Likelihood-free inference by ratio estimation

Ritabrata Dutta
InterDisciplinary Institute of Data Science, Universitá della Svizzera italiana
Jukka Corander
Department of Biostatistics, University of Oslo,
Samuel Kaski
Helsinki Institute for Information Technology,
Department of Computer Science, Aalto University
Michael U. Gutmann*
School of Informatics, University of Edinburgh

August 21, 2017

We consider the problem of parametric statistical inference when likelihood computations are prohibitively expensive but sampling from the model is possible. Several so-called likelihood-free methods have been developed to perform inference in the absence of a likelihood function. The popular synthetic likelihood approach infers the parameters by modelling summary statistics of the data by a Gaussian probability distribution. In another popular approach called approximate Bayesian computation, the inference is performed by identifying parameter values for which the summary statistics of the simulated data are close to those of the observed data. Synthetic likelihood is easier to use as no measure of "closeness" is required but the Gaussianity assumption is often limiting. Moreover, both approaches require judiciously chosen summary statistics. We here present an alternative inference approach that is as easy to use as synthetic likelihood but not as restricted in its assumptions, and that, in a natural way, enables automatic selection of relevant summary statistic from a large set of candidates. The basic idea is to frame the problem of estimating the posterior as a problem of estimating the ratio between the data generating distribution and the marginal distribution. This problem can be solved by logistic regression, and including regularising penalty terms enables automatic selection of the summary statistics relevant to the inference task. We illustrate the general theory on toy problems and use it to perform inference for stochastic nonlinear dynamical systems.

Keywords: approximate Bayesian computation, density-ratio estimation, likelihood-free inference, logistic regression, probabilistic classification, stochastic dynamical systems, summary statistics selection, synthetic likelihood

*Corresponding author: michael.gutmann@ed.ac.uk

1 Introduction

We consider the problem of inferring the posterior probability density function (pdf) of some model parameters $\theta \in \mathbb{R}^d$ given observed data $x_0 \in \mathcal{X}$ when computation of the likelihood function is too costly but data can be sampled from the model. In particular, we assume that the model specifies the data generating pdf $p(x|\theta)$ not explicitly, e.g. in closed form, but only implicitly in terms of a stochastic simulator that generates samples x from the model $p(x|\theta)$ for any value of the parameter θ. The simulator can be arbitrarily complex so that we do not impose any particular conditions on the data space $\mathcal{X}$. Such simulator-based (generative) models are used in a wide range of scientific disciplines to simulate different aspects of nature on the computer, ranging from sub-atomic particles (Martinez *et al.*, 2016) to human societies (Turchin *et al.*, 2013), or universes (Schaye *et al.*, 2015).

Denoting the prior pdf of the parameters by $p(\theta)$, the posterior pdf $p(\theta|x_0)$ can be obtained from Bayes' formula,

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}, \qquad p(x) = \int p(\theta)p(x|\theta)\,\mathrm{d}\theta, \tag{1}$$

for $x = x_0$. Exact computation of the posterior pdf is, however, impossible if the likelihood function $L(\theta) \propto p(x_0|\theta)$ is too costly to compute. Several approximate inference methods have appeared for simulator-based models. They are collectively known as likelihood-free inference methods, and include approximate Bayesian computation (Tavaré *et al.*, 1997; Pritchard *et al.*, 1999; Beaumont *et al.*, 2002) and the synthetic likelihood approach (Wood, 2010). For a comprehensive introduction to the field, we refer the reader to the review papers by Lintusaari *et al.* (2017); Marin *et al.* (2012); Hartig *et al.* (2011); Beaumont (2010).

Approximate Bayesian computation (ABC) relies on finding parameter values for which the simulator produces data that are similar to the observed data. Similarity is typically assessed by reducing the simulated and observed data to summary statistics and comparing their distance. While the summary statistics are classically determined by expert knowledge about the problem at hand, there have been recent pursuits in choosing them in an automated manner (Aeschbacher *et al.*, 2012; Fearnhead and Prangle, 2012; Blum *et al.*,

2013; Gutmann *et al.*, 2014, 2017). While ABC can be considered to implicitly construct a nonparametric approximation of $p(x|\theta)$ (e.g. Hartig *et al.*, 2011; Gutmann and Corander, 2016), synthetic likelihood assumes that the summary statistics for a given parameter value follow a Gaussian distribution (Wood, 2010). The synthetic likelihood approach is applicable to a diverse set of problems (Price *et al.*, 2016), but the Gaussianity assumption may not always hold and the original method does not include a mechanism for choosing summary statistics automatically.

In this paper, we propose a framework and practical method to directly approximate the posterior distribution in the absence of a tractable likelihood function. The proposed approach includes automatic selection of summary statistics in a natural way.

The basic idea is to frame the original problem of estimating the posterior as a problem of estimating the ratio $r(x, \theta)$ between the data generating pdf $p(x|\theta)$ and the marginal distribution $p(x)$,

$$r(x, \theta) = \frac{p(x|\theta)}{p(x)}. \tag{2}$$

By definition of the posterior distribution, an estimate $\hat{r}(x, \theta)$ for the ratio implies an estimate $\hat{p}(\theta|x_0)$ for the posterior,

$$\hat{p}(\theta|x_0) = p(\theta)\hat{r}(x_0, \theta). \tag{3}$$

In addition, the estimated ratio also yields an estimate $\hat{L}(\theta)$ of the likelihood function,

$$\hat{L}(\theta) \propto \hat{r}(x_0, \theta), \tag{4}$$

as the denominator $p(x)$ in the ratio does not depend on θ. We can thus perform likelihood-free inference by ratio estimation, and we call this framework in short "LFIRE". If approximating the likelihood function is the goal, also other distributions than the marginal can be chosen in the denominator. While we do not further address the question of what distributions can be chosen, for reasons of stability, however, at first glance it seems reasonable to prefer distributions that have heavier tails than $p(x|\theta)$ in the numerator.

There are several methods in the literature available for the estimation of density ratios (Sugiyama *et al.*, 2012), of which estimation through logistic regression is widely used and has some favourable asymptotic properties (Qin, 1998; Cheng and Chu, 2004; Bickel *et al.*,

2007). Logistic regression is very closely related to probabilistic classification and we use it in the paper to estimate the ratio $r(x, \theta)$.

Logistic regression and probabilistic classification have been employed before to address computational problems in statistics. Gutmann and Hyvärinen (2012) used it to estimate unnormalised models, Pham *et al.* (2014); Cranmer *et al.* (2015) used it to estimate likelihood ratios, and Goodfellow *et al.* (2014) employed it for training neural networks to generate samples following the same distribution as given reference data. More general methods for ratio estimation are now considered for training such neural networks (see e.g. the review by Mohamed and Lakshminarayanan, 2016), and they were used before to estimate unnormalised models (Pihlaja *et al.*, 2010; Gutmann and Hirayama, 2011). Classification in general has been shown to yield a natural distance function in terms of the classifiability between simulated and observed data, which can be used for ABC (Gutmann *et al.*, 2014, 2017). While this earlier approach is very general, the classification problem can be difficult to set up when the observed data consist of very few data points only. The method proposed in this paper avoids this difficulty.

The details on how to generally estimate the ratio $r(x, \theta)$ and hence the posterior by logistic regression are presented in Section 2. In Section 3, we model the ratio as a linear superposition of summary statistics and show that this assumption yields an exponential family approximation of the intractable model pdf. As Gaussian distributions are part of the exponential family, our approach thus includes the synthetic likelihood approach as a special case. We then show in Section 4 that including a penalty term in the logistic regression enables automatic selection of relevant summary statistics. In Section 5, we validate the resulting method on toy examples, and in Section 6, we apply it to challenging inference problems in ecology and weather forecasting. All simulation studies include a comparison with the synthetic likelihood approach. The new method yielded consistently more accurate inference results.

2 Posterior estimation by logistic regression

We here show that the ratio $r(x, \theta)$ in Equation (2) can be estimated by logistic regression, which yields estimates for the posterior and the likelihood function together with Equations

(3) and (4). Figure 1 provides an overview.

By assumption we can generate data from the pdf $p(x|\theta)$ in the numerator of the ratio $r(x, \theta)$; let $X^\theta = \{x_i^\theta\}_{i=1}^{n_\theta}$ be such a set with n_θ independent samples generated with a fixed value of θ. Additionally we can also generate data from the marginal pdf $p(x)$ in the denominator of the ratio; let $X^m = \{x_i^m\}_{i=1}^{n_m}$ be such a set with n_m independent samples. As the marginal $p(x)$ is obtained by integrating out θ, see Equation (1), the samples can be obtained by first sampling from the joint distribution of (x, θ) and then ignoring the sampled parameters,

$$\theta_i \sim p(\theta), \qquad\qquad x_i^m \sim p(x|\theta_i). \tag{5}$$

We now formulate a classification problem where we aim to determine whether some data x were sampled from $p(x|\theta)$ or from $p(x)$. This classification problem can be solved via (nonlinear) logistic regression (e.g. Hastie *et al.*, 2001), where the probability for x to belong to X^θ, for instance, is parametrised by some nonlinear function $h(x)$,

$$\mathbb{P}(x \in X^\theta; h) = \frac{1}{1 + \nu \exp(-h(x))}, \tag{6}$$

with $\nu = n_m/n_\theta$ compensating for unequal class sizes. A larger value of h at x indicates a larger probability for x to originate from X^θ. A suitable function h is typically found by minimising the loss function $\mathcal{J}$ on the training data X^θ and X^m,

$$\mathcal{J}(h, \theta) = \frac{1}{n_\theta + n_m} \left\{ \sum_{i=1}^{n_\theta} \log \left[1 + \nu \exp(-h(x_i^\theta))\right] + \sum_{i=1}^{n_m} \log \left[1 + \frac{1}{\nu} \exp(h(x_i^m))\right] \right\}. \tag{7}$$

The dependency of the loss function on θ is due to the dependency of the training data X^θ on θ.

We prove in Appendix A that for large n_m and n_θ, the minimising function h^* is given by the log-ratio between $p(x|\theta)$ and $p(x)$, that is

$$h^*(x, \theta) = \log r(x, \theta). \tag{8}$$

For finite sample sizes n_m and n_θ, the minimising function $\hat{h}$,

$$\hat{h} = \arg\min_h \mathcal{J}(h, \theta), \tag{9}$$

thus provides an estimate $\hat{r}(x, \theta)$ of the ratio $r(x, \theta)$,

$$\hat{r}(x, \theta) = \exp(\hat{h}(x, \theta)), \tag{10}$$

and Equations (3) and (4) yield the corresponding estimates for the posterior and likelihood function, respectively,

$$\hat{p}(\theta|x_0) = p(\theta) \exp(\hat{h}(x_0, \theta)), \qquad \hat{L}(\theta) \propto \exp(\hat{h}(x_0, \theta)). \tag{11}$$

In case samples from the posterior are needed, we can use standard sampling schemes with $\hat{p}(\theta|x_0)$ as the target pdf (Andrieu and Roberts, 2009). The estimates can also be used together with Bayesian optimisation for fast likelihood-free inference (Gutmann and Corander, 2016).

