On the identifiability of transmission dynamic models for infectious disease

Jarno Lintusaari[*], Michael U. Gutmann[*§], Samuel Kaski[*], Jukka Corander[§]

July 3, 2015

[*]Helsinki Institute for Information Technology HIIT,
Department of Computer Science, Aalto University

[§]Helsinki Institute for Information Technology HIIT,
Department of Mathematics and Statistics, University of Helsinki

Running Head: Transmission model identifiability

Key Words: Identifiability, transmission dynamic models, intractable likelihood, approximate Bayesian computation, tuberculosis

Corresponding Author:

Jarno Lintusaari

Department of Computer Science

Aalto University

00076 Espoo

Finland

jarno.lintusaari@aalto.fi

Abstract

Understanding the transmission dynamics of infectious disease 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 causes two challenges: Firstly, computationally intensive simulation-based inference methods need to be employed. Secondly, 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 which are more driven by the prior assumptions than the data themselves. Moreover, problems with identifiability may be hard to recognize, in particular when simulation-based inference methods are used. We here consider a popular and relatively simple, yet analytically intractable model for the spread of tuberculosis based on classical IS6110 fingerprinting data. We report on the identifiability of the model, presenting also some methodological advances regarding the inference. It is shown that the model does not allow for accurate inference about the reproductive value, and that the posterior distributions obtained in previous work have likely been substantially dominated by the assumed prior distribution. It is further shown that the inferences are influenced by the assumed infectious population size, which has generally 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.

Statistical models for transmission dynamics are very widely employed to answer fundamental questions about infectivity of bacteria and viruses, and to make predictions for intervention policies, such as vaccines, de-colonization and case containment. For some types of infectious disease, the complexity of the transmission process and the corresponding model, combined with the characteristics of the available data, make 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; Stadler 2011; Fearnhead and Prangle 2012; Del Moral *et al.* 2012; Baragatti *et al.* 2013, who considered the transmission dynamics of *Mycobacterium tuberculosis* based on IS6110 fingerprinting data from tuberculosis (*M. tuberculosis*) cases in San Francisco, reported earlier by Small *et al.* (1994). Except for Stadler (2011), who proposed a likelihood-based inference scheme, 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 the estimation of epidemiological parameters of *M. tuberculosis* with the model of Tanaka *et al.* (2006) has been widely studied, important open issues still remain. Only more recently, for instance, Aandahl *et al.* (2014) reconciled a major difference in the estimates of the reproductive value R between the ABC approach of Tanaka *et al.* (2006) and the method of Stadler (2011). In order to accomplish this, Aandahl *et al.* (2014) simplified the inference task by either providing an informative prior for the death rate δ or, alternatively, by setting it to a fixed value. The need of using such prior knowledge in the inference indicates that the model may not be fully identifiable.

We here address the reliability of the ABC estimates by evaluating the accuracy of the approximate inference method and the influence of the prior, for genotype data of the kind available from the San Francisco study (Small *et al.* 1994). As the above-mentioned studies have assumed a fixed infectious population size of $10,000$ for the data, we further investigate how this choice influences the estimation of the epidemiological parameters, and whether it is

possible to infer the population size from this kind of genotype data without access to more extensive surveillance data about incidence. Since comparable data are widely considered for many different kinds of bacteria, the issue of model identifiability is of wider general interest beyond the particular case discussed here.

MODELS 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 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 with 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 has been earlier noted by Stadler (2011) that due to the time scaling of the model, at least one of the rate parameters must be fixed. As in the earlier studies, we use a time scale of one year and fix the mutation rate to $\tau = 0.198$ 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. In addition, we will consider 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 to 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 form of the allele vector $a = (a_1, a_2, \ldots, a_n) \in \mathbb{N}^n$, where element a_i is equal to 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 which belong to the cluster. For example, the vector $a = (4, 0, 1)$ 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 a is defined as the sample size n, which can be written in terms of 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 consists of an allele vector 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. But for the model considered in this paper, 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 which is related to ABC but which makes explicit its inherent approximations (Blum 2010; Gutmann and Corander 2015).

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

$$\hat{L}^N_{d,\epsilon}(\theta) \propto \frac{1}{N} \sum_{i=1}^{N} \chi_\epsilon \left(d(a^{(i)}, a^*) \right), \tag{1}$$

where $\chi_\epsilon(d)$ is an indicator function which equals one if the distance d is less than a threshold ϵ and zero otherwise, $a^{(i)}$ is an allele vector simulated using parameter θ, and N is the number of such simulations performed (Gutmann and Corander 2015). The distance $d(a, a^*)$ is a non-negative function which measures the similarity between the simulated allele vector

a and the observed allele vector a^*. Possible choices of d are discussed below. Equation 1 means that the likelihood is approximated by the fraction of times the simulated allele vector a is within distance ϵ from the observed allele vector 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 $d(a, a^*)$ is less than the threshold ϵ. Under certain conditions, the retained parameters correspond to samples from the posterior. In ABC, the approximate likelihood is generally never explicitly constructed. While this can be advantageous in some cases, it can also mask the influence of the prior on the posterior. In our approach, on the other hand, using Equation 1, we are able to disentangle the contribution of the approximate likelihood and the prior to the posterior. This is important because the approximate likelihood provides information about the identifiability of the parameters.

