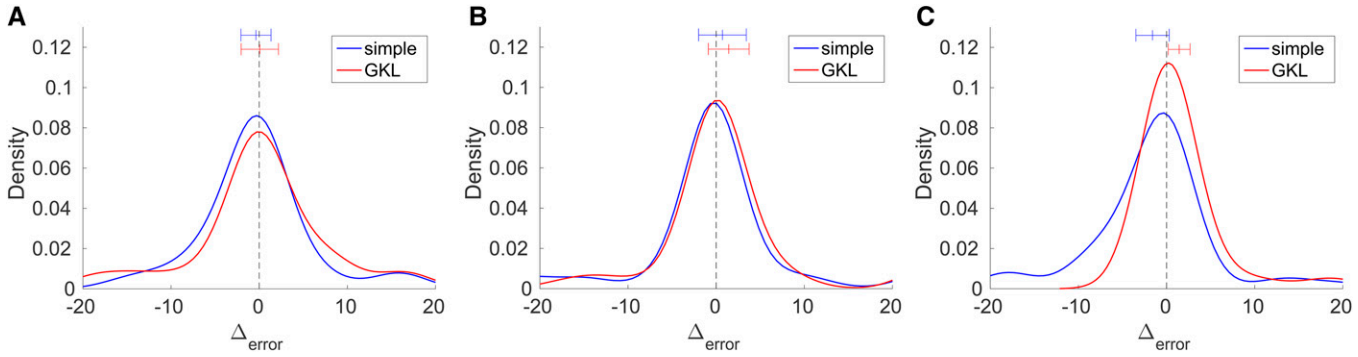


Figure 1 The distribution of the difference Δ_{error} between the relative estimation errors for the baseline and two alternative distance measures. A positive value of Δ_{error} indicates that the alternative method performs better. The intervals at the top show the estimated mean and the 95% confidence interval [(A) θ^S = (3.4, 0.69), (B) θ^S = (2.1, 0.57), and (C) θ^S = (1.7, 0.35)].

obtaining significantly different estimates with her method. This concern was later reconciled by Aandahl *et al.* (2014), who showed that the ABC method was valid and, moreover, also computationally more efficient. However, their confirmatory experiments with synthetic data only covered the setting with a single free parameter, which leaves the question about estimableness open for models with multiple free parameters.

In certain cases, the outcomes of ABC may not be accurate because the method includes several approximations and because practical algorithms can have several tuning parameters. Multiple validation methods thus have been developed. Many operate by using synthetic data generated with known parameter values in place of the observed data and comparing the inference results with the known parameter values. Wegmann *et al.* (2009), for example, used the absolute difference between the data-generating parameter values and the posterior point estimates to test whether the inferred posterior distribution is concentrated around the right parameter values; to test whether the spread of the distribution is not overly large or small, they suggested computing the proportion of times the credible interval contains the data-generating parameter values [see also the work by Prangle *et al.* (2014)]. Csilléry *et al.* (2012), however, recommended comparing the observed data with data simulated from the posterior predictive distribution. Further validation methods include confidence intervals, interquantile ranges, visualizations of the posterior distributions (*e.g.*, Tanaka *et al.* 2006; Toni *et al.* 2009; Blum 2010), and principal-components analysis (Toni *et al.* 2009; Cornuet *et al.* 2010).

Failure to pass some of the validation tests can occur for several reasons. The approximation may not be accurate enough, the settings of the inference algorithm could be the problem, or the issue could be deeper: the model may not be fully identifiable in the first place. In this paper, we assess the identifiability of the epidemiological model of Tanaka *et al.* (2006) for genotype data of the kind available from the San Francisco study of Small *et al.* (1994). Because the likelihood function indicates the informativeness of the data, we approach the identifiability problem directly by approximating the likelihood. Further, because previous ABC studies dealing with the same epidemiological model have assumed a fixed infectious population size for the data, we investigated how this choice influences estimation of the epidemiological parameters and whether it is possible to infer the population size from these kinds of genotype data without access to more extensive surveillance data about incidence. Because comparable data are widely considered for many different kinds of pathogens, the issue of model identifiability—and our approach of addressing it by approximating the likelihood function—is of wider general interest beyond the particular case discussed here.

Materials and Methods

Model for disease transmission

The model considered in the paper is a linear birth-death process with mutations (BDM) introduced by Tanaka *et al.* (2006). The process model is defined as follows: each infected individual, hereafter called *host*, carries the pathogen characterized by an allele at a single locus of its genome. The host transmits the pathogen and the corresponding allele with rate α and dies or recovers with rate δ. For simplicity, we call α the *birth rate* and δ the *death rate*. In addition, the pathogen mutates within the host at rate τ, resulting each time in a novel allele in the population of hosts (infinite-alleles model). When simulating the process, one begins with a single host and stops when either the population of hosts X reaches a predetermined size m or the pathogen goes extinct. The observation model assumes sampling of $n < m$ hosts from X without replacement. It was noted earlier by Stadler (2011) that owing to the time scaling of the model, at least one of the rate parameters must be fixed. Similar to many of the earlier studies, we use a time scale of 1 year and fix the mutation rate to τ = 0.198 per year throughout the experiments. Likewise, the infectious population size m is set to 10,000 unless otherwise stated.

The epidemiological parameters of interest in this study are the reproductive value R and the net transmission rate t.

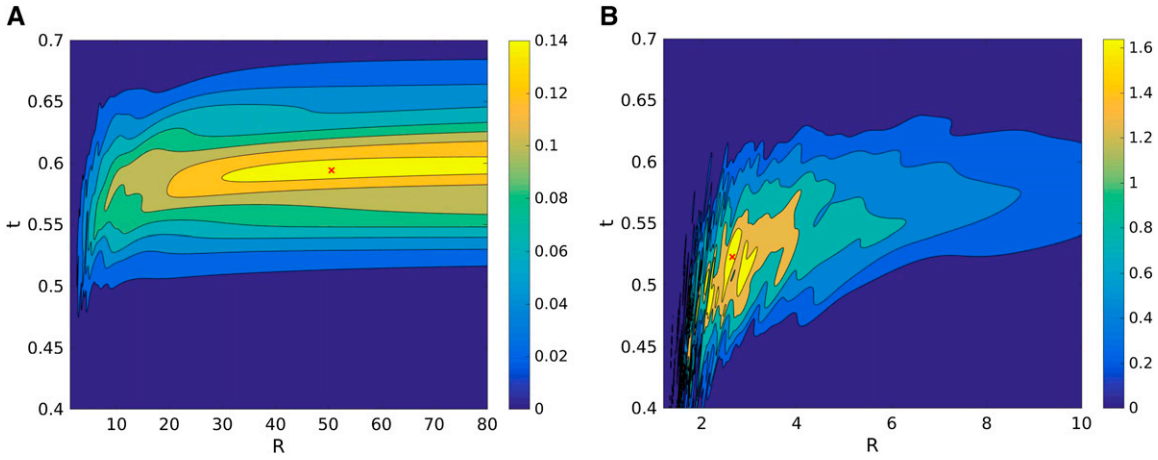


Figure 2 Likelihood (A) and posterior distribution when using a uniform prior in the (α, δ) space (B). The approximate posterior is obtained by multiplying the approximate likelihood by the prior on the grid. The red cross denotes the mode. Note the different scales of the *x*-axes.