When estimating the posterior or likelihood function by logistic regression as outlined above, the sample sizes n_m and n_θ are entirely under our control. Their values reflect the trade-off between computational and statistical efficiency. We note that both X^θ and X^m can be constructed in a perfectly parallel manner. Moreover, while X^θ needs to be re-constructed for each value of θ, X^m is independent of θ and needs to be generated only once.

Different models can be used for probabilistic classification; equivalently, different assumptions can be made on the family of functions to which the log-ratio h belongs. While non-parametric families or deep architectures can be used, we next consider a simple parametric family that is spanned by a set of summary statistics.

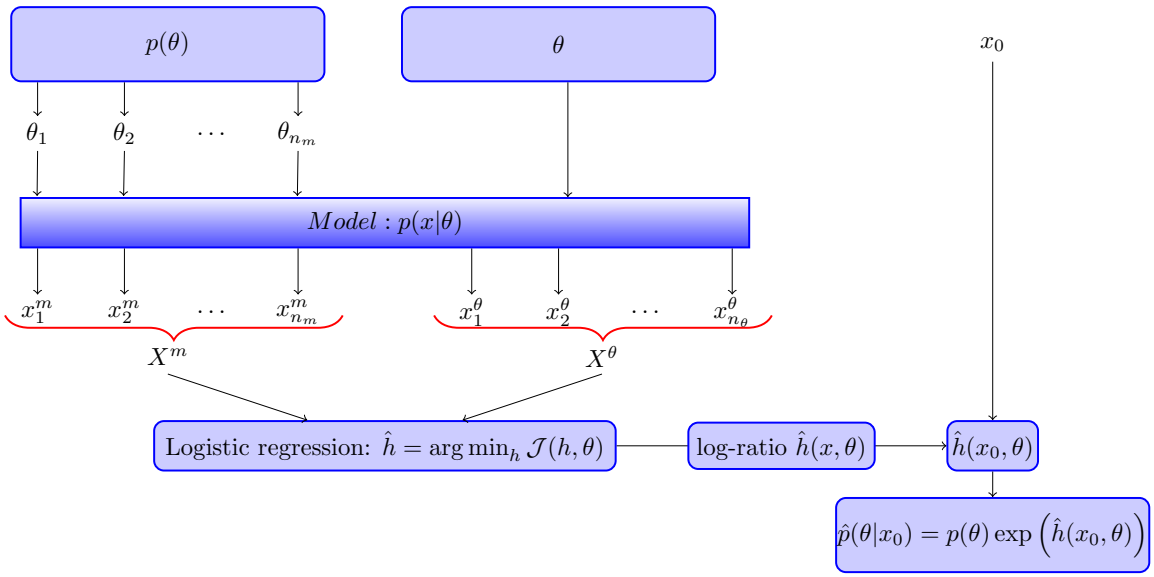


Figure 1: A schematic view of likelihood-free inference by ratio estimation (LFIRE). Logistic regression is used to estimate the posterior distribution $\hat{p}(\theta|x_0)$ as explained in Equations (5) to (11).

3 Exponential family approximation

We here restrict the search in Equation (9) to functions h that are members of the family spanned by b summary statistics $\psi_i(x)$, each mapping data $x \in \mathcal{X}$ to $\mathbb{R}$,

$$h(x) = \sum_{i=1}^{b} \beta_i \psi_i(x) = \beta^\top \psi(x), \tag{12}$$

with $\beta_i \in \mathbb{R}$, $\beta = (\beta_1, \ldots, \beta_b)$, and $\psi(x) = (\psi_1(x), \ldots, \psi_b(x))$. This corresponds to performing logistic regression with a linear basis expansion (Hastie *et al.*, 2001). The observed data x_0 may be used in the definition of the summary statistics, as for example with the Ricker model in Section 6, and thus influence the logistic regression part of the likelihood-free inference pipeline in Figure 1 (not shown in the figure).

When we assume that $h(x)$ takes the functional form in Equation (12), estimation of the ratio $r(x, \theta)$ boils down to the estimation of the coefficients β_i. This is done by minimising $J(\beta, \theta) = \mathcal{J}(\beta^\top \psi, \theta)$ with respect to β,

$$\hat{\beta}(\theta) = \arg\min_{\beta \in \mathbb{R}^b} J(\beta, \theta), \tag{13}$$

$$J(\beta,\theta) = \frac{1}{n_\theta + n_m}\left\{\sum_{i=1}^{n_\theta}\log\left[1+\nu\exp(-\beta^\top\psi_i^\theta)\right] + \sum_{i=1}^{n_m}\log\left[1+\frac{1}{\nu}\exp(\beta^\top\psi_i^m)\right]\right\} \quad (14)$$

The terms $\psi_i^\theta = \psi(x_i^\theta)$ and $\psi_i^m = \psi(x_i^m)$ denote the summary statistics of the simulated data sets $x_i^\theta \in X^\theta$ and $x_i^m \in X^m$, respectively. The estimated coefficients $\hat{\beta}$ depend on θ because the training data $x_i^\theta \in X^\theta$ depend on θ. With the model assumption in Equation (12), the estimate for the ratio in Equation (10) thus becomes

$$\hat{r}(x,\theta) = \exp(\hat{\beta}(\theta)^\top\psi(x)) \quad (15)$$

and the estimates for the posterior and likelihood function in Equation (11) are

$$\hat{p}(\theta|x_0) = p(\theta)\exp(\hat{\beta}(\theta)^\top\psi(x_0)), \qquad \hat{L}(\theta) \propto \exp(\hat{\beta}(\theta)^\top\psi(x_0)), \quad (16)$$

respectively.

As $r(x,\theta)$ is the ratio between $p(x|\theta)$ and $p(x)$, we can consider the estimate $\hat{r}(x,\theta)$ in Equation (15) to provide an implicit estimate $\hat{p}(x|\theta)$ of the intractable model pdf $p(x|\theta)$,

$$p(x|\theta) \approx \hat{p}(x|\theta), \qquad \hat{p}(x|\theta) = \hat{p}(x)\exp(\hat{\beta}(\theta)^\top\psi(x)). \quad (17)$$

The estimate is implicit because we have not explicitly estimated the marginal pdf $p(x)$. Importantly, the equation shows that $\hat{p}(x|\theta)$ belongs to the exponential family with $\psi(x)$ being the sufficient statistics for the family, and $\hat{\beta}(\theta)$ the vector of natural parameters.

In previous work, Wood (2010) in his synthetic likelihood approach, as well as Leuenberger and Wegmann (2010), approximated the model pdf by a member from the Gaussian family. As the Gaussian family belongs to the exponential family, the approximation in Equation (17) includes this previous work as a special case. Specifically, a synthetic likelihood approximation with summary statistics ϕ corresponds to an exponential family approximation where the summary statistics ψ are the individual ϕ_k, all pairwise combinations $\phi_k\phi_{k'}$, $k \geq k'$, and a constant. While in the synthetic likelihood approach, the weights of the summary statistics are determined by the mean and covariance matrix of ϕ, in our approach, they are determined by the solution of the optimisation problem in (14). Hence, even if equivalent summary statistics are used, the two approaches can yield different approximations if the summary statistics are actually not Gaussian. Computationally, both methods require generating the data set X^θ, which most often will dominate

the computational cost. The proposed method has the additional cost of constructing the set X^m once and the cost of performing logistic regression for each θ. Synthetic likelihood, on the other hand, requires inversion of the covariance matrix of the summary statistics for each θ. We thus consider the computational cost of both methods to be roughly equal. Later in the paper, we compare the posteriors estimated by the two approaches. We will see that for equivalent summary statistics, relaxing the Gaussianity assumption typically leads to better inference results.

4 Data-driven selection of summary statistics

The estimated coefficients $\hat{\beta}(\theta)$ are weights that determine to which extent a summary statistic $\psi_i(x)$ contributes to the approximation of the posterior. As the number of simulated data sets n_m and n_θ increases, the error in the estimates $\hat{\beta}(\theta)$ decreases and the importance of each summary statistic can be determined more accurately. Increasing the number of simulated data sets, however, increases the computational cost too. As an alternative to increasing the number of simulated data sets, we here use an additional penalty term in the logistic regression to determine the importance of each summary statistic. This approach enables us to work with a large list of candidate summary statistics and automatically select the relevant ones in a data-driven manner. This makes the posterior inference more robust and less dependent on subjective user input.

While many choices are possible, we use the L_1 norm of the coefficients as penalty term, like in lasso regression (Tibshirani, 1994). The coefficients β in the basis expansion in Equation (12) are thus determined as the solution of a L_1-regularised logistic regression problem,

$$\hat{\beta}_{\text{reg}}(\theta, \lambda) = \underset{\beta \in \mathbb{R}^b}{\arg\min}\, J(\beta, \theta) + \lambda \sum_{i=1}^{b} |\beta_i|. \tag{18}$$

The value of λ determines the degree of the regularisation. Sufficiently large values cause some of the coefficients to be exactly zero. Different schemes to choose λ have been proposed that aim at minimising the prediction risk (Zou *et al.*, 2007; Wang and Leng, 2007; Tibshirani and Taylor, 2012; Dutta *et al.*, 2012). Following common practice and recommendations (Tibshirani, 1994; Hastie *et al.*, 2001; Greenshtein and Ritov, 2004; Efron *et al.*,

2004; Zou *et al.*, 2007; van de Geer, 2008; Friedman *et al.*, 2010; Tibshirani, 2011; van de Geer and Lederer, 2013)), we here choose λ by minimising the prediction risk $\mathcal{R}(\lambda)$,

$$\mathcal{R}(\lambda) = \frac{1}{n_\theta + n_m} \left\{ \sum_{i=1}^{n_\theta} \mathbb{1}_{\Pi_\lambda(x_i^\theta)<0.5} + \sum_{i=1}^{n_m} \mathbb{1}_{\Pi_\lambda(x_i^m)>0.5} \right\}, \tag{19}$$

estimated by ten-fold cross-validation, where $\Pi_\lambda(x) = \mathbb{P}(x \in X^\theta; h(x) = \hat{\beta}_{\text{reg}}(\theta, \lambda)^\top \psi(x))$. The minimising value $\lambda_{\min}$ determines the coefficient $\hat{\beta}(\theta)$,

$$\hat{\beta}(\theta) = \hat{\beta}_{\text{reg}}(\theta, \lambda_{\min}), \tag{20}$$

which is used in the estimate of the density ratio in Equation (15), and thus the posterior and likelihood in Equation (17). Algorithm 1 presents pseudo-code that summarises the procedure for joint summary statistics selection and posterior estimation. Algorithm 1 is a special case of the scheme described in Figure 1 when $h(x)$ is a linear combination of the summary statistics $\psi(x)$ as described in Equation (12).