Since 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 121×137 evenly spaced grid over a subspace $\Delta_\alpha \times \Delta_\delta = [0.3, 2] \times [2, 1.5]$ of the (α, δ) BDM parameter space. At each node of the grid, $N = 3000$ allele vectors were simulated to approximate the likelihood. For inference of the population size m, we used a grid with $N = 30,000$ simulated allele vectors in each node. The amount of simulated data was selected to ensure the stability of the likelihood approximations. Results for the stabilization of the marginal likelihoods are shown in Appendix A. All of the grids were found to cover the relevant parameter values for the data used in this paper.

Distance measures used to approximate the likelihood: The likelihood approximation in Equation 1 relies on a distance measure $d(a, a^*)$ between the observed sample a^* and the sample a produced by the simulation process with the parameter vector θ. 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 $\epsilon \to 0$ and $N \to \infty$, one can easily define distance measures $d(a, a^*)$ which lead to exact likelihoods $L(\theta)$. The only requirement is that $d(a, a^*) = 0$ if and only if $a = a^*$. 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 $\hat{L}_{d,\epsilon}(\theta)$ dictated by the distance measure d, a non-zero 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 which are as close as possible to $L(\theta)$. Because the likelihood $L(\theta)$ is unknown in the first place, the 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 a^s where the data generating parameters θ^s are known. In particular, the mode of the likelihood approximation should, on average, be near θ^s if the sample a^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

$$\mathrm{d}_{\mathrm{base}}(a, a') = \frac{1}{n}|g_a - g_{a'}| + |H_a - H_{a'}|, \tag{2}$$

where $g_a = \sum_i a_i$ is the number of distinct alleles in a and $H_a = 1 - \sum_i a_i(i/n)^2$ is a gene diversity measure.

The second distance measure called "simple" is defined as

$$\mathrm{d}_{\mathrm{sim}}(a, a') = |a_1 - a'_1| + |M_a - M_{a'}|, \tag{3}$$

where a and a' are allele vectors and $M_a = \max\{i|a_i \neq 0\}$ is the largest cluster size in 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 considered is a generalized Kullback-Leibler divergence

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

where $f_a : \mathbb{R} \to \mathbb{R}$ is a smoothed continuous function which approximates a in the discrete locations i, that is, $f_a(i) \approx a_i$. The idea is to consider vectors a and a' as unnormalized probability vectors of cluster sizes and to compare their distribution while allowing small differences in the number of clusters of similar size. Unlike in the usual Kullback-Leibler divergence, f_a and $f_{a'}$ can be unnormalized, that is, they do not need to integrate to one. The generalized Kullback-Leibler divergence above belongs to the family of Bregman divergences (Bregman 1967) which have been shown to have a number of desirable properties and useful applications (see for example Collins *et al.* 2002; Frigyik *et al.* 2008; Gutmann and Hirayama 2011).

RESULTS

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

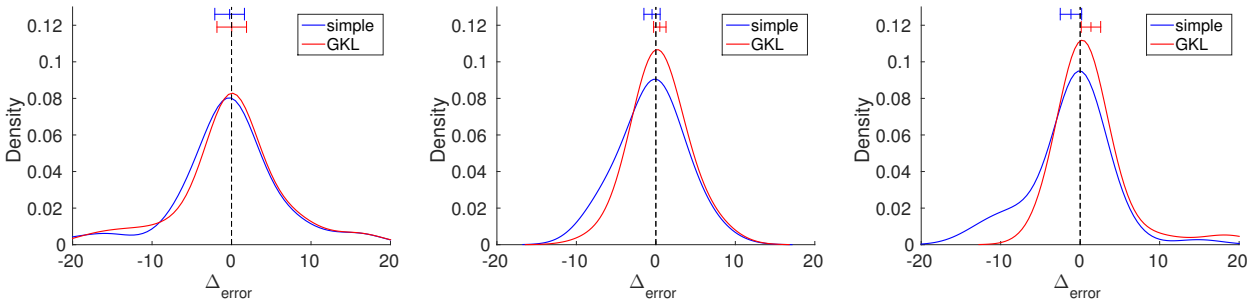
The relative error was defined as $\sum_i(|\theta_i^s - \hat{\theta}_i|/\theta_i^s)$, where θ^s is the true generating parameter used to simulate the synthetic data a^s, and $\hat{\theta}$ is the estimate obtained by maximizing the approximate likelihood. Maximization was performed in a simple way, by searching for the maximal value over the two-dimensional grid.