In addition, we will consider the inference of the underlying infectious population size m given some estimate of R and t. In what follows, we will often use θ to denote the tuple (R, t). The epidemiological parameters R and t are in a one-to-one correspondence with the event-rate parameters of the BDM process: $R = \alpha/\delta$, $t = \alpha - \delta$, and $\delta = t/(R-1)$, $\alpha = tR/(R-1)$.

Data

The alleles of the pathogen carried by the n sampled hosts are summarized in the form of the allele vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_n) \in \mathbb{N}^n$, where element a_i equals the number of allele clusters of size i present in the sample. An allele cluster is a set of hosts having the same allele of the pathogen, and its size is the number of hosts that belong to the cluster. For example, the vector $\boldsymbol{a} = (4, 0, 1)$ implies that there are four singleton clusters and one cluster with three hosts in the sample. In other words, there are four different alleles, each found in only one host and one allele shared by three hosts. The size of $\mathbf{a}$ is defined as the sample size n, which can be written in terms of $\mathbf{a}$ as $n = \sum_i i a_i$.

For inference of the parameters, as in Tanaka *et al.* (2006), we used the San Francisco data of Small *et al.* (1994), which consist of an allele vector $\mathbf{a}^*$ of size $n = 473$. Its nonzero elements are $a_1^* = 282$, $a_2^* = 20$, $a_3^* = 13$, $a_4^* = 4$, $a_5^* = 2$, $a_8^* = 1$, $a_{10}^* = 1$, $a_{15}^* = 1$, $a_{23}^* = 1$, and $a_{30}^* = 1$.

Inference method

The likelihood function plays a central role in statistical inference. For the model considered in this paper, however, it cannot be expressed analytically in closed form (Tanaka *et al.* 2006). Tanaka *et al.* (2006) thus used approximate Bayesian computation (ABC) for the inference. We here approximate the likelihood function using kernel density estimation with a uniform kernel and a distance measure d, an approach that is related to ABC but that makes explicit its inherent approximations (Blum 2010; Gutmann and Corander 2015).

For a fixed value of the population size m, the likelihood function $L(\theta)$ is approximated as $L(\theta) \approx \hat{L}^N_{d,\varepsilon}(\theta)$, that is,

$$\hat{L}^N_{d,\varepsilon}(\theta) \propto \frac{1}{N} \sum_{i=1}^{N} 1_{[0,\varepsilon]}\left[d\left(\boldsymbol{a}^*, \boldsymbol{a}^{(i)}\right)\right] \tag{1}$$

where $1_{[0,\varepsilon]}(\cdot)$ is an indicator function that equals 1 if the distance d is less than a threshold ε and 0 otherwise, $\mathbf{a}^{(i)}$ is an allele vector simulated using parameter θ, and N is the number of such simulations performed (Gutmann and Corander 2015). The distance $d(\boldsymbol{a}^*, \boldsymbol{a})$ is a nonnegative function that measures the similarity between the observed allele vector $\boldsymbol{a}^*$ and the simulated allele vector $\mathbf{a}$. Possible choices of d are discussed later. Equation 1 means that the likelihood is approximated by the fraction of times the simulated allele vector $\mathbf{a}$ is within distance ε of the observed allele vector $\boldsymbol{a}^*$. The approximate likelihood function for inference of the population size m for fixed θ is defined in an analogous manner.

While there are several variants of the inference procedure of ABC, they are essentially built out of sampling candidate parameter values θ and retaining those for which the distance $d(\boldsymbol{a}^*, \boldsymbol{a})$ is less than the threshold ε. Under certain conditions, the retained parameters correspond to samples from the posterior. In ABC, the approximate likelihood function is generally never explicitly constructed but is implicitly represented by the sampled parameter values for which the simulated data are close to the observed data. By contrast, in this paper, the likelihood function is explicitly approximated using Equation 1, which is important because it provides information about the identifiability of the parameters.

Because the parameter space is low dimensional, we can evaluate the approximate likelihood function by varying the parameters on a grid. Several grids were created based on the different inference tasks considered:

For inference of θ, we formed a 137×120 evenly spaced grid over a subspace $\Delta_\alpha \times \Delta_\delta = [0.3, 2] \times [0.0125, 1.5]$ of

Table 1 Effect of the prior on the posterior mode, mean, and credible interval of the epidemiological parameters of *M. tuberculosis* for the San Francisco data

Prior	Parameter	Mode	Mean	Credible interval (95%)
Uniform prior in (R, t)	R	50.6	44.1	(9.5, 80.0)
	t	0.59	0.59	(0.51, 0.67)
Uniform prior in (α, δ)	R	2.7	10.5	(1.4, 39.0)
	t	0.52	0.56	(0.46, 0.66)

the (α, δ) BDM parameter space. At each node of the grid, $N = 3000$ allele vectors were simulated to approximate the likelihood in that location by using Equation 1.

To study the effect of m on R, four different grids (one for each value of m) with 41 nodes were created, which were subsets of the interval $\Delta_\alpha = [0.53, 1.4]$ and had $N = 3000$ simulations in each node.

For evaluation of the feasibility of inference of m, four grids also were used, two with size 50 and two with size 51, $N = 3000$, all grids being subsets of the interval $\Delta_N = [500, 25, 000]$.

For the final inference of the population size m, we used two grids, one with 15 nodes and the other with 26 nodes, both subsets of $\Delta_N = [500, 35, 000]$, with $N = 30{,}000$ simulated allele vectors in each node.

The amount of simulated data in the preceding grids was selected to ensure the stability of the likelihood approximations. Results for stabilization of the marginal likelihoods for the grid over the (α, δ) parameter space are shown in the Supporting Information, File S1. The dimensions of the two-dimensional grid were chosen such that the grid includes the modes of the likelihoods used in the experiments. The one-dimensional grids were chosen so that they included all the significant mass of the approximated likelihood functions and thus numerical computation of the posterior means is possible.