Algorithm 1 Posterior estimation by penalised logistic regression

1: Consider b-dimensional summary statistics $\psi : x \in \mathcal{R} \mapsto \mathbb{R}^b$.

2: Simulate n_m samples $\{x_i^m\}_{i=1}^{n_m}$ from the marginal density $p(x)$.

3: To estimate the posterior pdf at parameter value θ do:

   a. Simulate n_θ samples $\{x_i^\theta\}_{i=1}^{n_\theta}$ from the model pdf $p(x|\theta)$

   b. Estimate $\hat{\beta}_{\text{reg}}(\theta, \lambda)$ by solving the optimisation problem in Equation (18) for $\lambda \in [10^{-4}\lambda_0, \lambda_0]$ where λ_0 is the smallest λ value for which $\hat{\beta}_{\text{reg}} = 0$.

   c. Find the minimiser $\lambda_{\min}$ of the prediction risk $\mathcal{R}(\lambda)$ in Equation (19) as estimated by ten-fold cross-validation, and set $\hat{\beta}(\theta) = \hat{\beta}_{\text{reg}}(\theta, \lambda_{\min})$.

   d. Compute the value of the estimated posterior pdf $\hat{p}(\theta|x_0)$ according to Equation (16).

For the results in this paper, we always used $n_\theta = n_m$. To implement steps b and c we used the R package 'glmnet' (Friedman *et al.*, 2010).

5 Validation on toy inference problems

We here validate and illustrate the presented theory on a set of toy inference problems.

5.1 Gaussian distribution

We illustrate the proposed inference method on the simple example of estimating the posterior pdf of the mean of a Gaussian distribution. The observed data x_0 is a single observation that was sampled from a uni-variate Gaussian with mean $\mu_o = 2.3$ and standard deviation $\sigma_o = 3$. Assuming a uniform prior $\mathcal{U}(-20, 20)$ on the unknown mean μ, the log posterior density of μ given x_0 is

$$\log p(\mu|x_0) = \alpha_0(\mu) + \alpha_1(\mu)x_0 + \alpha_2(\mu)x_0^2 \tag{21}$$

if $\mu \in (-20, 20)$, and zero otherwise. The model is thus within the family of models specified in Equation (16). Coefficient $\alpha_0(\mu)$ equals

$$\alpha_0(\mu) = -\frac{\mu^2}{2\sigma_0^2} - \log\left(\sqrt{2\pi\sigma_0^2}\right) - \log\left(\Phi\left(\frac{20 - x_0}{\sigma_0}\right) - \Phi\left(\frac{-20 - x_0}{\sigma_0}\right)\right), \tag{22}$$

where Φ is the cumulative distribution function of the standard normal distribution, and the coefficients $\alpha_1(\mu)$ and $\alpha_2(\mu)$ are

$$\alpha_1(\mu) = \frac{\mu}{\sigma_0^2}, \qquad \alpha_2(\mu) = -\frac{1}{2\sigma_0^2}. \tag{23}$$

For Algorithm 1, we used ten-dimensional summary statistics $\psi(x) = (1, x^2, \ldots, x^{b-1})$, so that $b = 10$, and fixed $n_m = n_\theta = 1000$. As an illustration of step c of Algorithm 1, we show the prediction error $\mathcal{R}(\lambda)$ in Figure 2a as a function of λ for a fixed value of μ. The chosen λ_{min} minimises the prediction error. Repeating step 3 in Algorithm 1 for different values of μ on a grid over the interval $[-5, 5]$, we estimated the ten-dimensional coefficient vector $\hat{\beta}(\mu)$ as a function of μ, which corresponds to an estimate $\hat{\alpha}(\mu)$ of $\alpha(\mu)$, and hence of the posterior, by Equation 16.

In Figure 2b, we plot $\hat{\alpha}(\mu)$ and $\alpha_0(\mu)$, $\alpha_1(\mu)$, $\alpha_2(\mu)$ from Equation (21) for $\mu \in [-5, 5]$. We notice that the estimated coefficients α_k are exactly zero for $k > 2$ while for $k \leq 2$, they match the true coefficients up to random fluctuations. This shows that our inference procedure can select the summary statistics that are relevant for the estimation of the posterior distribution from a larger set of candidates.

In Figure 2c, we compare the estimated posterior pdf (yellow) with the true posterior pdf (blue). We can see that the estimate matches the true posterior up to random fluctuations.

The figure further depicts the posterior obtained by the synthetic likelihood approach of Wood (2010) (red) where the summary statistics $\phi(x)$ are equal to x. Here, working with Gaussian data, the performance of the proposed inference scheme based on penalised logistic regression and the performance of the existing synthetic likelihood approach are practically equivalent.

5.2 Autoregressive model with conditional heteroskedasticity

The observed data are a time-series $x_0 = \left(y^{(t)},\ t = 1, \ldots, T\right)$ produced by a lag-one autoregressive model with conditional heteroskedasticity (ARCH(1)),

$$y^{(t)} = \theta_1 y^{(t-1)} + e^{(t)}, \qquad e^{(t)} = \xi^{(t)} \sqrt{0.2 + \theta_2 (e^{(t-1)})^2}, \qquad t = 1, \ldots, T, \qquad y^{(0)} = 0, \tag{24}$$

where $T = 100$, and $\xi^{(t)}$ and $e^{(0)}$ are independent standard normal random variables. The parameters in the model, θ_1 and θ_2, are correspondingly the mean and variance process coefficients. The observed data were generated with $\theta^0 = (\theta_1^o, \theta_2^o) = (0.3, 0.7)$ and we assume uniform priors $\mathcal{U}(-1, 1)$ and $\mathcal{U}(0, 1)$ on the unknown parameters θ_1 and θ_2, respectively. The true posterior distribution of $\theta = (\theta_1, \theta_2)$ can be computed numerically (e.g. Gutmann *et al.*, 2017, Appendix 1.2.4). This enables us to compare the estimated posterior with the true posterior using the symmetrised Kullback-Leibler divergence (sKL), where sKL between two continuous distributions with densities p and q is defined as

$$sKL(p||q) = \frac{1}{2} \int p(x) \log \frac{p(x)}{q(x)} dx + \frac{1}{2} \int q(x) \log \frac{q(x)}{p(x)} dx. \tag{25}$$

For estimating the posterior distribution with Algorithm 1, we used summary statistics ψ that measure the (nonlinear) temporal correlation between the time-points, namely the auto-correlations with lag one up to five, all pairwise combinations of them, and additionally a constant. For checking the robustness of the approach, we also considered the case where almost 50% of the summary statistics are noise by augmenting the above set of summary statistics by 15 white-noise random variables. For synthetic likelihood, we used the auto-correlations as the summary statistics without any additional noise variables, as synthetic likelihood approach is typically not adapted to selecting among relevant and irrelevant summary statistics. As explained in Section 4, synthetic likelihood always uses the pairwise combinations of the summary statistics due to its underlying Gaussianity assumption.

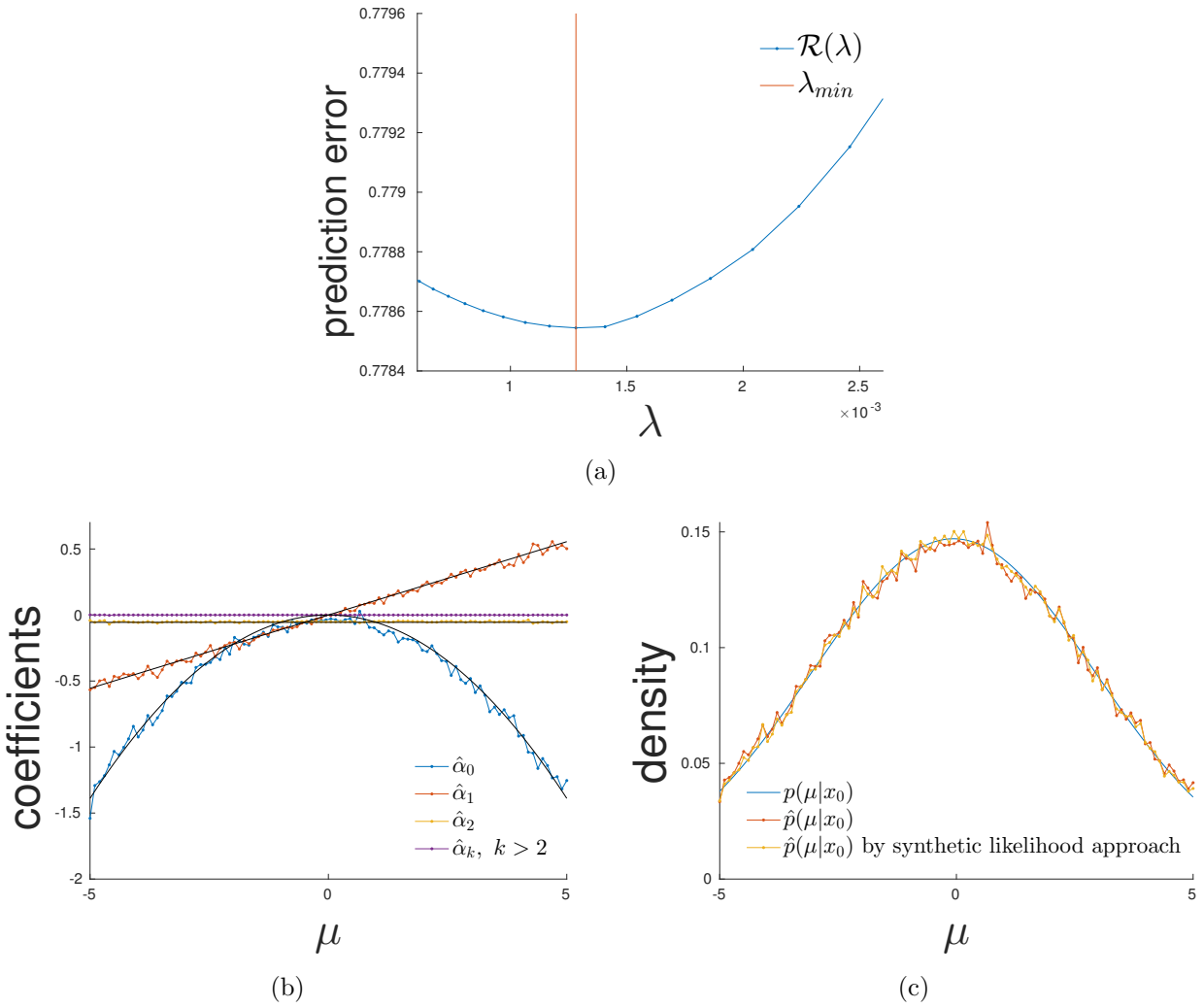


Figure 2: Steps for estimating the posterior distribution of the mean of a Gaussian. (a) For any fixed value of μ, λ_{min} minimises the estimated prediction error $\mathcal{R}(\lambda)$ (vertical line). (b) The figure shows the true coefficients from Equation (21) in black and the coefficients estimated by Algorithm 1 in colour. The algorithm sets the coefficients of unnecessary summary statistics automatically to zero. (c) Comparison of the estimated posterior with the posterior by the synthetic likelihood approach and the true posterior.