In total 50 simulated sets of observations a^s with a known generating parameter θ^s were used in three different setups. In the first setup, the value of θ^s was set to the estimate

$(0.69, 3.4)$ of TANAKA *et al.* (2006). As we were also interested to see if the values of the actual epidemiological parameters had an effect on the estimation accuracy, we considered two alternative values for θ^s. In the first alternative, we doubled the reproductive value R, and in the second, both the transmission rate t and the reproductive value R were divided by two.

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 zero 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 generalized Kullback-Leibler distance d_{GKL}, on the other hand, performs slightly better than the baseline and the difference is significant in the last setup (the zero mean hypothesis of Δ_{error} can be rejected at the p-value 0.0303). It should nevertheless be noted that the absolute errors tend to be rather large with all of the distance measures in this last setting as shown in Appendix C.

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



(a) $\theta^s = (3.4, 0.69)$ (b) $\theta^s = (6.8, 0.69)$ (c) $\theta^s = (1.7, 0.35)$

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 better performance of the alternative method. The intervals at the top show the estimated mean and the 95% confidence interval.

Effect of the prior: The simulator operates genuinely in the (α, δ) space, where α and δ are the birth rate and the death rate in the model. Accordingly, all of the studies so far have assumed a uniform prior for the region $0 < \delta < \alpha$ in the Bayesian framework. As the objective is to provide estimates for the epidemiological parameters t and R, it is important to understand the effect of this prior choice to the estimates.

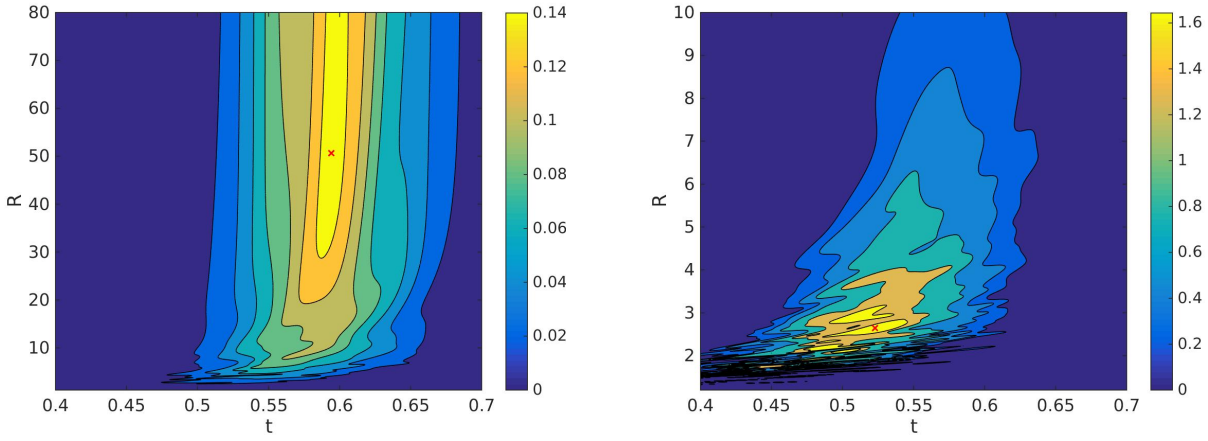
The law of transformation of random variables implies that choosing a uniform prior for (α, δ) is equivalent to choosing the following prior for (R, t),

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

The prior is visualized in Appendix B. The formula and the figure show that its probability mass is concentrated on small values of the reproductive value R.

We show next that the prior in Equation 5, equivalent to the the uniform prior for (α, δ), has a substantial impact on the posterior distribution of the epidemiological parameters t and R for the San Fransisco data of Small *et al.* (1994). Figure 2 visualizes the difference in shape between the likelihood and the posterior distribution in a rectangular support

$1.2 < R < 80$ and $0.4 < t < 0.7$. It can be seen that the uniform prior for (α, δ) leads to a substantial shift of the probability mass towards the lower end of values of R. The difference between the modes of the likelihood and the posterior is striking: $R = 50.6$ versus $R = 2.7$ 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 less than 80. The shape of the likelihood in Figure 2 (a) suggests that computations with larger values of R would lead to a corresponding increase of the credibility intervals.



(a) Likelihood

(b) Posterior with uniform prior in (α, δ) space

Figure 2: Likelihood and posterior distribution when using a uniform prior in the (α, δ) space. The red cross denotes the mode. Note the different scales of the y-axes.