Distance measures used to approximate the likelihood

The likelihood approximation in Equation 1 relies on a distance measure $d(\boldsymbol{a}*, \boldsymbol{a})$ between the observed sample $\boldsymbol{a}*$ and the sample $\mathbf{a}$ produced by the simulation process with the parameter vector $\boldsymbol{\theta}$. Consequently, different distance measures may lead to different estimation results depending on how much information about the generating process they are able to capture from the data. It is thus natural to ask which distance measures would be optimal for a model of the type considered here.

In the limit of $\varepsilon \to 0$ and $N \to \infty$, one can easily define distance measures $d(\boldsymbol{a}*, \boldsymbol{a})$, which lead to exact likelihoods $L(\theta)$. The only requirement is that $d(\boldsymbol{a}*, \boldsymbol{a}) = 0$ if and only if $\boldsymbol{a} = \boldsymbol{a}*$. In practice, however, too small thresholds ε and a very large number of simulations N are computationally not feasible. Therefore, one has to rely on likelihood approximations $\widehat{L}_{\mathrm{d},\varepsilon}(\theta)$ dictated by the distance measure d, a nonzero threshold value ε, and a finite N.

Given a fixed number of simulations N, differences in the quality of the approximations arise from the ability of the distance measures d to produce approximate likelihood functions that are as close as possible to $L(\theta)$. Because the likelihood $L(\theta)$ is unknown in the first place, evaluation of the approximations is challenging. One method to evaluate the goodness of an approximate likelihood function is to measure the goodness of the corresponding estimates using synthetic observed data $\boldsymbol{a}^{\mathrm{s}}$ where the data-generating parameters θ^{s} are known. In particular, the mode of the likelihood approximation, on average, should be near θ^{s} if the sample $\boldsymbol{a}^{\mathrm{s}}$ is, in general, informative enough.

In this study, we evaluate the performance of three different distance measures. The baseline distance measure is the one introduced by Tanaka *et al.* (2006) and is defined as

$$d_{\mathrm{base}}\left(\mathbf{a}, \mathbf{a}'\right) = \frac{1}{n}|g_{\mathrm{a}} - g_{a'}| + |H_{\mathrm{a}} - H_{a'}| \tag{2}$$

where $H_{\mathrm{a}} = 1 - \sum_{\mathrm{i}} a_{\mathrm{i}}(i/n)^2$ is a gene diversity summary statistic, and $g_{\mathrm{a}} = \sum_{\mathrm{i}} a_{\mathrm{i}}$ is the number of distinct alleles in $\mathbf{a}$ or, in other words, the total number of clusters present in the data.

The second distance measure, called *simple*, is defined as

$$d_{sim}\left(\boldsymbol{a}, \boldsymbol{a}'\right) = \left|a_1 - a_1'\right| + |M_a - M_{a'}| \tag{3}$$

where $\mathbf{a}$ and $\mathbf{a}'$ are allele vectors, and $M_{\mathrm{a}} = \max\{i | a_{\mathrm{i}} \neq 0\}$ is the largest cluster size in $\mathbf{a}$. This measure compares the number of singleton clusters in the sample and the sizes of the largest clusters. It can be seen as a simplified version of the baseline distance measure by excluding some information.

The third distance measure d_{GKL} is motivated by the observation that element a_{i} of the allele vector $\mathbf{a}$ is proportional to the frequency of occurrence of cluster size i in the data, with the proportionality factor $g_{\mathrm{a}} = \sum_{\mathrm{i}} a_{\mathrm{i}}$ being equal to the total number of clusters present. We thus can consider a_{i} to indicate the probability of observing cluster size i in the data. This probabilistic interpretation of vector $\mathbf{a}$ opens up the possibility of using known divergence measures from probability theory to gauge the similarity between the two vectors $\mathbf{a}$ and $\mathbf{a}'$. In more detail, because both $\mathbf{a}$ and $\mathbf{a}'$ correspond to estimated probabilities, we will discount small differences in the number of clusters of similar size, which may be present owing to chance alone, and represent $\mathbf{a}$ and $\mathbf{a}'$ by smooth continuous functions f_{a} and $f_{\mathrm{a}'}$ approximating the two vectors in the discrete locations i; that is, $f_a(i) \approx a_i$ and $f_{a'}(i) \approx a_i'$. Because $\mathbf{a}$ and $\mathbf{a}'$ correspond to unnormalized probability vectors, f_{a} and $f_{\mathrm{a}'}$ correspond to unnormalized probability densities, and we can assess their similarity by the generalized Kullback-Leibler (GKL) divergence

$$d_{\mathrm{GKL}}(f_a, f_{a'}) = \int f_a(x)[\log f_a(x) - \log f_{a'}(x)]dx - \int f_a(x)dx + \int f_{a'}(x)dx \tag{4}$$

The GKL divergence belongs to the family of Bregman divergences (Bregman 1967), which have a number of desirable

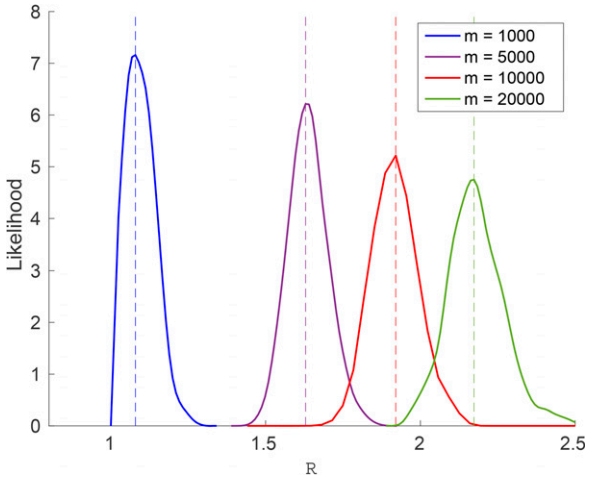


Figure 3 Likelihoods of the reproductive value R with fixed death rate $\delta = 0.52$ and four alternative population sizes m using the San Francisco data. The vertical lines indicate the modes of the likelihoods.

properties, *e.g.*, nonnegativity and being equal to 0 if $f_a = f_{a'}$, as well as useful applications (see, *e.g.*, Collins *et al.* 2002; Frigyik *et al.* 2008; Gutmann and Hirayama 2011). Like the ordinary Kullback-Leibler divergence, the GKL is also asymmetrical. However the integrals over f_a and $f_{a'}$ need not equal 1. In fact, by construction of f_a and $f_{a'}$, the difference between their integrals assesses the difference between the total number of clusters g_a and $g_{a'}$, much as in d_{base}. Conceptually, our probabilistic interpretation of **a** boils down to using the cluster size as a summary statistic and comparing its probability distribution for observed and simulated data in a nonparametric way.