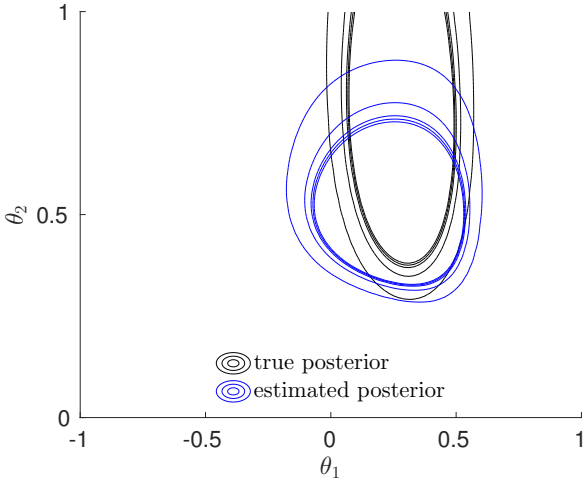
Method	$n_s = 100$	$n_s = 500$	$n_s = 1000$
Synthetic likelihood	1.82	1.80	2.25
Proposed method	2.04	1.57	1.48
Proposed method with 50% noise	3.24	1.60	1.51

Table 1: ARCH(1): Average symmetrised Kullback-Leibler divergence between the true and estimated posterior for $n_\theta = n_m = n_s \in \{100, 500, 1000\}$. Smaller values of the divergence mean better results.

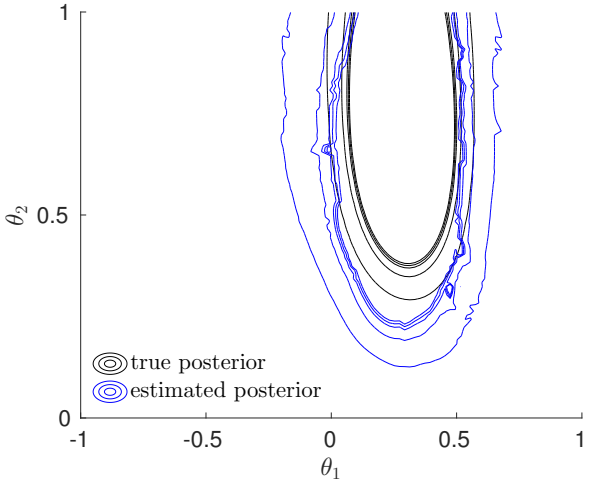
We estimated the posterior distribution on a 100 by 100 mesh-grid over the parameter space $[-1, 1] \times [0, 1]$ both for the proposed and the synthetic likelihood method. A comparison between two estimates is shown in Figure 3. The figure shows that the proposed approach yields a better approximation than the synthetic likelihood approach. Moreover, the posterior estimated with our method remains stable in the presence of the irrelevant summary statistics.

In order to assess the performance more systematically, we next performed posterior inference for 100 observed time-series that were each generated from Equation (24) with $\theta^0 = (\theta_1^o, \theta_2^o) = (0.3, 0.7)$. Table 1 shows the average value of the symmetrised Kullback-Leibler divergence for $n_\theta = n_m \in \{100, 500, 1000\}$. The average divergence decreases as the number of simulated data sets increases for our method, in contrast to the synthetic likelihood approach. We can attribute the better performance of our method to its ability to better handle non-Gaussian summary statistics and its ability to select the summary statistics that are relevant.

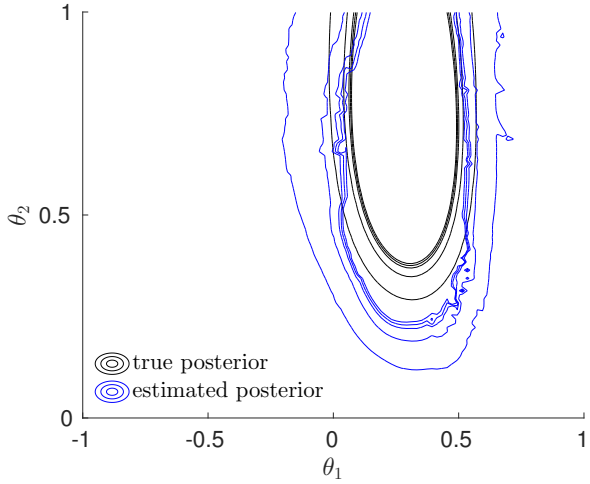
We further compared the performance of the proposed method and synthetic likelihood case-by-case for the 100 different observed data sets. For that purpose, we computed the difference Δ_{sKL} between $sKL(\hat{p}(\theta|x_0)||p(\theta|x_0))$ when $\hat{p}(\theta|x_0)$ is estimated by the proposed method and by synthetic likelihood. A value of $\Delta_{\text{sKL}} < 0$ indicates a better performance of the proposed method while a value $\Delta_{\text{sKL}} > 0$ indicates that synthetic likelihood is performing better. As Δ_{sKL} depends on x_0, it is a random variable and we can compute its empirical distribution on the 100 different inference problems corresponding to different



(a) synthetic likelihood



(b) proposed method



(c) proposed method subject to noise

Figure 3: ARCH(1): Contour plots of the posterior $\hat{p}(\theta|x_0)$ estimated by (a) synthetic likelihood, (b) the proposed method, and (c) the proposed method subject to 50% irrelevant summary statistics. The range of the axes indicates the domain of the uniform prior. We used $n_\theta = n_m = 1000$ for all results. The proposed approach yields a better approximation than the synthetic likelihood approach and remains stable in the presence of irrelevant summary statistics.

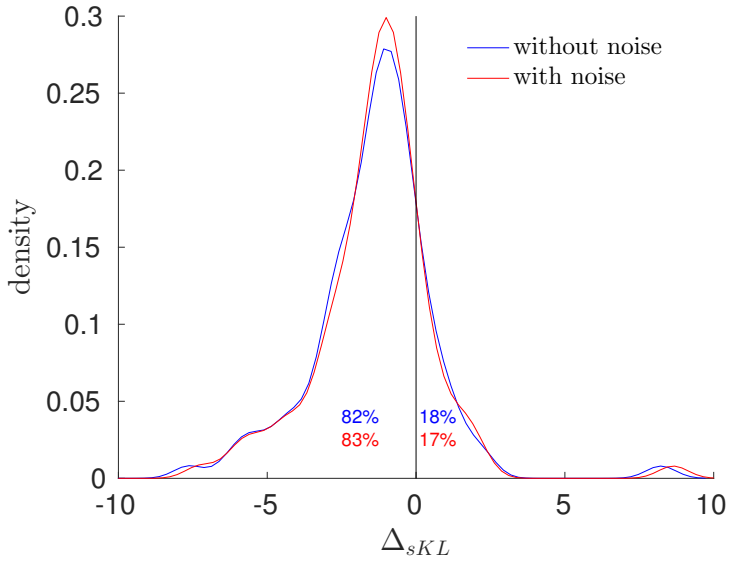


Figure 4: ARCH(1): Estimated density of the difference Δ_{sKL} between the symmetrised Kullback-Leibler divergence of the proposed method and synthetic likelihood for $n_\theta = n_m = 1000$. A negative value of Δ_{sKL} indicates that the proposed method has a smaller divergence and thus is performing better. Depending on whether white noise summary statistics are absent (blue) or present (red) in the proposed method, it is 82% or 83% better than synthetic likelihood. The densities were estimated with a Gaussian kernel density estimator with bandwidth 0.5.

observed data sets. Figure 4 shows the distribution of Δ_{sKL} when the white noise variables are absent (blue) and present (red) for the proposed method. The area under the curve on the negative-side of the x-axis is 82% (white-noise absent) and 83% (white-noise present), which indicates a superior performance of the proposed method over synthetic likelihood and robustness to the perturbing irrelevant summary statistics.

6 Bayesian inference for nonlinear dynamical systems

We here apply Algorithm 1 to two realistic models with intractable likelihood functions and compare the inference results with the results for the synthetic likelihood approach by Wood (2010). The first one is the ecological model of Ricker (1954) that was also previously used by Wood (2010). The second one is the widely used weather prediction model of Lorenz

(1995) with a stochastic reparametrisation (Wilks, 2005), which we simply call "Lorenz model". Both are time series models, and the inference is difficult due to unobserved variables and their strongly nonlinear dynamics.

6.1 Models

Ricker model. This is a model from ecology that describes the size of some animal population over time. The observed population size at time t, $y^{(t)}$, is assumed to be a stochastic observation of the actual but unobservable population size $N^{(t)}$. Conditional on $N^{(t)}$, the observable $y^{(t)}$ is assumed Poisson distributed,

$$y^{(t)}|N^{(t)}, \phi \sim \text{Poisson}(\phi N^{(t)}), \tag{26}$$

where ϕ is a scaling parameter. The dynamics of the unobservable population size $N^{(t)}$ is described by a stochastic version of the Ricker map (Ricker, 1954),

$$\log N^{(t)} = \log r + \log N^{(t-1)} - N^{(t-1)} + \sigma e^{(t)}, \qquad t = 1, \ldots, T, \qquad N^{(0)} = 0, \tag{27}$$

where $T = 50$, $e^{(t)}$ are independent standard normal random variables, $\log r$ is related to the log population growth rate, and σ is the standard deviation of the innovations. The model has in total three parameters $\theta = (\log r, \sigma, \phi)$. The observed data x_0 are the time-series $(y^{(t)}, t = 1, \ldots, T)$, generated using $\theta^0 = (\log r^0, \sigma^0, \phi^0) = (3.8, 0.3, 10)$. We have assumed uniform prior for all parameters: $\mathcal{U}(3, 5)$ for $\log r$, $\mathcal{U}(0, 0.6)$ for σ, and $\mathcal{U}(5, 15)$ for ϕ.

