Bayesian Analysis (2020) **TBA**, Number TBA, pp. 1–31

Likelihood-Free Inference by Ratio Estimation

Owen Thomas[*], Ritabrata Dutta[†], Jukka Corander[‡],
Samuel Kaski[§], and Michael U. Gutmann[¶,‖]

Abstract. 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 canonical examples and employ it to perform inference for challenging stochastic nonlinear dynamical systems and high-dimensional summary statistics.

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

1 Introduction

We consider the problem of estimating 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

[*]Department of Biostatistics, University of Oslo, Norway, o.m.t.thomas@medisin.uio.no
[†]Department of Statistics, University of Warwick, UK, ritabrata.dutta@warwick.ac.uk
[‡]Department of Biostatistics, University of Oslo, Norway, jukka.corander@medisin.uio.no
[§]Helsinki Institute for Information Technology, Department of Computer Science, Aalto University, Finland, samuel.kaski@aalto.fi
[¶]School of Informatics, University of Edinburgh, UK, michael.gutmann@ed.ac.uk
[‖]Corresponding author.

 https://doi.org/10.1214/20-BA1238

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, for example in genetics (Pritchard et al., 1999; Arnold et al., 2018), ecology (Wood, 2010; Sirén et al., 2018), or epidemiology of infectious diseases (Tanaka et al., 2006; Corander et al., 2017).

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 Beaumont (2010); Hartig et al. (2011); Marin et al. (2012); Lintusaari et al. (2017); Sisson et al. (2018).

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, 2018). While ABC can be considered to implicitly construct a nonparametric approximation of $p(x|\theta)$ (e.g. Hartig et al., 2011; Lintusaari et al., 2017), a wide range of parametric surrogate models are being used to accelerate the inference or improve its accuracy. The models employed include regression models and neural networks, Gaussian processes as well as normalising flows (Beaumont et al., 2002; Blum, 2010; Wilkinson, 2014; Gutmann and Corander, 2016; Papamakarios and Murray, 2016; Papamakarios et al., 2017, 2019; Chen and Gutmann, 2019). Synthetic likelihood, on the other hand, 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 (Meeds and Welling, 2014; Price et al., 2017), 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 (1) a framework, "LFIRE", and (2) a practical method, "linear LFIRE", to directly approximate the posterior distribution in the absence of a tractable likelihood function.[1] As we will see, the proposed approach includes the

[1]The ideas in this paper were first communicated on arXiv in 2016 (Thomas et al., 2016). The reader may wonder about the several years difference between the arXiv paper and this paper. This is largely due to three review periods that took 8, 9, and 7 months, respectively, and the introduction of a new first author. The core content has stayed the same. We thus would like to ask you to please also acknowledge (Thomas et al., 2016) when citing this paper.

synthetic likelihood as a special case and further enables 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 $p(x)$, in the context of a Bayesian belief update

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

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".

In the LFIRE framework, other distributions than the marginal $p(x)$ can also be used in the denominator, in particular if approximating the likelihood function or identifying its maximiser is the goal. While we do not further address the question of what distributions can be chosen for estimation of the posterior, initially it seems reasonable to prefer distributions that have heavier tails than $p(x|\theta)$ in the numerator because of stability reasons.

Closely related work was done by Pham et al. (2014) and Cranmer et al. (2015) who estimated likelihood ratios. Pham et al. (2014) estimated the ratio between the likelihoods of two parameters appearing in the acceptance probability of the Metropolis-Hastings MCMC sampling scheme. If we used the approximate posterior distribution in (3) to estimate the acceptance probability, we would also end up with a density ratio that can be used for MCMC sampling. A key difference is that our approach results in estimates of the posterior and not in a single accepted, or rejected, parameter value. Cranmer et al. (2015) estimated the ratio between the likelihood at a freely varying parameter value and a fixed reference value in the context of frequentist inference. The goals are thus somewhat different, which, as we will see, account well for the differences in the results in our empirical comparison in Section 5.3.