Results

Evaluations of the distance measures

To compare the performance of the alternative distance measures d_{sim} and d_{GKL} to the baseline distance measure d_{base} in likelihood approximation, the difference Δ_{error} between the relative errors in their respective estimates was computed. A value $\Delta_{\text{error}} > 0$ indicates that the relative error is larger for the baseline compared with the alternative method, in which case the alternative method would be preferable.

The relative error was defined as $\sum_i(|\theta_i^s - \theta_i|/\theta_i^s)$, where θ^s is the true data-generating parameter value used to simulate the synthetic data $\boldsymbol{a}^s$, and $\widehat{\theta}$ is the estimate obtained by maximizing the approximate likelihood. Maximization was performed in a simple way, by searching for the maximal value over the two-dimensional grid. To compute the relative error, a total of 50 synthetic observations $\boldsymbol{a}^s$ were simulated using θ^s, and the likelihood function was approximated with the grid for each observation $\boldsymbol{a}^s$. The threshold ε was set for each distance measure to the value minimizing the sum of the relative errors over the 50 trials.

We considered three different setups for the data-generating parameter value θ^s. In the first setup, the value of θ^s was set to the estimate (3.4, 0.69) of Tanaka *et al.* (2006). The second setup uses the estimate (2.1, 0.57) from Aandahl *et al.* (2014). To further see whether the values of the actual epidemiological parameters had an effect on estimation accuracy, we considered one more setup in which both the reproductive value R and the transmission rate t of the first setup were divided by 2.

Figure 1 shows the resulting distribution of Δ_{error} for the comparison between d_{sim} and d_{base} (blue curve) and the comparison between d_{GKL} and d_{base} (red curve). The simple distance measure performs slightly worse than the baseline, although the difference is not significant in any of the setups (the null hypothesis of a 0 mean of Δ_{error} cannot be rejected). This means that reducing the distance measure d_{base} of Tanaka *et al.* (2006) to the simpler d_{sim} does not cause a notable reduction in estimation accuracy. The GKL distance d_{GKL}, however, performs slightly better than the baseline, and the difference is significant in the last setup (the 0 mean hypothesis of Δ_{error} can be rejected at $P = 0.0237$). It should be noted, nevertheless, that the absolute errors tend to be rather large with all the distance measures, as shown in File S2.

Because d_{GKL} was found to perform at least as well as the other measures, it was used in the remaining parts of this paper unless stated otherwise. Furthermore, for simplicity, we will often drop the qualifier "approximate" and use "approximate likelihood" and "likelihood" interchangeably.

Relative effects of likelihood and prior on the posterior

The simulator operates genuinely in the (α, δ) space, where α and δ are the birth rate and the death rate in the model, respectively. Accordingly, all the ABC studies so far have assumed an uninformative (uniform) prior for the region $0 < \delta < \alpha$ in the Bayesian framework. The law of transformation of random variables implies that choosing a uniform prior for (α, δ) is equivalent to choosing the following prior for the epidemiological parameters (R, t):

$$p(R,t) \propto \begin{cases} \frac{t}{(R-1)^2} & \text{if} \quad R > 1, t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

The prior is shown in File S3. The formula and the figure show that its probability mass is concentrated on small values of the reproductive value R.

Figure 2A shows the likelihood function of (R, t) for the San Francisco data of Small *et al.* (1994) on the rectangle defined by $1.2 < R < 80$ and $0.4 < t < 0.7$. We used the same grid as in the preceding section with threshold equal to the 10^{-4} quantile of all the distances. The likelihood function is flat over large areas of the parameter space, and many values of R are equally likely, which means that the data are not informative enough to identify the parameter R of the model.

Figure 2B shows the posterior distribution of (R, t) for the prior in Equation 5, *i.e.*, for the uniform prior on (α, δ). It can

Table 2 Mean estimated population size $\overline{m}$ for 50 trials and the respective confidence intervals (CIs) under two alternative configurations of *R* and population size *m*

	m	$\overline{m}$	CI (95%)
$R = 2.1$	1,000	1,051	(990, 1,112)
	10,000	10,020	(9,380, 10,660)
$R = 1.1$	1,000	1,007	(976, 1,038)
	10,000	10,510	(10,003, 11,017)

Death rate was fixed to $\delta = 0.52$; results are for synthetic data.

be seen that the prior leads to a substantial shift of the probability mass toward the lower end of values of R. The difference between the modes of the likelihood and the posterior is striking: $R = 50.6$ *vs.* $R = 2.7$, as shown in Table 1. The table also shows the posterior means and credible intervals for the case that the likelihood is interpreted as the posterior distribution with a uniform prior in the (R, t) space. It should be noted that the upper value of the credible intervals for R is an artifact of limiting our computation to values of $R < 80$. The shape of the likelihood in Figure 2A suggests that computations with larger values of R would lead to a corresponding increase in the credibility intervals. The results mean that the prior dominates the posterior distribution, which confirms previous findings by Blum (2010).

Effect of the infectious population size

The preceding sections suggest that data of the kind considered in the San Francisco study do not carry enough information for accurate inference of R but that the prior plays a major role. To make the inference of R possible, Aandahl *et al.* (2014) obtained an estimate 0.52 for the death rate δ by summing the rates of self-cure, death from causes other than tuberculosis, and death from untreated tuberculosis as estimated in other studies. The parameter δ then either was fixed to this value or an informative prior centered on it was used. Also previously, in all the corresponding ABC studies, the infectious population size m has been fixed, commonly to the value $m = 10{,}000$ following Tanaka *et al.* (2006). We were thus interested in whether reducing the infectious population size to a smaller, possibly more realistic number has an influence on the estimated value of R. To ease the comparison with previous studies, we used the distance d_{base} in the likelihood approximation. The thresholds were set to the 5×10^{-3} quantile of the distances in the respective grids.

Figure 3 shows the likelihoods of R for $m \in \{1000, 5000, 10,000, 20,000\}$ using the San Francisco data (Small *et al.* 1994) and $\delta = 0.52$. The difference in location of the likelihoods is clear, with modes shifting to the right with increasing m. The mode locations were (1.1, 1.6, 1.9, 2.2) given in the order of increasing m. Posterior means with uniform prior over the support of the likelihoods were the same. The results show that the assumed infectious population size m affects the inference of R.

Inference of the infectious population size

The observed effect of the infectious population size m on the inference of R means that there is some (statistical) dependency between m and R. This suggests that it might be possible to infer the size of the underlying infectious population from the data when R and δ are known. Alternatively, given the relationship between the parameters, knowing any two of α, δ, R, or t would be sufficient.