For our method, we use the set of 13 summary statistics ϕ suggested by Wood (2010) as well as all their pairwise combinations and a constant in order to make the comparison with synthetic likelihood fair – as pointed in Section 4, synthetic likelihood implicitly uses the pairwise combinations of the summary statistics due to its underlying Gaussianity assumption. The set of 13 summary statistics ϕ are: the mean observation $\bar{y}$, the number of zero observations, auto-covariances with lag one up to five, the coefficients of a cubic regression of the ordered differences $y^{(t)} - y^{(t-1)}$ on those of the observed data, and the least squares estimates of the coefficients for the model $(y^{(t+1)})^{0.3} = b_1(y^{(t)})^{0.3} + b_2(y^{(t)})^{0.6} + \epsilon^{(t)}$, see (Wood, 2010) for details.

Lorenz model. This model is a modification of the original weather prediction model of Lorenz (1995) when fast weather variables are unobserved (Wilks, 2005). The model

assumes that weather stations measure a high-dimensional time-series of slow weather variables $(y_k^{(t)}, k = 1, \ldots, 40)$, which follow a coupled stochastic differential equation (SDE), called the forecast model (Wilks, 2005),

$$\frac{dy_k^{(t)}}{dt} = -y_{k-1}^{(t)}(y_{k-2}^{(t)} - y_{k+1}^{(t)}) - y_k^{(t)} + F - g(y_k^{(t)}, \theta) + \eta_k^{(t)} \tag{28}$$

$$g(y_k^{(t)}, \theta) = \sum_{i=1}^{2} \theta_i \left(y_k^{(t)}\right)^{i-1}, \tag{29}$$

where $\eta_k^{(t)}$ is stochastic and represents the uncertainty due to the forcing of the unobserved fast weather variables. The function $g(y_k^{(t)}, \theta)$ represents the deterministic net effect of the unobserved fast variables on the observable $y_k^{(t)}, k = 1, \ldots, 40$, and $F = 10$. The model is cyclic in the variables $y_k^{(t)}$, e.g. in Equation (28) for $k = 1$ we have $k - 1 = 40$ and $k - 2 = 39$. We assume that the initial values $y_k^{(0)}, k = 1, \ldots, 40$ are known, and that the model is such that the time interval $[0, 4]$ corresponds to 20 days.

The above set of coupled SDEs does not have an analytical solution. We discretised the 20 days time-interval $[0, 4]$ into $T = 160$ equal steps of $\Delta t = 0.025$, equivalent to 3 hours, and solved the SDEs by using a 4th order Runge-Kutta solver at these time-points (Carnahan *et al.*, 1969, Section 6.5). In the discretised SDEs, following Wilks (2005), the stochastic forcing term is updated for an interval of Δt as

$$\eta_k^{(t+\Delta t)} = \phi\eta_k^{(t)} + \sqrt{1-\phi^2}e^{(t)},\ t \in \{0, \Delta t, 2\Delta t, \ldots, 160\Delta t\},$$

where the $e^{(t)}$ are independent standard normal random variables and $\eta^{(0)} = \sqrt{1-\phi^2}e^{(0)}$.

The inference problem that we solve here is the estimation of the posterior distribution of the parameters $\theta = (\theta_1, \theta_2)$, called closure parameters in weather modelling, from the 40 slow weather variables $y_k^{(t)}$, recorded over twenty days. We simulated such observed data x_0 from the model by solving the SDEs numerically as described above with $\theta^0 = (\theta_1^o, \theta_2^o) = (2.0, 0.1)$ over a period of twenty days. The uniform priors assumed for the parameters were $\mathcal{U}(0.5, 3.5)$ for θ_1 and $\mathcal{U}(0, 0.3)$ for θ_2.

For the inference of the closure parameters θ of the Lorenz model, Hakkarainen *et al.* (2012) suggested six summary statistics: (1) the mean of $y_k^{(t)}$, (2) the variance of $y_k^{(t)}$, (3) the auto-co-variance of $y_k^{(t)}$ with time lag one, (4) the co-variance of $y_k^{(t)}$ with its neighbour $y_{k+1}^{(t)}$, and (5, 6) the cross-co-variance of $y_k^{(t)}$ with its two neighbours $y_{k-1}^{(t)}$ and $y_{k+1}^{(t)}$ for time

lag one. These values were computed and averaged over all k due to the symmetry in the model. We used the six summary statistics for synthetic likelihood, and, to make the comparison fair, for the proposed method, we also used their pairwise combinations as well as a constant like in the previous sections.

6.2 Results

We used an importance sampling scheme (Ripley, 1987, IS) by sampling 8000 samples from the prior distribution and computed their weights using Algorithm 1, which is equivalent to one generation of the SMC algorithm (Cappé *et al.*, 2004; Del Moral *et al.*, 2006, SMC). As suggested by Wood (2010), for the synthetic likelihood approach we used a robust variance-covariance matrix estimation scheme for a better estimation of the likelihood function.

Figure 5 shows example results for the Ricker model, and Figure 6 example results for the Lorenz model. While the results look reasonable, assessing their accuracy rigorously is difficult due to the intractability of the likelihood functions and the lack of ground truth posterior distributions. We here used the (relative) error between the posterior mean and the true data generating parameter value to gauge the accuracy of the inference.

The posterior mean $\hat{\mathbb{E}}(\theta|x_0)$ for the proposed approach and $\hat{\mathbb{E}}_{\mathrm{SL}}(\theta|x_0)$ for the synthetic likelihood approach were computed from the posterior samples, and their relative errors were computed for each element of the parameter vector θ,

$$\mathcal{RE}(x_0) = \sqrt{\frac{(\hat{\mathbb{E}}(\theta|x_0) - \theta^0)^2}{(\theta^0)^2}}, \qquad \mathcal{RE}_{\mathrm{SL}}(x_0) = \sqrt{\frac{(\hat{\mathbb{E}}_{\mathrm{SL}}(\theta|x_0) - \theta^0)^2}{(\theta^0)^2}}. \tag{30}$$

The squaring and division should be understood as element-wise operations. As the relative error depends on the observed data x_0, we computed the error for 100 different observed datasets. We performed a point-wise comparison between the proposed method and synthetic likelihood by computing the difference $\Delta_{\mathrm{rel-error}}$ between the relative errors for all elements in the parameter vector θ,

$$\Delta_{\mathrm{rel-error}} = \mathcal{RE}(x_0) - \mathcal{RE}_{\mathrm{SL}}(x_0). \tag{31}$$

A value of $\Delta_{\mathrm{rel-error}} < 0$ means that the relative error for the proposed method is smaller than the relative error for the synthetic likelihood approach. A value of $\Delta_{\mathrm{rel-error}} > 0$, on

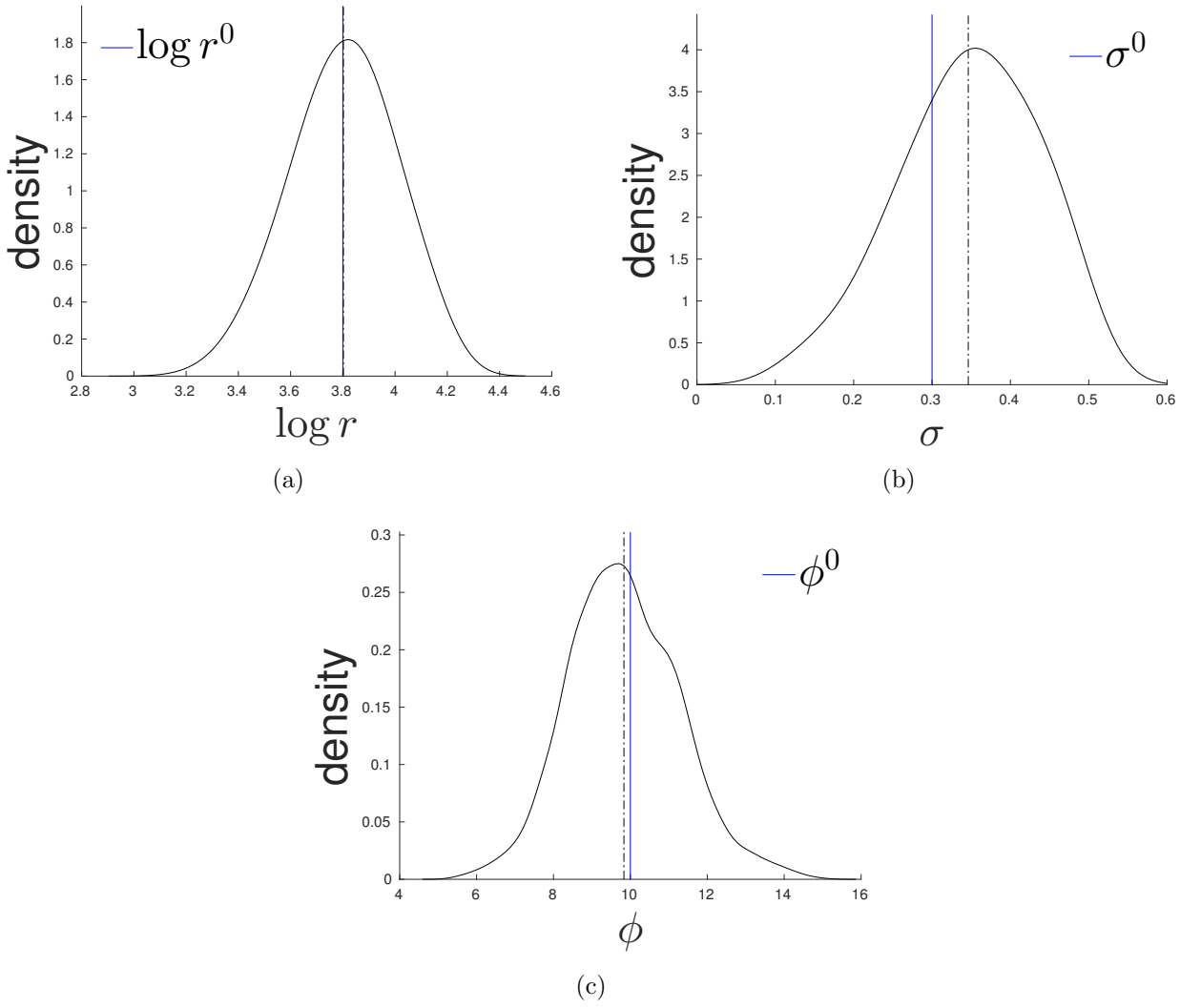


Figure 5: Ricker model: Example marginal posterior distribution of (a) $\log r$, (b) σ and (c) ϕ, estimated with Algorithm 1 using $n_\theta = n_m = 100$. The blue vertical lines show the true parameter values $(\log r^0, \sigma^0, \phi^0)$ that we used to simulate the observed data and the black-dashed vertical lines show the corresponding estimated posterior means. The densities in (a-c) were estimated from posterior samples using a Gaussian kernel density estimator with bandwidths 0.1, 0.04, and 0.3, respectively.