Since we have first communicated the LFIRE framework as an arXiv paper (Thomas et al., 2016), there have been a number of developments within this framework. For instance, Dinev and Gutmann (2018) tailored the approach to the special case of time-series models, Rogers-Smith et al. (2018) adapted the framework to a sequential population Monte Carlo scheme with adaptive proposals, and Hermans et al. (2020) perform (sequential) LFIRE for amortised likelihood-free MCMC sampling. Durkan et al. (2020) discuss how inference methods in the LFIRE framework relate to conditional density

estimation methods, in particular drawing connections between the work by Hermans et al. (2020) and Greenberg et al. (2019).

There are several methods in the literature available for the estimation of density ratios (e.g. Gutmann and Hirayama, 2011; Sugiyama et al., 2012; Izbicki et al., 2014), of which estimation through logistic regression is widely used and has some favourable asymptotic properties (Geyer, 1994; 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 other computational problems in statistics. Gutmann and Hyvärinen (2012) used this kind of "contrastive learning" to estimate unnormalised models and Goodfellow et al. (2014) employed it for training neural networks to generate samples similar to given reference data. More general methods for ratio estimation are also used 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 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, 2018). While this earlier approach is very general, the classification problem is difficult to set up when the observed data consist of very few data points only. The related work by Pham et al. (2014) and Cranmer et al. (2015) and the method proposed in this paper do not have this shortcoming.

The rest of the paper is organised as follows: Section 2 presents the details on how to generally estimate the ratio $r(x, \theta)$ and hence the posterior by logistic regression. In Section 3, we model the ratio as a linear superposition of summary statistics, yielding the "linear LFIRE" method, and show that this assumption corresponds to 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 canonical examples, and in Sections 6 and 7, we apply it to challenging inference problems in ecology, weather forecasting, and cell proliferation modelling. All simulation studies include a comparison with the synthetic likelihood approach, with their relative computational costs analysed in Section 8. We find that the new method yielded consistently more accurate inference results than synthetic likelihood.

2 Posterior estimation by logistic regression

We here show that the ratio $r(x, \theta)$ in (2) can be estimated by logistic regression, which yields estimates for the posterior and the likelihood function together with (3) and (4). Figure 1 provides an overview.

As we assumed working with a simulator-based model, 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

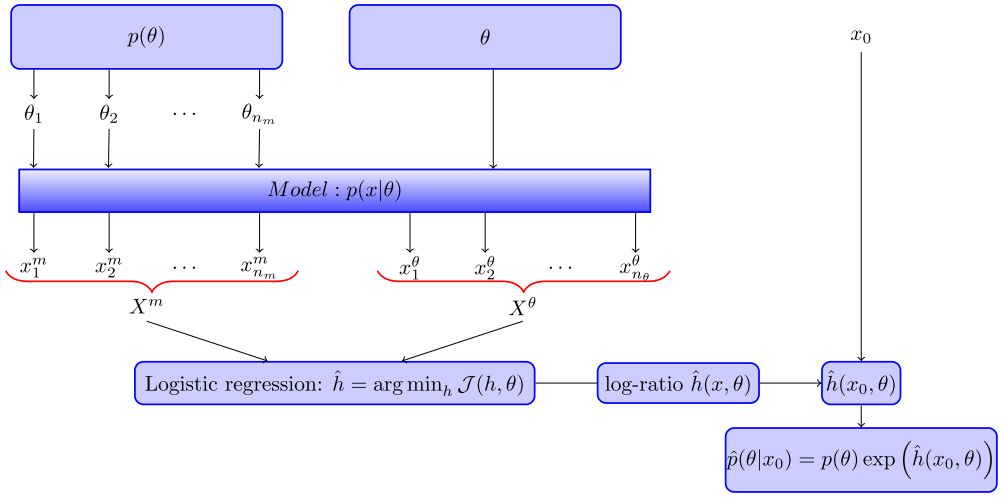


Figure 1: A schematic view of likelihood-free inference by ratio estimation (LFIRE) by means of logistic regression, as explained in (5) to (11).