We fixed $\delta = 0.52$ as earlier and considered two alternative configurations: $R = 2.1$ and $R = 1.1$. The former is also the estimate of Aandahl *et al.* (2014). To test whether inference of the infectious population size parameter is possible, we ran 50 trials with synthetic data as before and determined each time the parameter value on the grid which maximized the approximate likelihood. The thresholds ε were set to the 10^{-2} quantile of the distances. Table 2 shows the results of these experiments: the estimated population sizes are reasonable, and the actual population size m is covered by the 95% confidence interval of the mean in all but one of the cases. Only in the last case is the true m just barely outside the interval. These results thus illustrate that estimation of m is possible, provided that reliable information is available about the other epidemiological parameters.

We next estimated the infectious population size for *M. tuberculosis* in the San Francisco area during the time the data of Small *et al.* (1994) were collected. Threshold ε was selected as the 10^{-3} quantile of the distances. Figure 4 shows the likelihood functions for two different values of R: $R = 1.1$ and $R = 2.1$ (with $\delta = 0.52$ as before). Assuming that $R = 1.1$ produced the posterior mean $\widehat{m} = 2106$ and the 95% credible interval (1166, 3226) with uniform prior over the support of the likelihood function. Assuming that $R = 2.1$, however, produced the posterior mean $\widehat{m} = 21,108$ and the 95% credible interval (13,408, 30,155).

Discussion

Statistical inference plays an important role in the study of the transmission dynamics of infectious diseases. In this paper, we considered some of the challenges arising from model identifiability and from the expert choices necessary for approximate inference using the relatively simple yet analytically intractable model of Tanaka *et al.* (2006) for the transmission dynamics of *M. tuberculosis*. It is reasonable to assume that these problems persist for more complex transmission models unless molecular and epidemiological data are detailed enough to mitigate their effect.

Given the intractability of the transmission model, an approximate inference approach was used that belongs to the framework of approximate Bayesian computation (ABC), which relies on a distance measure gauging similarity between observed and simulated data. An alternative approach was presented by Stadler (2011) based on likelihood and Markov chain Monte Carlo (MCMC) approximations. Also

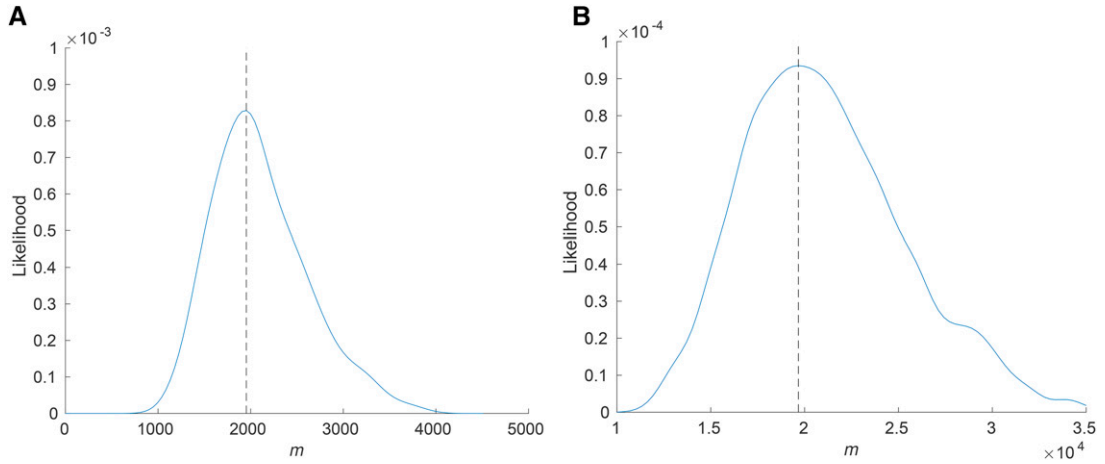


Figure 4 Likelihoods of population size m with fixed death rate $\delta = 0.52$ and two alternative values of R using the San Francisco data: (A) $R = 1.1$ and (B) $R = 2.1$. Vertical lines indicate the modes of the likelihoods.

in this approach, the model has an analytically intractable likelihood function, meaning that the probability of the observed data as a function of the parameters of interest is not available in closed form. The reason for the intractability is the presence of unobserved variables (forming the transmission tree in Stadler's model), which are integrated out using MCMC. While feasible for a small number of variables, this technique runs into problems when the number of unobserved variables is large (see, *e.g.*, Green *et al.* 2015). The problems manifest in the form of increased convergence issues of the Markov chain. Aandahl *et al.* (2014) resolved such issues in the approach of Stadler (2011) but concluded that the ABC method has a similar accuracy and better computational efficiency than the amended version.

In all the tests comparing distance measures in ABC, the GKL distance measure attained a lower or equal estimation error compared to the baseline measure introduced by Tanaka *et al.* (2006), suggesting that one can reduce the estimation error to some degree by the choice of the distance measure only (see also Fearnhead and Prangle 2012). A possible disadvantage of the approach with the GKL distance is a slight increase in computational cost. Although not an issue in our study, where the computational bottleneck was the simulator, this may be an issue when the inference procedure is repeated for a large number of data sets or when the simulation of the data is not computationally expensive. While the measure used by Tanaka *et al.* (2006) has some clear biological meaning, the GKL distance is based on a more general information-theoretical construction. The observed increase in performance is thus interesting because, in ABC, distances are usually strongly based on application-specific knowledge, even though some exceptions do exist (Gutmann *et al.* 2014).

Our explicit construction of the approximate likelihood function put on display the difficulties in the estimation of R when inferring both the reproductive value R and the transmission rate t (Tanaka *et al.* 2006; Blum 2010). The credible intervals for R were (1.4, 38.0) and (9.5, 80.0) when using a uniform prior over the (α, δ) and (R, t) space, respectively. The large upper end points of the credible intervals reflect the extreme flatness of the approximate likelihood function with respect to the reproductive value parameter. An uninformative prior over the (α, δ) space has been, to our knowledge, the standard choice in all the related ABC studies. Comparing the posterior with the approximate likelihood function shows how significantly the prior contributed to the posterior, altering the shape of the likelihood function greatly. The results highlight the usefulness of likelihood approximation as an identifiability check when performing inference for models with intractable likelihoods.

A uniform prior is usually considered uninformative, so it may seem paradoxical that the prior had such a strong influence on the posterior. The apparent paradox is readily resolved by noting that the uniform prior was not imposed on the actual epidemiological quantities of interest but on a nonlinear transformation of them.

The standard assumption in previous ABC studies concerned with the tuberculosis data from San Francisco (Small *et al.* 1994) has been that the infectious population size m equals 10,000 individuals. We showed that the infectious population size influences the estimation of R, with the estimate of R increasing with m. Assuming that $m = 10{,}000$, the posterior mean was $\widehat{R} = 1.9$. Smaller assumed population sizes $m = 1000$ and $m = 5000$ yielded smaller estimates $\widehat{R} = 1.1$ and $\widehat{R} = 1.6$, respectively, while larger assumed population size $m = 20{,}000$ increased the estimate to $\widehat{R} = 2.2$. The corresponding likelihoods (proportional to posterior distributions with uniform prior) with varying m were clearly distinct from each other, as seen in Figure 3.