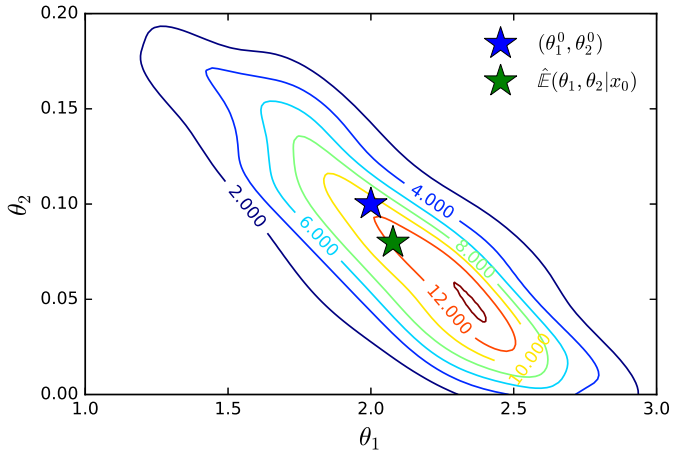
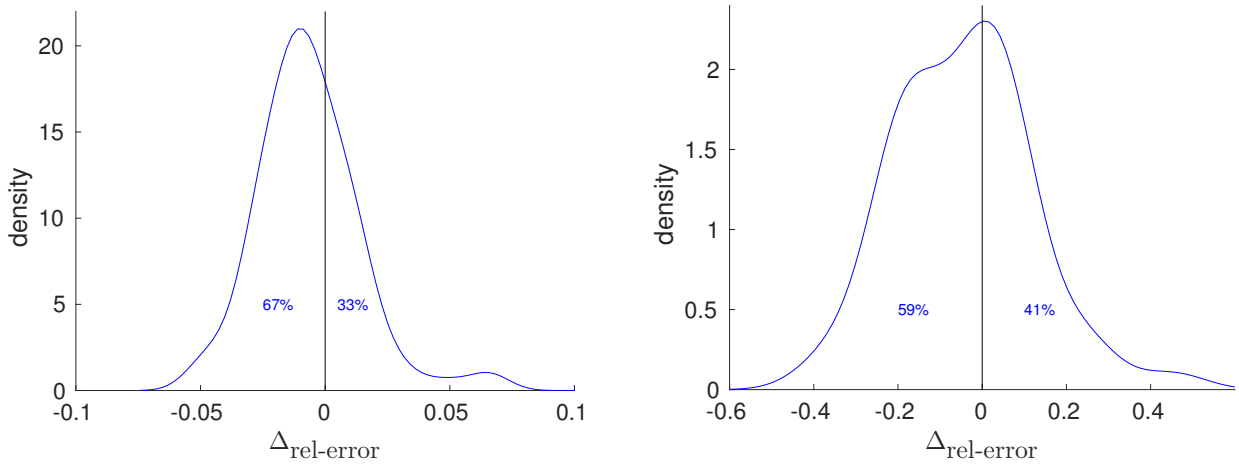


Figure 6: Lorenz model: Example posterior distribution of the closure parameters (θ_1, θ_2) estimated with Algorithm 1 using $n_\theta = n_m = 100$. The blue and green asterisk indicate the true parameter values (θ_1^0, θ_2^0) that were used to simulate the observed data and the estimated posterior mean of the parameters, respectively. The contour plot was generated from posterior samples by a weighted Gaussian kernel density estimator with bandwidth 0.5.

the other hand, indicates that the synthetic likelihood is performing better. As $\Delta_{\text{rel}-\text{error}}$ is a function of x_0, in Figures 7 and 8 we show the empirical distribution of $\Delta_{\text{rel}-\text{error}}$ computed from the 100 different observed data sets x_0. The distribution is tilted toward negative values of $\Delta_{\text{rel}-\text{error}}$ for all the parameters in both the Ricker and the Lorenz model, which indicates that the proposed method is performing better in both applications. As the proposed and the synthetic likelihood method use exactly the same summary statistics, we do not expect large improvement in the performance. Nevertheless, it can be seen that the proposed method is able to reduce the error for most parameters, including the difficult σ parameter of the Ricker model.

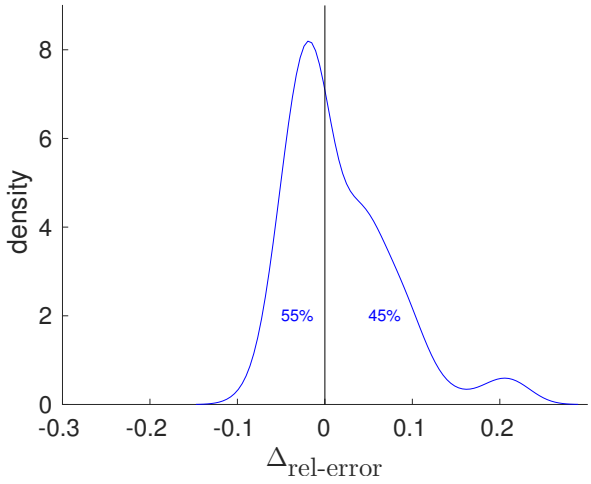
We next analysed the impact of the improved inference on weather prediction, which is the main area of application of the Lorenz model. Having observed weather variables for $t \in [0, 4]$, or 20 days, we would like to predict the weather of the next days. We here consider prediction over a horizon of ten days, which corresponds to $t \in [4, 6]$.

Given x_0, we first estimated the posterior mean of the parameters using the proposed



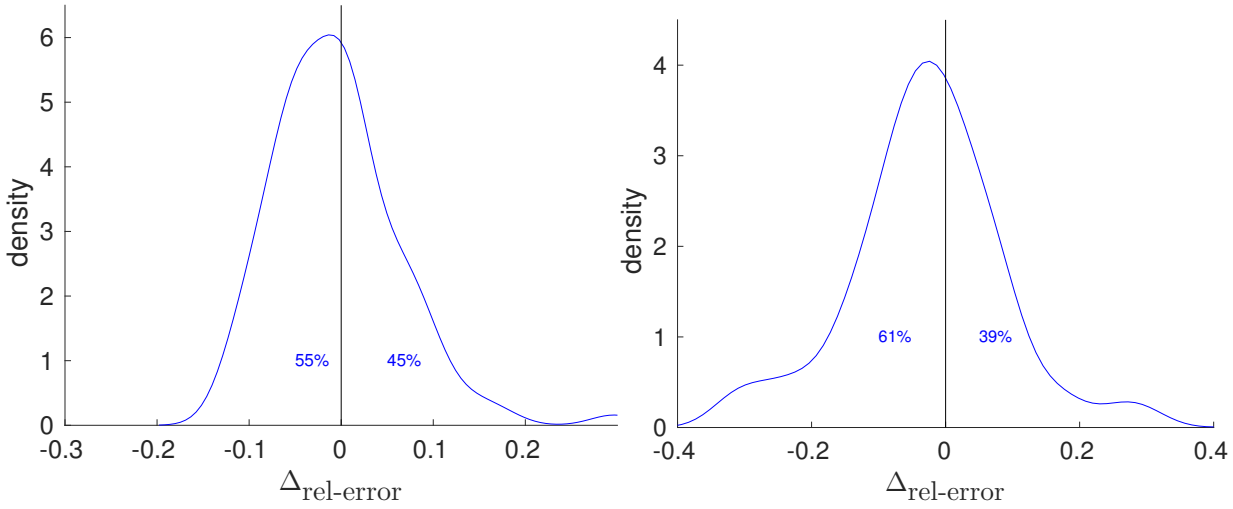
(a) $\log r$

(b) σ



(c) ϕ

Figure 7: Ricker model: Empirical pdf of $\Delta_{\text{rel}-\text{error}}$ for the parameters (a) $\log r$, (b) σ and (c) ϕ. More area under the curve on the negative side of the x-axis indicates a better performance of the proposed method compared to the synthetic likelihood. We used Algorithm 1 and synthetic likelihood with $n_\theta = n_m = 100$ to estimate the posterior pdf. The densities in (a-c) were estimated using a Gaussian kernel density estimator with bandwidth 0.01, 0.07 and 0.02, respectively.



(a) θ_1 (b) θ_2

Figure 8: Lorenz model: Empirical pdf of $\Delta_{\text{rel}-\text{error}}$ for the parameters (a) θ_1 and (b) θ_2. More area under the curve on the negative side of the x-axis indicates a better performance of the proposed method. We used Algorithm 1 and synthetic likelihood with $n_\theta = n_m = 100$ to estimate the posterior pdf. The densities in (a-b) were estimated using a Gaussian kernel density estimator with bandwidth 0.025 and 0.037, respectively.