Effect of the infectious population size: The previous 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) fixed the death rate to $\delta = 0.52$, a value obtained by summing the rates of self cure, death from causes other than tuberculosis and death from untreated tuberculosis as estimated in other studies. Moreover, as to our knowledge in all of the other

Prior	Parameter	Mode	Mean	Credible Interval (95%)
Uniform prior in (t, R)	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	$(0.46, 0.66)$

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.

relevant studies, an infectious population size m of $10,000$ individuals was assumed. 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 $\mathrm{d_{base}}$ in the likelihood approximation.

Figure 3 shows the likelihoods of R for $m = 1000$ and $m = 10,000$ using the San Francisco data (Small *et al.* 1994) and $\delta = 0.52$. The difference in location of the likelihoods is clear with modes at 1.1 for $m = 1000$ and 1.9 for $m = 10,000$ (posterior means assuming uniform prior were the same). Therefore the assumed infectious population size m affects the inference of R.

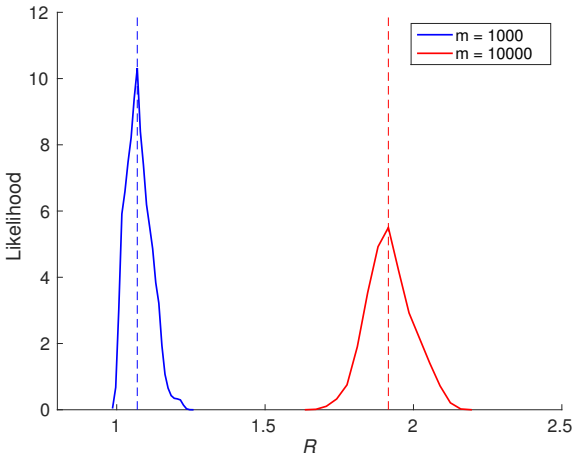


Figure 3: Likelihoods of the reproductive value R with fixed death rate $\delta = 0.52$ and two alternative population sizes m using the San Francisco data. The vertical lines indicate the modes of the likelihoods.

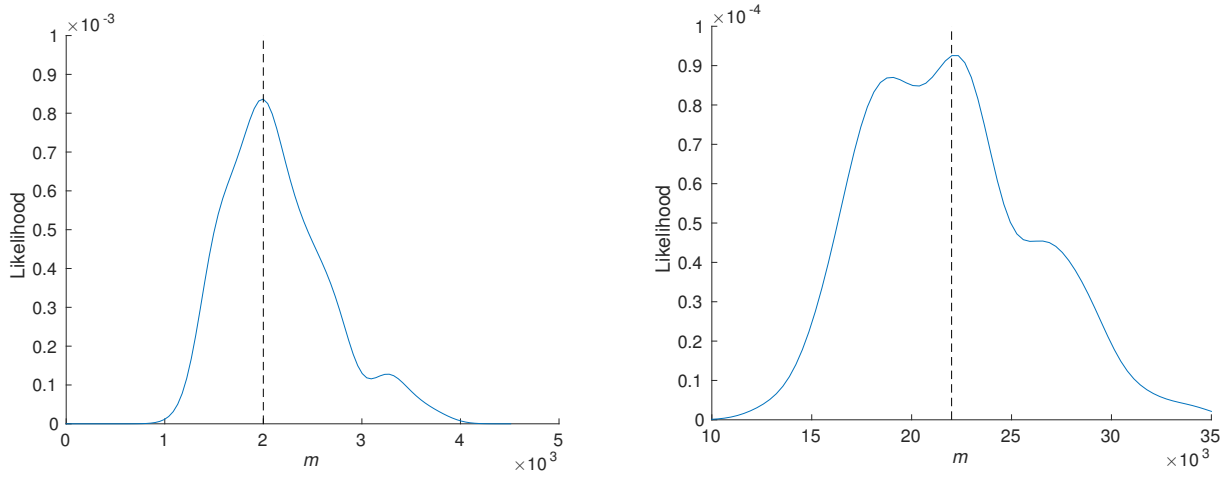
Inference of the infectious population size: The observed effect of the infectious population size m on the inference of R means that there is some (statistical) dependency between m and R. This suggests that it might be possible to infer the size of the underlying infectious population from the data when R and δ are known. Alternatively, due to 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 run 50 trials with synthetic data as before, and determined each time the maximizer of the approximate likelihood on the grid. 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, the true m is just barely outside of the interval. These results thus illustrate that estimation of m is possible, provided that reliable information is available about the other epidemiological parameters.

	m	$\bar{m}$	CI (95%)
$R = 2.1$	1000	1051	$(990, 1112)$
	10000	10020	$(9380, 10660)$
$R = 1.1$	1000	1007	$(976, 1038)$
	10000	10510	$(10003, 11017)$

Table 2: Mean estimated population size $\bar{m}$ for 50 trials and the respective confidence interval (CI) under two alternative configurations of R and population size m. Death rate was fixed to $\delta = 0.52$. Results are for synthetic data.