Taking advantage of the dependency between R and m, we showed that it is possible to estimate the infectious population size m when R and δ are known. Using the estimate $\delta = 0.52$ (Aandahl *et al.* 2014) and assuming either $R = 2.1$ or $R = 1.1$, the posterior means of m were 21,100 or 2100, respectively, for the San Francisco data of Small *et al.*

(1994). Further biological expertise can be used to assess the reasonability of different inferred population sizes in a comparable modeling setting.

We noticed that for small values of m, the generative model was unable to produce data with a similar number of distinct alleles as the San Francisco data while also containing large clusters, supporting the observation of Tanaka *et al.* (2006) that $m = 1000$ does not result in an appropriate level of diversity. In the San Francisco data, large clusters were present originating from groups of people with conditions affecting the immune system, *e.g.*, AIDS. Among such groups, the transmission rate of *M. tuberculosis* can be expected to be notably higher and thus rapidly producing large clusters. The simple model, however, does not account for these situations. In future work it would be interesting to consider approximate inference for models with possibly heterogeneous reproductive values that depend on auxiliary epidemiological data (Bacaër *et al.* 2008). However, given the apparent identifiability issues with the simple model studied here, it would be of utmost importance to ensure that the molecular and epidemiological data are jointly informative enough to perform reliable inferences.

Acknowledgments

We acknowledge the computational resources provided by the Aalto Science–IT Project. This research was funded by the Academy of Finland [Finnish Centre of Excellence in Computational Inference Research (COIN)].

Literature Cited

Aandahl, R. Z., T. Stadler, S. A. Sisson, and M. M. Tanaka, 2014 Exact *vs.* approximate computation: reconciling different estimates of Mycobacterium tuberculosis epidemiological parameters. Genetics 196: 1227–1230.

Albert, C., H. Knsch, and A. Scheidegger, 2015 A simulated annealing approach to approximate Bayes computations. Stat. Comput. 25: 1217–1232.

Bacaër, N., R. Ouifki, C. Pretorius, R. Wood, and B. Williams, 2008 Modeling the joint epidemics of TB and HIV in a South African township. J. Math. Biol. 57: 557–593.

Baragatti, M., A. Grimaud, and D. Pommeret, 2013 Likelihood-free parallel tempering. Stat. Comput. 23: 535–549.

Blum, M. G. B., 2010 Approximate Bayesian computation: a nonparametric perspective. J. Am. Stat. Assoc. 105: 1178–1187.

Bregman, L., 1967 The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. 7: 200–217.

Collins, M., R. Schapire, and Y. Singer, 2002 Logistic regression, AdaBoost and Bregman distances. Mach. Learn. 48: 253–285.

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

Csilléry, K., O. Franois, and M. G. B. Blum, 2012 abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3: 475–479.

Del Moral, P., A. Doucet, and A. Jasra, 2012 An adaptive sequential Monte Carlo method for approximate Bayesian computation. Stat. Comput. 22: 1009–1020.

Fearnhead, P., and D. Prangle, 2012 Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J. R. Stat. Soc. B Stat. Methodol. 74: 419–474.

Frigyik, B., S. Srivastava, and M. Gupta, 2008 Functional Bregman divergence and Bayesian estimation of distributions. IEEE Trans. Inf. Theory 54: 5130–5139.

Green, P., K. Latuszynski, M. Pereyra, and C. P. Robert, 2015 Bayesian computation: a summary of the current state, and samples backwards and forwards. Stat. Comput. 25: 835–862.

Gutmann, M., and J. Corander, 2015 Bayesian optimization for likelihood-free inference of simulator-based statistical models. J. Mach. Learn. Res. (in press) arXiv:1501.03291 [stat.ML].

Gutmann, M., R. Dutta, S. Kaski, and J. Corander, 2014 Statistical inference of intractable generative models via classification. arXiv:1407.4981 [stat. CO].

Gutmann, M. U., and J. Hirayama, 2011 Bregman divergence as general framework to estimate unnormalized statistical models, pp. 283–290 in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, Belfast, Northern Ireland, UK.

Prangle, D., M. G. B. Blum, G. Popovic, and S. A. Sisson, 2014 Diagnostic tools for approximate Bayesian computation using the coverage property. Aust. N.Z. J. Stat. 56: 309–329.

Sisson, S. A., Y. Fan, and M. M. Tanaka, 2007 Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA 104: 1760–1765.

Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet *et al.*, 1994 The epidemiology of tuberculosis in San Francisco: a population-based study using conventional and molecular methods. N. Engl. J. Med. 330: 1703–1709.

Stadler, T., 2011 Inferring epidemiological parameters on the basis of allele frequencies. Genetics 188: 663–672.

Tanaka, M. M., A. R. Francis, F. Luciani, and S. A. Sisson, 2006 Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173: 1511–1520.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf, 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface 6: 187–202.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182: 129–141.

Communicating editor: W. Valdar

GENETICS

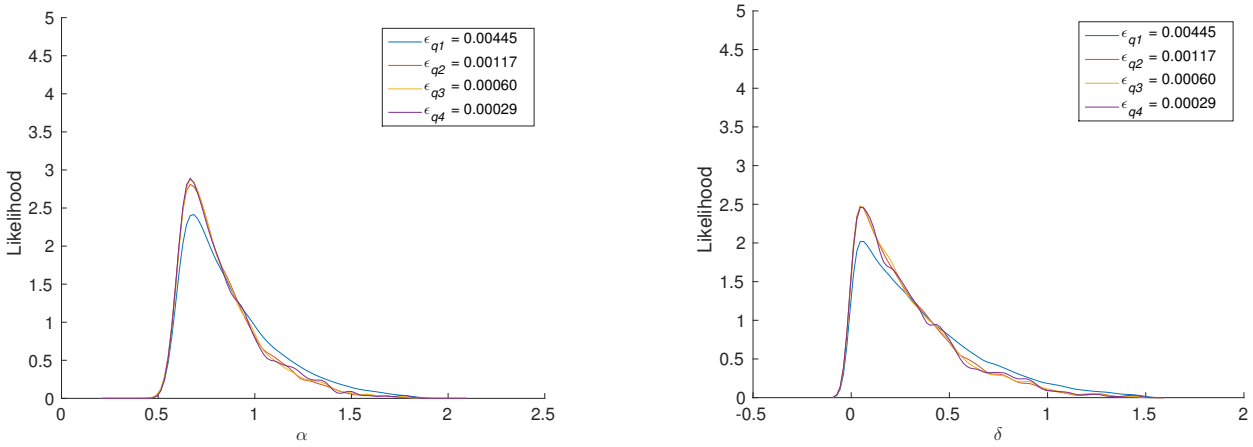
Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.180034/-/DC1