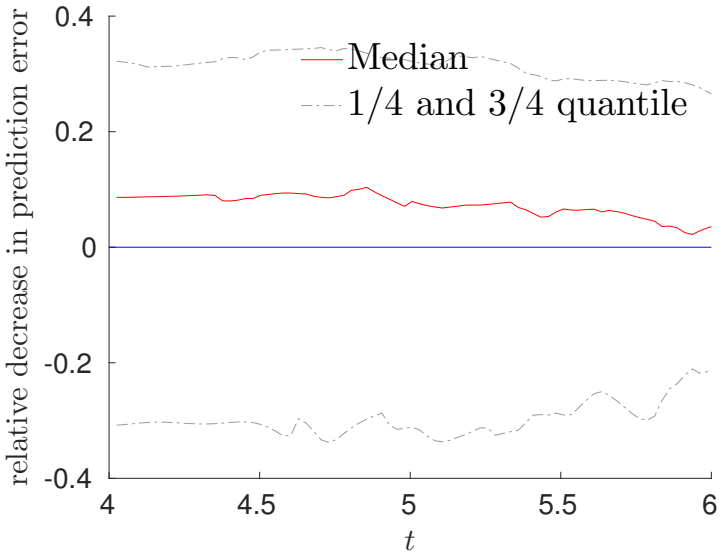


Figure 9: Lorenz Model: Median, 1/4 and 3/4 quantile of the relative decrease in the prediction error $\zeta^{(t)}$ for $t \in [4, 6]$ corresponding to 1 to 10 days in the future. We used Algorithm 1 and synthetic likelihood with $n_\theta = n_m = 100$ to estimate the posterior pdf. As the median is always positive, the proposed method obtains, on average, a smaller prediction error than synthetic likelihood.

and the synthetic likelihood approach. Taking the final values of the observed data $(y_k^{(4)}, k = 1, \ldots, 40)$ as initial values, we then simulated the future weather development using the SDE in Equation (28) for both the true parameter value θ^0, as well as for the two competing sets of estimates. Let us denote the 40-dimensional time series corresponding to θ^0, $\hat{\mathbb{E}}(\theta|x_0)$ and $\hat{\mathbb{E}}_{\mathrm{SL}}(\theta|x_0)$ at time t by $y^{(t)}$, $\hat{y}^{(t)}$, and $\hat{y}_{\mathrm{SL}}^{(t)}$, respectively. We then compared the proposed and the synthetic likelihood method by comparing their prediction error. Denoting the Euclidean norm of a vector by $||.||$, we computed

$$\zeta^{(t)}(x_0) = \frac{||y^{(t)} - \hat{y}_{\mathrm{SL}}^{(t)}|| - ||y^{(t)} - \hat{y}^{(t)}||}{||y^{(t)} - \hat{y}_{\mathrm{SL}}^{(t)}||}, \quad t \in (4, 6], \tag{32}$$

which measures the relative decrease in the prediction error achieved by the proposed method over synthetic likelihood. As the estimates depend on the observed data x_0, $\zeta^{(t)}(x_0)$ depends on x_0. We assessed its distribution by computing its values for 100 different x_0.

In Figure 9, we plot the median, the 1/4 and the 3/4 quantile of $\zeta^{(t)}(x_0)$ for $t \in [4, 6]$ corresponding to one to ten days in the future. We achieve on average a clear improvement in prediction performance for the first days; for longer-term forecasts, the improvement

becomes smaller, which is due the inherent difficulty to make long-term predictions for chaotic time series.

7 Discussion

In the paper, we considered the problem of estimating the posterior density when the likelihood function is intractable but generating data from the model is possible. We framed the posterior density estimation problem as a density ratio estimation problem. The latter problem can be solved by (nonlinear) logistic regression and is thus related to classification. This approach for posterior estimation with generative models mirrors the approach of Gutmann and Hyvärinen (2012) for the estimation of unnormalised models. The main difference is that here, we classify between two simulated data sets while Gutmann and Hyvärinen (2012) classified between the observed data and simulated reference data. This difference reflects the fact that generating samples is relatively easy for generative models while typically difficult for unnormalised models.

For unnormalised models, Gutmann and Hyvärinen (2012) found that increasing the size of the reference data set increases the accuracy of the estimates. We here worked with two equally sized data sets. But the data set with samples from the marginal can be made larger without much overhead as it is only generated once, which will likely lead to more accurate posterior estimates.

In our previous work (Gutmann *et al.*, 2014, 2017), we showed how to perform likelihood-free inference via classification. The difference to the current work is that we there classified between observed and simulated data, and not between two simulated data sets as done here. The key advantage of working with two simulated data sets is that it supports posterior inference given a single observed datum only, as we are guaranteed to have enough data to train the classifier. Goodfellow *et al.* (2014) classified between observed and simulated data as well, in order to generate random samples via a neural network. Our approach thus distinguishes itself not only in the different goals but also in the fact that we work with two simulated data sets.

Working with two simulated data sets has the additional advantage that most computations can be performed offline before the observed data are seen, and that computations

can be recycled for newly observed data sets, which enables "crowd-sourcing" of computations. This kind of (shared) pre-computations can be particularly advantageous when the posterior needs to be estimated as part of a decision making process that is subject to time constraints.

Our method requires that several samples from the model are generated for the estimation of the posterior at any parameter value, like for synthetic likelihood (Wood, 2010). While the sampling can be performed perfectly in parallel, it constitutes the main computational cost. There are several ways to reduce it: First, the proposed approach is compatible with likelihood-free inference that uses Bayesian optimisation to intelligently decide where to evaluate the posterior (Gutmann and Corander, 2016), thus reducing unnecessary computations. Second, we can learn the relation between the parameters and the weights in the logistic regression model from the outcomes of the regressions performed for previous parameter values. An initial estimate of the posterior can thereby be obtained without any new sampling from the model, and additional computations may only be spent on fine-tuning that estimate. Third, for prior distributions much broader than the posterior, performing logistic regression with samples from the marginal distribution is not very efficient. Iteratively constructing a proposal distribution that is closer to the posterior, e.g. by re-using data sets and parameter values from previous simulations, will likely lead to computational gains. The estimated posterior will not be properly normalised but this is typically not a problem when used as the target distribution for posterior sampling by Monte Carlo algorithms.

Density ratio estimation has previously been used by Izbicki *et al.* (2014) and Cranmer *et al.* (2015) to approximate likelihood functions and to estimate likelihood-ratios in the framework of hypothesis testing, respectively. In other recent work, Pham *et al.* (2014) used classification to estimate the ratio of the likelihoods of two parameters appearing in the acceptance probability of the Metropolis-Hastings sampling scheme. If we used the approximate posterior distribution in Equation (11) to estimate the acceptance probability, we would end up with a similar density ratio as Pham *et al.* (2014). A key difference is that our approach results in estimates of the posterior and not in a single accepted, or rejected, parameter value, which is arguably less informative. Additionally, the MCMC approach is

computationally often not very efficient for likelihood-free inference, since long simulations from the model may nonetheless result in rejected parameter values. Importantly, our work goes beyond using density ratio estimation for posterior estimation in that we show how it can be used to automatically combine and select relevant summary statistics from a large pool of candidates.

The existing work related to summary statistics selection in likelihood-free inference (Aeschbacher *et al.*, 2012; Fearnhead and Prangle, 2012; Blum *et al.*, 2013; Gutmann *et al.*, 2014, 2017; Marin *et al.*, 2016) are generally concentrated on ABC. Our approach is more closely related to the synthetic likelihood approach by Wood (2010), including it as a special case. While the cited methods for summary statistics selection in ABC might be adaptable for use with synthetic likelihood, the summary statistics generally have to be transformed before use, in order to match the Gaussianity assumption of synthetic likelihood. This is in contrast to our approach that automatically adapts to non-Gaussianity of the summary statistics.

Our method is simple to use and relies on standard statistics or machine learning libraries. Ease of use is shared with existing likelihood-free inference methods that are based on a Gaussianity assumption of the summary statistics or the data generating process (Wood, 2010; Fan *et al.*, 2013; Price *et al.*, 2016). But even the simple approach of using a linear basis expansion in the logistic regression yields a richer statistical model than the Gaussian family, in the form of the exponential family. The advantage of working with the exponential family has been demonstrated on a number of applications, including an auto-regressive time series model (Section 5.2) and stochastic nonlinear dynamical systems (Section 6). Moreover, we are not restricted to (penalised) linear logistic regression. Our approach can be used with more general regression models too, as portrayed in Figure 1.

Gutmann and Hirayama (2011) used the Bregman divergence to estimate ratios of probability distributions thereby introducing a large family of estimators for unnormalised models, containing the method by Gutmann and Hyvärinen (2012) as a special case. Drawing on the discussed connection between (Gutmann and Hyvärinen, 2012) and the paper at hand suggests that an equally large family of inference methods can be obtained for the estimation of generative models, containing our method based on logistic regression as a

special case.

In conclusion, our paper opens up a direction to new likelihood-free inference methods based on logistic regression or other density ratio estimation schemes that can be used whenever the likelihood function is not available but sampling from the model is possible.

Acknowledgements

The work was partially done when RD and MUG were at the Department of Computer Science, Aalto University, and the Department of Mathematics and Statistics, University of Helsinki, respectively. The work was financially supported by the Academy of Finland (grants 294238 and 292334, and the Finnish Centre of Excellence in Computational Inference Research COIN). The authors thank Chris Williams for helpful comments and gratefully acknowledge the computational resources provided by the Aalto Science-IT project. RD is presently funded by Swiss National Science Foundation grant *no.*105218_163196.

Author contributions Idea and theory: MUG; implementation and simulations: RD; writing of the paper: MUG and RD; contributed to writing: JC and SK.

A Proof of Equation (8)

We here prove that $\log r(x,\theta) = \log\left(p(x|\theta)/p(x)\right)$ minimises $\mathcal{J}(h,\theta)$ in Equation (7) in the limit of large n_θ and n_m.

We first simplify the notation and denote x^θ by x, its pdf $p(x|\theta)$ by p_x, n_θ by n, x^m by y, its pdf $p(x)$ by p_y, and n_m by m. Moreover, as θ is considered fixed for this step, we drop the dependency of $\mathcal{J}$ on θ. Equation (7) thus reads

$$\mathcal{J}(h) = \frac{1}{n+m}\left\{\sum_{i=1}^{n}\log\left[1+\nu\exp(-h(x_i))\right] + \sum_{i=1}^{m}\log\left[1+\frac{1}{\nu}\exp(h(y_i))\right]\right\}. \tag{33}$$

We will consider the limit where n and m are large, with fixed ratio $\nu = m/n$. For that purpose we write $\mathcal{J}$ as

$$\mathcal{J}(h) = \frac{n}{n+m}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\nu\exp(-h(x_i))\right] + \frac{1}{n}\sum_{i=1}^{m}\log\left[1+\frac{1}{\nu}\exp(h(y_i))\right]\right\} \tag{34}$$

$$= \frac{n}{n+m}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\nu\exp(-h(x_i))\right]+\nu\frac{1}{m}\sum_{i=1}^{m}\log\left[1+\frac{1}{\nu}\exp(h(y_i))\right]\right\} \quad (35)$$

$$= \frac{1}{1+\nu}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\nu\exp(-h(x_i))\right]+\nu\frac{1}{m}\sum_{i=1}^{m}\log\left[1+\frac{1}{\nu}\exp(h(y_i))\right]\right\}. \quad (36)$$