We next estimated the infectious population size for the *M. tuberculosis* in San Francisco area during the time the data of Small *et al.* (1994) were collected. Figure 4 shows the likelihood functions, given by two alternative configurations of R (with $\delta = 0.52$ as before). Assuming $R = 1.1$ produced the posterior mean $\hat{m} = 2150$ and the 95% credible interval $(1302, 3300)$ with uniform prior over m. Alternatively, assuming $R = 2.1$ in turn produced the posterior mean $\hat{m} = 21,900$ and the 95% credible interval $(14640, 29770)$.



(a) $R = 1.1$ (b) $R = 2.1$

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. The vertical lines indicate the modes of the likelihoods.

DISCUSSION

Statistical inference plays an important role in the study of the transmission dynamics of infectious disease. 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.

Due to the intractability of the transmission model, an approximate inference approach was used, belonging to the framework of "approximate Bayesian computation", which relies on a distance measure gauging similarity between observed and simulated data. In all of the tests we performed, the generalized Kullback-Leibler distance measure attained 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). While the measure used by Tanaka *et al.* (2006) has some clear biological meaning, the generalized Kullback-Leibler distance is a more general information theoretical construction. The observed increase in the 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).

Investigation of the likelihood suggests that the estimation of R is difficult when inferring both the reproductive value R and the transmission rate t. The credible intervals for R were $(1.4, 38.0)$ or $(9.5, 80.0)$ when using a uniform prior over the (α, δ) or (t, R) 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. The uniform prior over the (α, δ) space was, to our knowledge, the standard choice in all of the studies with a related setting, and we showed that it has a substantial impact on the

posterior, altering the shape of the likelihood significantly. In previous work, this has likely remained unobserved because unlike in our work, the approximate likelihood function was not explicitly constructed. Our study thus highlights the importance of using identifiability checks when performing inference for models with intractable likelihoods.

A uniform prior is usually considered uninformative so that 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 studies was that the infectious population size m equals $10,000$ individuals. We showed that the infectious population size influences the estimation of R; for $m = 10,000$, the posterior mean $\hat{R} = 1.9$ was well within the credible interval $(1.54, 2.66)$ reported by Aandahl *et al.* (2014). For $m = 1000$, we obtained instead an estimate $\hat{R} = 1.1$ which is clearly outside that credible interval.

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 $21,900$ 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 containing also large clusters. 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.

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

This research was funded by the grant no. 251170 from Academy of Finland to the COIN centre of excellence. We acknowledge the computational resources provided by the Aalto Science-IT project.

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.

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

Blum, M. G. B., 2010 Approximate Bayesian Computation: A Nonparametric Perspective. Journal of the American Statistical Association **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 Computational Mathematics and Mathematical Physics **7:** 200–217.

Collins, M., R. Schapire, and Y. Singer, 2002 Logistic Regression, AdaBoost and Bregman Distances. Machine Learning **48:** 253–285.

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

Fearnhead, P. and D. Prangle, 2012 Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. Journal of

the Royal Statistical Society: Series B (Statistical Methodology) **74:** 419–474.

Frigyik, B., S. Srivastava, and M. Gupta, 2008 Functional Bregman Divergence and Bayesian Estimation of Distributions. IEEE Trans. on Information Theory **54:** 5130–5139.

Gutmann, M. U. and J. Corander, 2015 Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. arXiv:1501.03291.

Gutmann, M. U., R. Dutta, S. Kaski, and J. Corander, 2014 Likelihood-Free Inference via Classification. arXiv:1407.4981.

Gutmann, M. U. and J. Hirayama, 2011 Bregman Divergence as General Framework to Estimate Unnormalized Statistical Models. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, Corvallis, Oregon, pp. 283–290. AUAI Press.

Sisson, S. A., Y. Fan, and M. M. Tanaka, 2007 Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences **104:** 1760–1765.

Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet, D. C. Ruston, G. F. Schecter, C. L. Daley, and G. K. Schoolnik, 1994 The Epidemiology of Tuberculosis in San Francisco – A Population-Based Study Using Conventional and Molecular Methods. New England Journal of Medicine **330:** 1703–1709.