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 (1), the samples can be obtained by first sampling from the joint distribution of (x, θ) and then ignoring the sampled parameters,

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

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 Supplementary Material A (Thomas et al., 2020) 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

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

The proof shows that this holds for all θ for which we provide data in the classification. This means that we can average the loss function with respect to some distribution $f(\theta)$ defined on a domain where we would like to evaluate the ratio, which corresponds to performing the classification in the joint x, θ space. Choosing for $f(\theta)$ the prior $p(\theta)$, we would classify between data from $p(x|\theta)p(\theta)$ and $p(x)p(\theta)$, as recently done by Hermans et al. (2020), which enables amortisation of the computations with respect to θ.

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 (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), for instance MCMC (Hermans et al., 2020). The estimates can also be used together with Bayesian optimisation (Gutmann and Corander, 2016) or history matching (Wilkinson, 2014) to accelerate the inference. When estimating the posterior or likelihood function 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 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 (Dinev and Gutmann, 2018), we next consider a simple parametric family that is spanned by a set of summary statistics, yielding a particular inference method of the more general LFIRE framework.

3 Exponential family approximation

We here restrict the search in (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). Given the linear nature of the model in (12), we may call this instance of the LFIRE inference principle "linear LFIRE".

When we assume that $h(x)$ takes the functional form in (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,

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

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

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 (12), the estimate for the ratio in (10) thus becomes

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

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

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

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 (15) to provide an implicit estimate $\hat{p}(x|\theta)$ of the intractable model pdf $p(x|\theta)$,

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

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 the 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 (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. 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. Moreover, the selection of summary statistics through regularisation can substantially increase the interpretability of the inference: the number of data summaries identified as relevant may be small enough to be examined by statisticians and model experts individually, providing evidence-based insight into which summary statistics are relevant to the scientific question.

Algorithm 1 Linear LFIRE 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 (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 (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 (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).

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 (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 (e.g. Tibshirani, 1994; Hastie et al., 2001), 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 λ_{min} determines the coefficient $\hat{\beta}(\theta)$,

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

which is used in the estimate of the density ratio in (15), and thus the posterior and likelihood in (17). Algorithm 1 presents pseudo-code that summarises the linear LFIRE 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 (12).

The cross-validation adds computational cost and the dependency of λ_{min} on θ can make more detailed theoretical investigations more difficult. In order to reduce the cost or to facilitate theoretical analyses, working with a fixed value of λ as, for example, An et al. (2019) for synthetic likelihood with the graphical lasso may be appropriate.

5 Validation on canonical low-dimensional problems

We here validate and illustrate the presented theory on a set of canonical inference problems widely considered in the likelihood-free inference literature and empirically compare the proposed approach to an approach based on likelihood ratios.

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 with known variance. The observed data x_0 is a single observation that was sampled from a univariate 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 (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 a ten-dimensional summary statistic $\psi(x) = (1, x, \ldots, x^{b-1})$, with $b = 10$, 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 $\lambda_{\min}$ minimises the prediction error. Repeating step 3 in the algorithm 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 (16).

In Figure 2b, we plot $\hat{\alpha}(\mu)$ and $\alpha_0(\mu)$, $\alpha_1(\mu)$, $\alpha_2(\mu)$ from (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 linear LFIRE by penalised logistic regression and the performance of the existing synthetic likelihood approach are practically equivalent.

5.2 Autoregressive model with conditional heteroskedasticity

In this example, 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)),

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

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., 2018, 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