In the stated limit, $\mathcal{J}(h)$ thus equals $\mathcal{J}(h) = \tilde{\mathcal{J}}(h)/(1+\nu)$, where

$$\tilde{\mathcal{J}}(h) = \mathbb{E}_x \log\left[1+\nu\exp(-h(x))\right] + \nu\mathbb{E}_y \log\left[1+\frac{1}{\nu}\exp(h(y))\right]. \quad (37)$$

The function h^* that minimises $\tilde{\mathcal{J}}(h)$ also minimises $\mathcal{J}(h)$ in the limit of large n and m. To determine h^* we apply

$$\log\left(1+\frac{1}{\nu}\exp h\right) = \log(\nu\exp(-h)+1) - \log(\nu\exp(-h)) \quad (38)$$

and re-write $\tilde{\mathcal{J}}$ as

$$\begin{aligned}
\tilde{\mathcal{J}}(h) =& \mathbb{E}_x \log(1+\nu\exp(-h(x))) + \nu\mathbb{E}_y \log(\nu\exp(-h(y))+1) \\
&- \nu\mathbb{E}_y \log(\nu\exp(-h(y))) \quad (39)\\
=& \mathbb{E}_x \log(1+\nu\exp(-h(x))) + \nu\mathbb{E}_y \log\left(1+\nu\exp(-h(y))\right) \\
&- \nu\log\nu + \nu\mathbb{E}_y h(y) \quad (40)\\
=& \int p_x(u)\log(1+\nu\exp(-h(u)))du + \nu\int p_y(u)\log(1+\nu\exp(-h(u)))du \\
&- \nu\log\nu + \nu\int p_y(u)h(u)du \quad (41)\\
=& \int (p_x(u)+\nu p_y(u))\log(1+\nu\exp(-h(u)))du \\
&- \nu\log\nu + \nu\int p_y(u)h(u)du. \quad (42)
\end{aligned}$$

We now expand $\tilde{\mathcal{J}}(h+\epsilon q)$ around h for an arbitrary function q and a small scalar ϵ.With

$$\begin{aligned}
\log(1+\nu\exp\left(-h(u)-\epsilon q(u)\right)) =& \log(1+\nu\exp(-h(u)) \\
&- \epsilon q(u)\frac{\nu\exp(-h(u))}{1+\nu\exp(-h(u))} \\
&+ \frac{\epsilon^2 q(u)^2}{2}\frac{\nu\exp(-h(u))}{1+\nu\exp(-h(u))}\frac{1}{1+\nu\exp(-h(u))} \\
&+ O(\epsilon^3) \quad (43)
\end{aligned}$$

we have

$$\begin{aligned}\tilde{\mathcal{J}}(h+\epsilon q) = &\int (p_x(u)+\nu p_y(u))\log(1+\nu\exp(-h(u))du \\ &- \int (p_x(u)+\nu p_y(u))\,\epsilon q(u)\frac{\nu\exp(-h(u))}{1+\nu\exp(-h(u))}du \\ &+ \int (p_x(u)+\nu p_y(u))\,\frac{\epsilon^2 q(u)^2}{2}\frac{\nu\exp(-h(u))}{1+\nu\exp(-h(u))}\frac{1}{1+\nu\exp(-h(u))}du \\ &- \nu\log\nu + \nu\int p_y(u)h(u)du + \nu\int p_y(u)\epsilon q(u)du + O(\epsilon^3). \end{aligned} \tag{44}$$

Collecting terms gives

$$\begin{aligned}\tilde{\mathcal{J}}(h+\epsilon q) = &\tilde{\mathcal{J}}(h) - \epsilon\int q(u)\left((p_x(u)+\nu p_y(u))\,\frac{\nu\exp(-h(u))}{1+\nu\exp(-h(u))} - \nu p_y(u)\right)du \\ &+ \frac{\epsilon^2}{2}\int q(u)^2\,(p_x(u)+\nu p_y(u))\,\frac{\nu\exp(-h(u))}{1+\nu\exp(-h(u))}\frac{1}{1+\nu\exp(-h(u))}du \\ &+ O(\epsilon^3). \end{aligned} \tag{45}$$

The second-order term is positive for all (non-trivial) q and h. The first-order term is zero for all q if and only if

$$\nu p_y(u) = \frac{p_x(u)+\nu p_y(u)}{1+\frac{1}{\nu}\exp(h^*(u))} \quad \Leftrightarrow \quad \nu p_y(u) + p_y(u)\exp(h^*(u)) = p_x(u) + \nu p_y(u) \tag{46}$$

that is, if and only if

$$\exp(h^*(u)) = \frac{p_x(u)}{p_y(u)}, \tag{47}$$

which shows that $h^* = \log(p_x/p_y)$ minimises $\tilde{\mathcal{J}}$. With the notation from the main text, $h^* = \log(p(x|\theta)/p(x))$, which equals $\log r(x,\theta)$, and thus proves the claim.

References

Aeschbacher, S., Beaumont, M., and Futschik, A. (2012). A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, **192**(3), 1027–1047.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, **37**(2), 697–725.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**(1), 379–406.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**(4), 2025–2035.

Bickel, S., Brückner, M., and Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 81–88, New York, NY, USA. ACM.

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, **28**(2), 189–208.

Cappé, O., Guillin, A., Marin, J.-M., and Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, **13**(4), 907–929.

Carnahan, B., Luther, H. A., and Wilkes, J. O. (1969). *Applied Numerical Methods*. Wiley, New York.

Cheng, K. and Chu, C. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, **10**(4), 583–604.

Cranmer, K., Pavez, J., and Louppe, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. *ArXiv:1506.02169*.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Royal Statistical Society: Series B (Statistical Methodology)*, **68**(3), 411–436.

Dutta, R., Bogdan, M., and Ghosh, J. (2012). Model selection and multiple testing - a Bayesian and empirical Bayes overview and some new results. *Journal of Indian Statistical Association*, **50**, 105–142.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**(2), 407–499.

Fan, Y., Nott, D. J., and Sisson, S. A. (2013). Approximate Bayesian computation via regression density estimation. *Stat*, **2**(1), 34–48.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(3), 419–474.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10**(6), 971–988.

Gutmann, M. and Hirayama, J. (2011). Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 283–290, Corvallis, Oregon. AUAI Press.

Gutmann, M., Dutta, R., Kaski, S., and Corander, J. (2014). Likelihood-free inference via classification. *arXiv:1407.4981*.

Gutmann, M., Dutta, R., Kaski, S., and Corander, J. (2017). Likelihood-free inference via classification. *Statistics and Computing*, **in press**.

Gutmann, M. U. and Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, **17**(125), 1–47.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, **13**, 307–361.

Hakkarainen, J., Ilin, A., Solonen, A., Laine, M., Haario, H., Tamminen, J., Oja, E., and Järvinen, H. (2012). On closure parameter estimation in chaotic systems. *Nonlinear Processes in Geophysics*, **19**(1), 127–143.

Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., and Huth, A. (2011). Statistical inference for stochastic simulation models – theory and application. *Ecology Letters*, **14**(8), 816–827.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Izbicki, R., Lee, A., and Schafer, C. (2014). High-dimensional density ratio estimation with extensions to approximate likelihood computation. In *Proceedings of the 7th International Conference on Artificial Intelligence and Statistics*, volume 33 of *JMLR Proceedings*, pages 420–429. JMLR.org.

Leuenberger, C. and Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. *Genetics*, **184**(1), 243–252.

Lintusaari, J., Gutmann, M., Dutta, R., Kaski, S., and Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, **66**(1), e66–e82.

Lorenz, E. (1995). Predictability: a problem partly solved. In *Proceedings of the Seminar on Predictability, 4-8 September 1995*, volume 1, pages 1–18, Shinfield Park, Reading. European Center on Medium Range Weather Forecasting, European Center on Medium Range Weather Forecasting.

Marin, J.-M., Pudlo, P., Robert, C., and Ryder, R. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, **22**(6), 1167–1180.

Marin, J.M. ans Raynal, L., Pudlo, P., Ribatet, M., and Robert, C. (2016). ABC random forests for Bayesian parameter inference. *arXiv:1605.05537*.

Martinez, E. A., Muschik, C. A., Schindler, P., Nigg, D., Erhard, A., Heyl, M., Hauke, P., Dalmonte, M., Monz, T., Zoller, P., and Blatt, R. (2016). Real-time dynamics of lattice gauge theories with a few-qubit quantum computer. *Nature*, **534**(7608), 516–519.

Mohamed, S. and Lakshminarayanan, B. (2016). Learning in implicit generative models. *arXiv:1610.03483*.

Pham, K. C., Nott, D. J., and Chaudhuri, S. (2014). A note on approximating ABC-MCMC using flexible classifiers. *Stat*, **3**(1), 218–227.

Pihlaja, M., Gutmann, M., and Hyvärinen, A. (2010). A family of computationally efficient and simple estimators for unnormalized statistical models. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Price, L., Drovandi, C., Lee, A., and Nott, D. (2016). Bayesian synthetic likelihood. Unpublished, Queensland University of Technology, Brisbane, Australia, Working paper.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**(12), 1791–1798.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, **85**(3), 619–630.

Ricker, W. E. (1954). Stock and recruitment. *Journal of the Fisheries Research Board of Canada*, **11**(5), 559–623.

Ripley, B. D. (1987). *Stochastic simulation*. John Wiley & Sons Inc., New York, USA.

Schaye, J., Crain, R. A., Bower, R. G., Furlong, M., Schaller, M., Theuns, T., Dalla Vecchia, C., Frenk, C. S., McCarthy, I. G., Helly, J. C., Jenkins, A., Rosas-Guevara, Y. M., White, S. D. M., Baes, M., Booth, C. M., Camps, P., Navarro, J. F., Qu, Y., Rahmati, A., Sawala, T., Thomas, P. A., and Trayford, J. (2015). The eagle project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, **446**(1), 521–554.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press, New York, NY, USA, 1st edition.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, **145**(2), 505–518.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **58**, 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(3), 273–282.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Annals of Statistics*, **40**(2), 1198–1232.

Turchin, P., Currie, T. E., Turner, E. A. L., and Gavrilets, S. (2013). War, space, and the evolution of old world complex societies. *Proceedings of the National Academy of Sciences*, **110**(41), 16384–16389.

van de Geer, S. and Lederer, J. (2013). The lasso, correlated design, and improved oracle inequalities. In M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii, and M. H. Maathuis, editors, *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, volume 9 of *Collections*, pages 303–316. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics*, **36**(2), 614–645.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, **102**(479), 1039–1048.

Wilks, D. S. (2005). Effects of stochastic parametrizations in the Lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society*, **131**(606), 389–407.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, **466**(7310), 1102–1104.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Annals of Statistics*, **35**(5), 2173–2192.