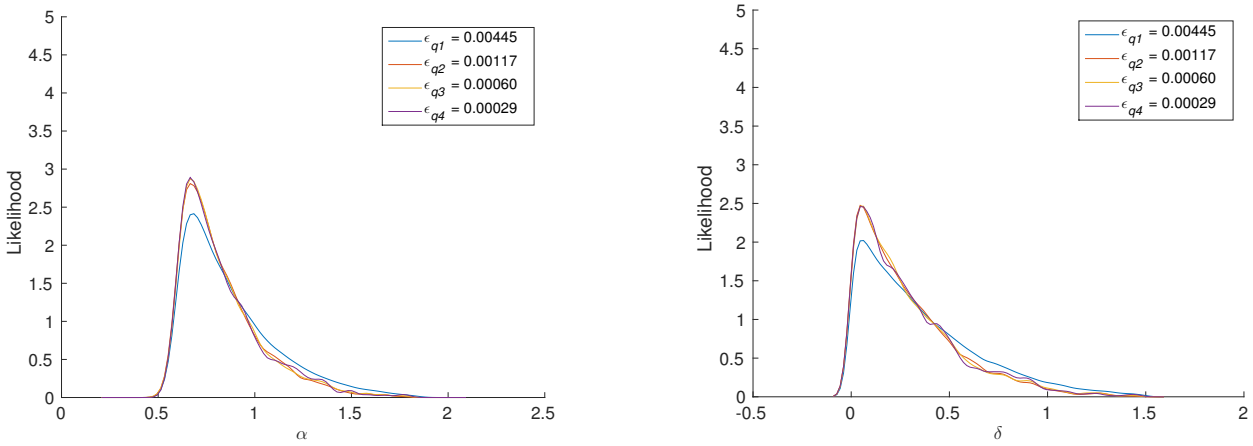
Stadler, T., 2011 Inferring Epidemiological Parameters on the Basis of Allele Frequencies. Genetics **188:** 663–672.

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.

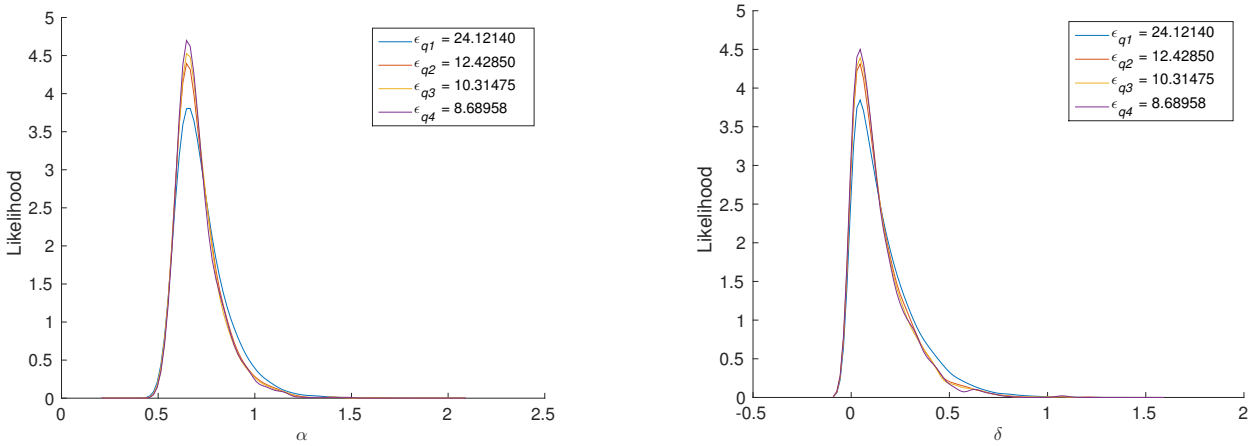
Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring Coalescence Times From DNA Sequence Data. Genetics **145:** 505–518.

APPENDIX A

Stabilization of the approximate likelihoods:



(a) Tanaka distance measure



(b) GKL distance measure

Figure 5: 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.

APPENDIX B

Transformed uniform prior:

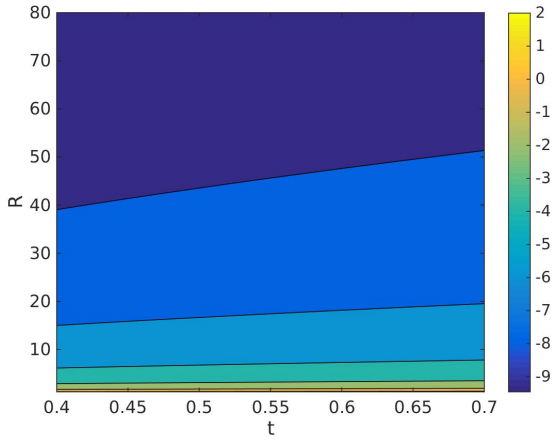
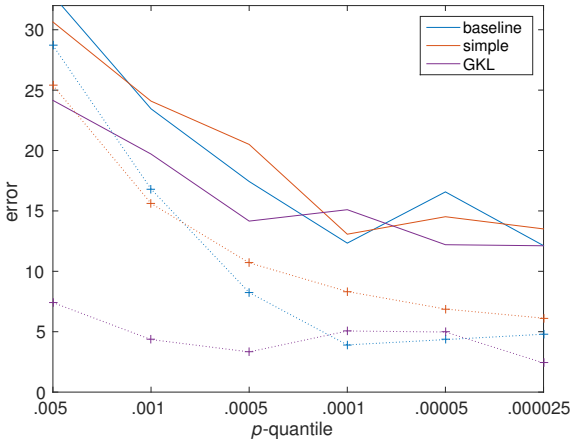