On the Identifiability of Transmission Dynamic Models for Infectious Diseases

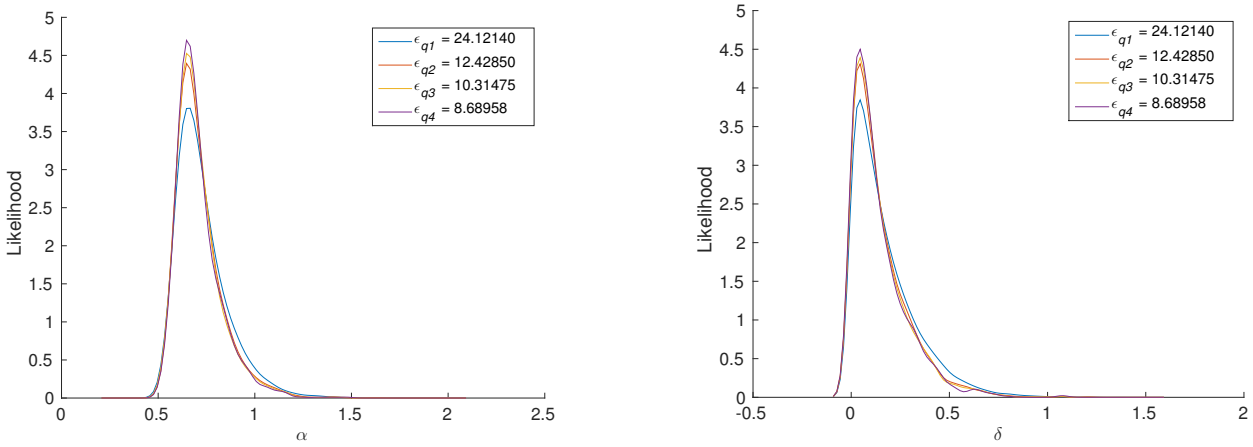
Jarno Lintusaari, Michael U. Gutmann, Samuel Kaski, and Jukka Corander

DOI: 10.1534/genetics.115.180034

FILE S1: STABILIZATION OF THE APPROXIMATE LIKELIHOODS



(a) Tanaka distance measure

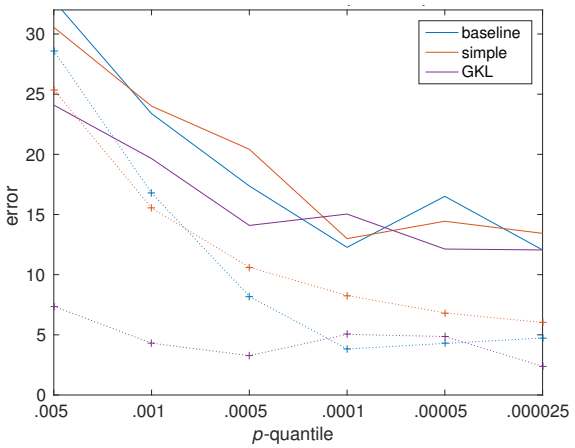


(b) GKL distance measure

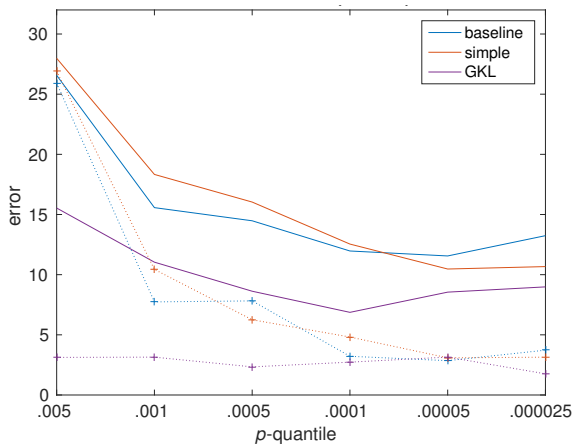
Figure S1: Stabilization (convergence) of the approximate marginal likelihoods for decreasing thresholds. The thresholds ϵ_{qi} were obtained from the quantiles $(q_1, q_2, q_3, q_4) = (0.001, 0.0001, 0.00005, 0.000025)$ of the distribution of the distances.

FILE S2: EVALUATIONS OF THE DISTANCE MEASURES

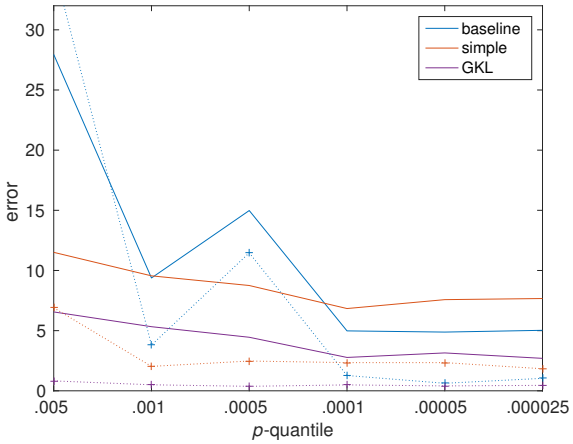
Mean and median absolute errors:



(a) $\theta^s = (3.4, 0.69)$



(b) $\theta^s = (2.1, 0.57)$

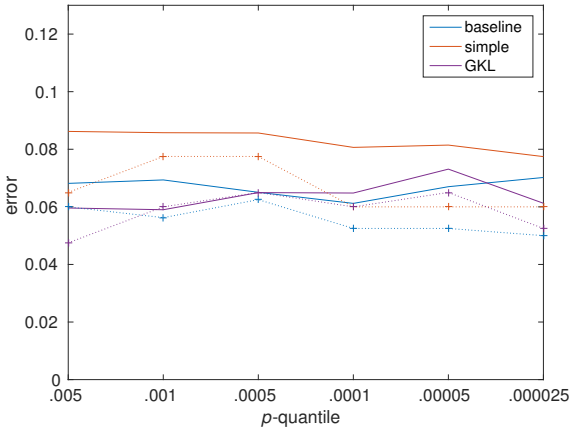


(c) $\theta^s = (1.7, 0.35)$

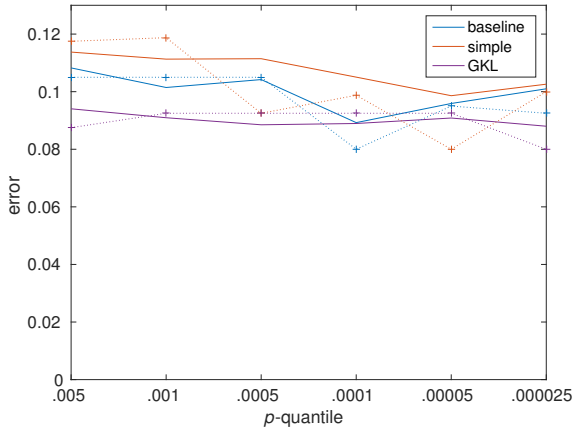
Figure S2: Mean (solid lines) and median (dotted lines) errors of the approximate maximum likelihood estimates with the three different distance measures $\mathrm{d}_{\mathrm{base}}$ (blue lines), $\mathrm{d}_{\mathrm{sim}}$ (red lines), $\mathrm{d}_{\mathrm{GKL}}$ (purple lines) and a decreasing threshold ϵ given by the p-quantile. The error is the L_1 distance between R^s and $\hat{R}$, that is, $|R^s - \hat{R}|$.