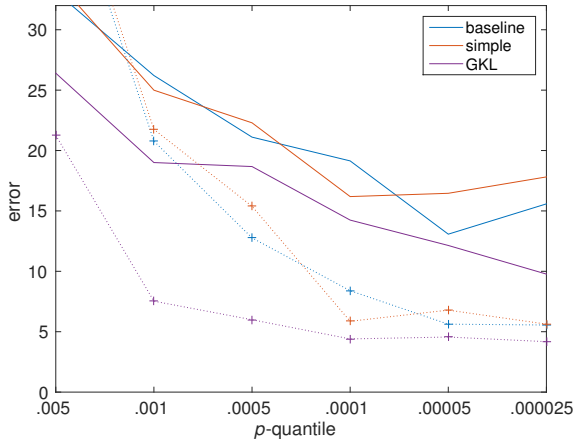
Figure 6: Visualization of the logarithm of the prior in Equation 5 corresponding to the uniform prior on (α, δ). Note the concentration of probability mass on small R.

APPENDIX C

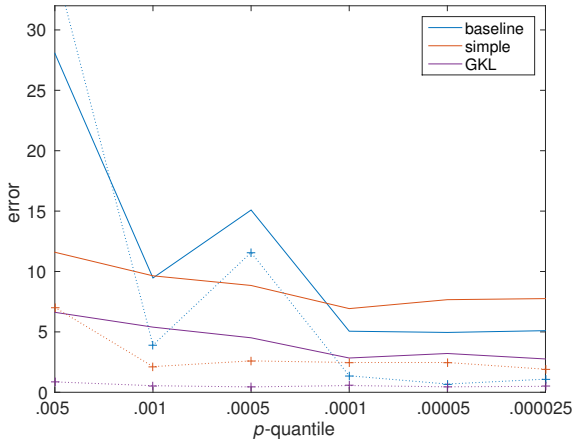
Mean and median absolute errors:



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



(b) $\theta^s = (6.8, 0.69)$

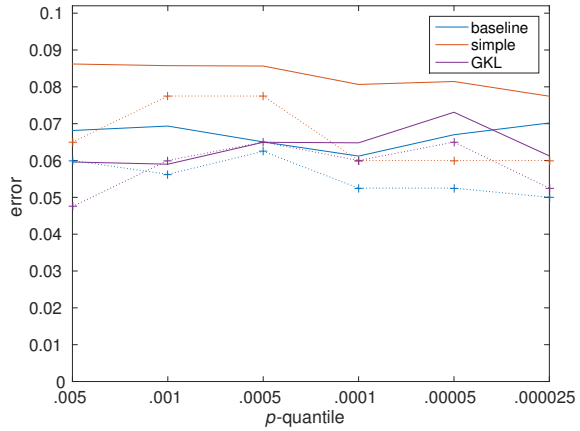


(c) $\theta^s = (1.7, 0.35)$

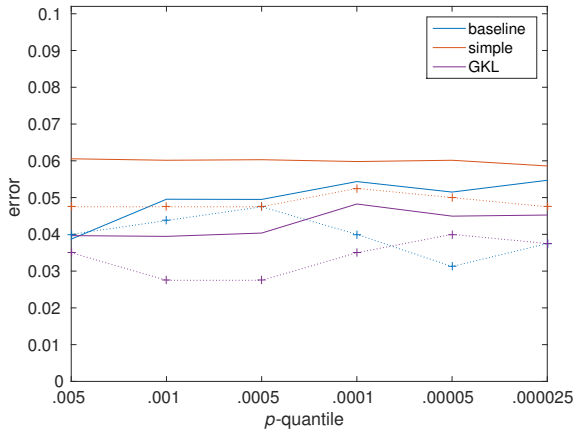
Figure 7: Mean (solid lines) and median (dotted lines) errors of the approximate maximum likelihood estimates with the three different distance measures $\mathrm{d}_{\mathrm{base}}$ (blue lines), $\mathrm{d}_{\mathrm{sim}}$ (red lines), $\mathrm{d}_{\mathrm{GKL}}$ (purple lines) and a decreasing threshold ϵ given by the p-quantile. The error is the L_1 distance between the vectors θ^s and $\hat{\theta}$, that is, $|\theta^s_1 - \hat{\theta}_1| + |\theta^s_2 - \hat{\theta}_2|$.

The rather large difference between the mean and median errors in Figure 7 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. The parameter values in Figure 7 (c) correspond to a larger death rate than in (a) and (b). This seems to significantly

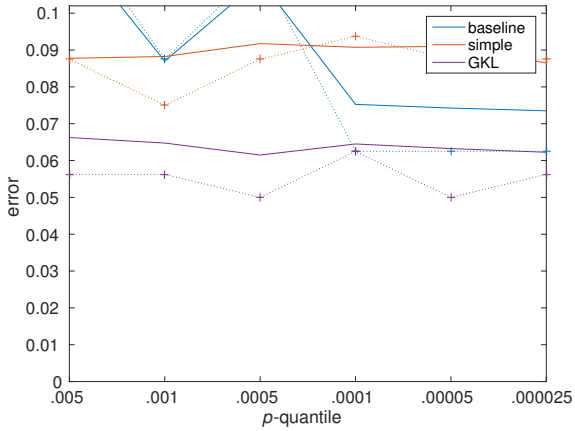
decrease the error, possibly because the generating parameter δ is further away from zero. Although the total errors in Figure 7 are large, the small errors in Figure 8 indicate that the estimation of the transmission rate t can still be done accurately and is not affected by the error-prone estimate of R.



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



(b) $\theta^s = (6.8, 0.69)$

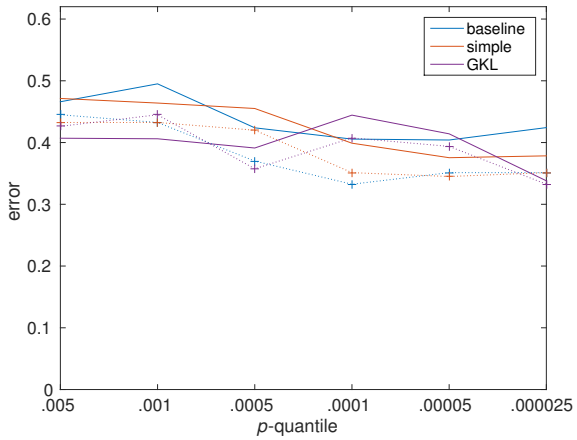


(c) $\theta^s = (1.7, 0.35)$

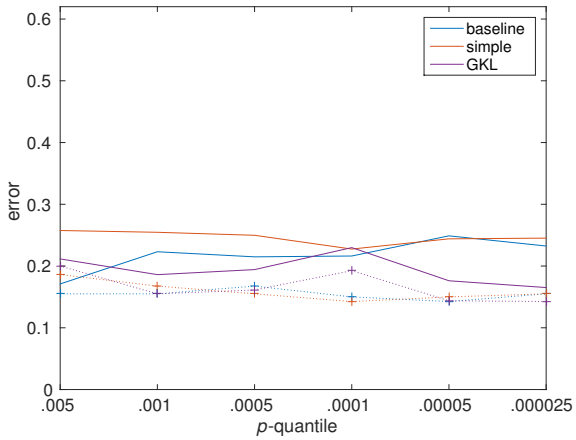
Figure 8: Mean and median errors of the approximate maximum likelihood estimate of t. Visualization is as in Figure 7. The error is the L_1 distance between t^s and $\hat{t}$, that is, $|t^s - \hat{t}|$.

APPENDIX D

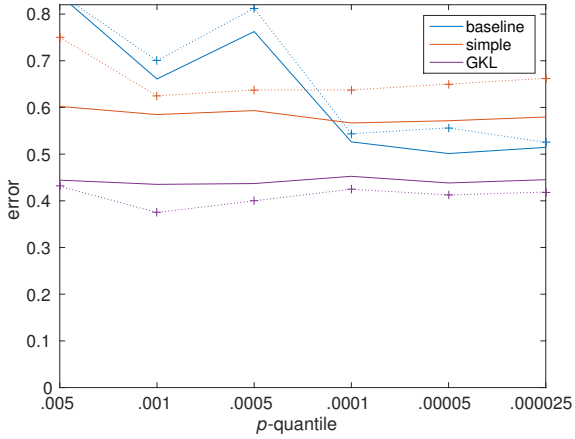
Absolute errors in the rate parameter space:



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



(b) $\phi^* = (0.81, 0.12)$



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

Figure 9: Mean and median errors in the estimated $\phi = (\alpha, \delta)$. Visualization and setup is as in Figure 7. The error is the L_1 distance between the vectors ϕ^* and $\hat{\phi}$.

We noticed a general tendency of acquiring small estimates of δ irrespective of the setup. This is a plausible reason for the slightly reduced error in Figure 9 (b) compared to the other setups as δ^* is the smallest in that setup.