The rather large difference between the mean and median errors of R in Figure S2 indicates that there are some large errors which pull the mean error up. This is mostly due the tendency of $R = \alpha/\delta$ to be large when the estimate of the death rate δ is small. Although the errors for R in Figure S2 are large, the small errors in Figure S3 indicate that the estimation

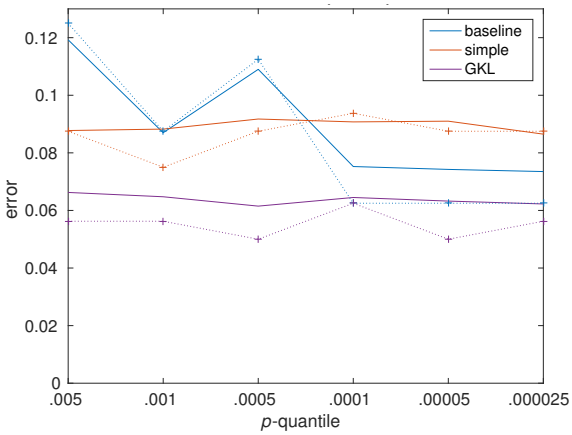
of the transmission rate t can still be done accurately and is not affected by the error-prone estimate of R.



(a) $\theta^s = (3.4, 0.69)$



(b) $\theta^s = (2.1, 0.57)$

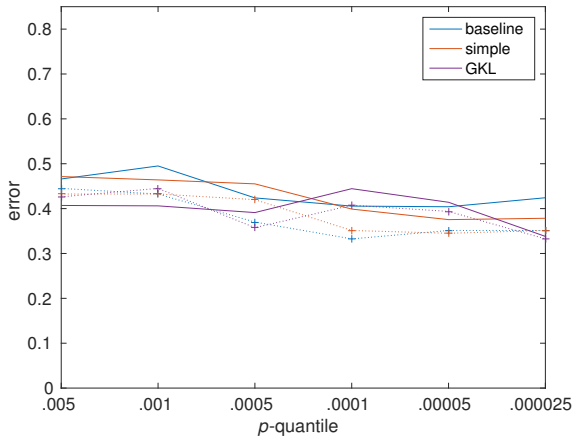


(c) $\theta^s = (1.7, 0.35)$

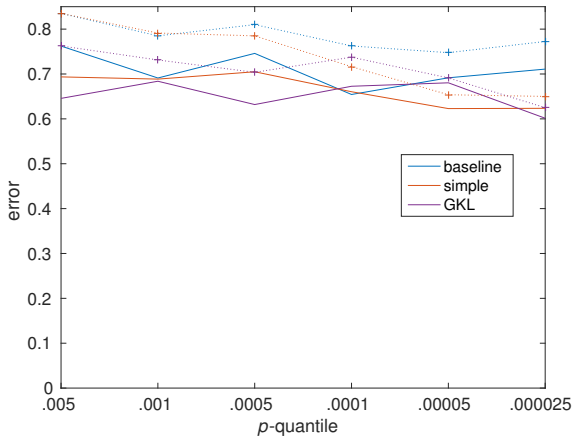
Figure S3: Mean and median errors of the approximate maximum likelihood estimate of t. Visualization is as in Figure S2. The error is the L_1 distance between t^s and $\hat{t}$, that is, $|t^s - \hat{t}|$.

Absolute errors in the rate parameter space:

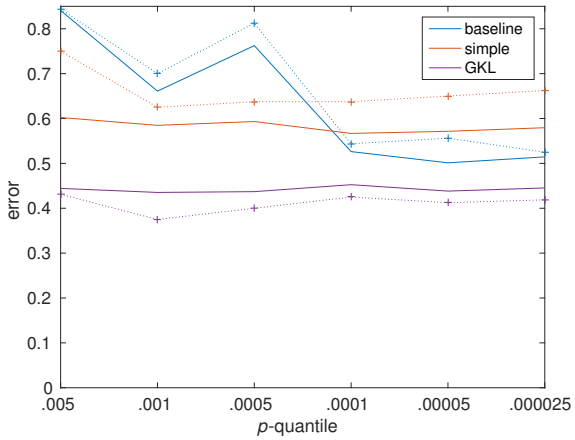
We noticed a general tendency of acquiring small estimates of δ irrespective of the setup. This is a plausible reason for the slightly reduced errors in Figure S4 (a) compared to the other setups as δ^* is the smallest in that setup.



(a) $\phi^* = (0.98, 0.29)$



(b) $\phi^* = (1.09, 0.52)$



(c) $\phi^* = (0.85, 0.50)$

Figure S4: Mean and median errors in the estimated $\phi = (\alpha, \delta)$. Visualization and setup is as in Figure S2. The error is the L_1 distance between the vectors ϕ^* and $\hat{\phi}$.

FILE S3: TRANSFORMED UNIFORM PRIOR

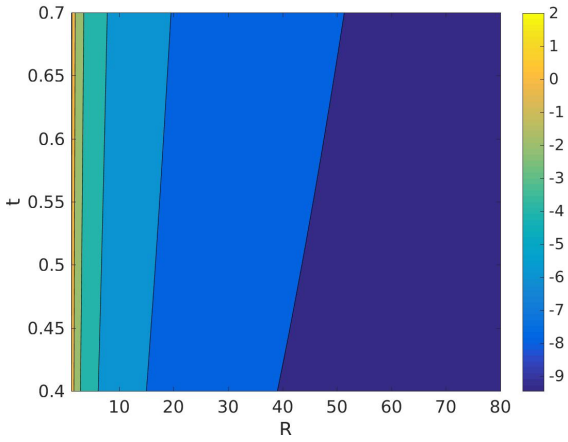


Figure S5: Visualization of the logarithm of the prior in Equation 5 corresponding to the uniform prior on (α, δ). Note the concentration of probability mass on small R.