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

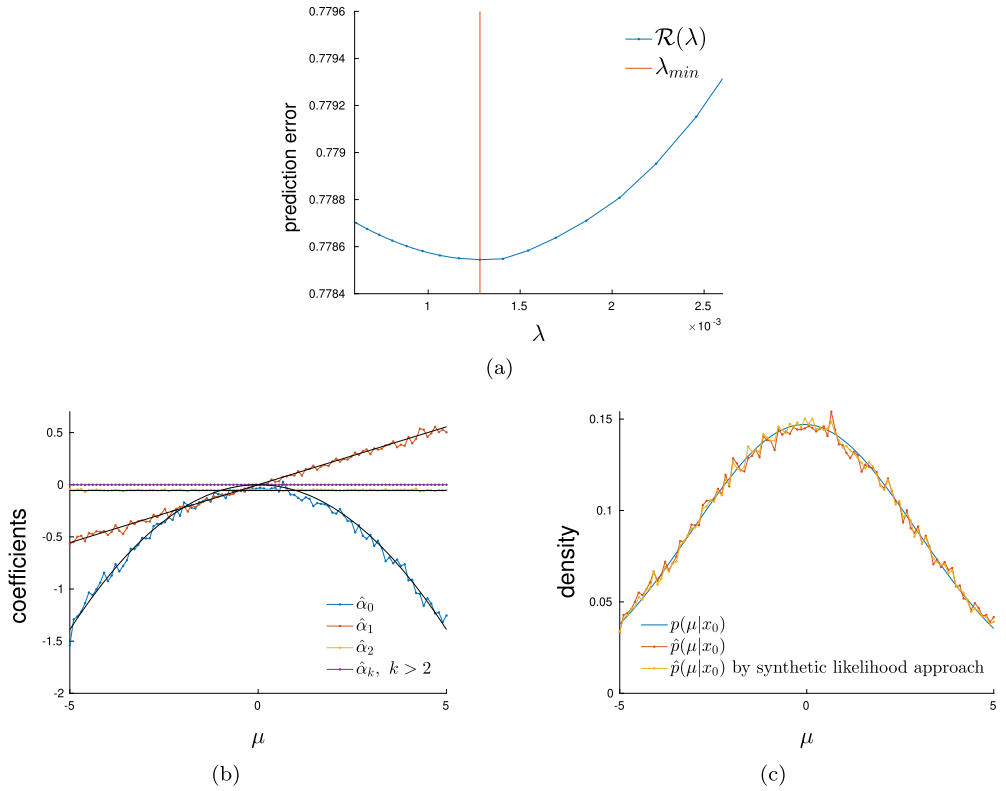
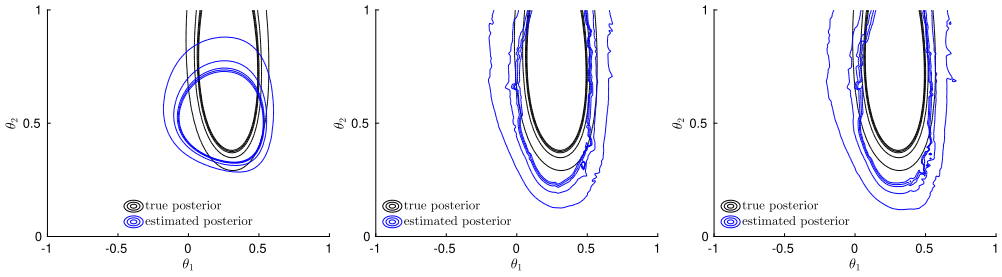


Figure 2: Steps for estimating the posterior distribution of the mean of a Gaussian. (a) For any fixed value of μ, λ_{min} minimises the estimated prediction error $\mathcal{R}(\lambda)$ (vertical line). (b) The figure shows the true coefficients from (21) in black and the coefficients estimated by Algorithm 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.

Instead of comparing to the true posterior, one could compare to an approximate posterior computed by conditioning on the observed value of the summary statistics rather than the full data. We here focus on the comparison to the true posterior in order to assess the overall accuracy. The effect of the employed summary statistics is analysed in Supplementary Material B, and for intractable models considered later in the paper, we construct reference posteriors via expensive rejection ABC runs.

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 a constant. For checking the robustness of the approach, we also considered the case where almost 50% of the summary statistics are irrelevant by augmenting the above set of summary statistics by 15 white-noise random variables. For synthetic likelihood, we



(a) synthetic likelihood (b) linear LFIRE (c) linear LFIRE with irrelevant summaries

Figure 3: ARCH(1): Contour plots of the posterior $\hat{p}(\theta|x_0)$ estimated by (a) synthetic likelihood, (b) the linear LFIRE method in Algorithm 1, and (c) linear LFIRE 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 linear LFIRE approach yields a better approximation than the synthetic likelihood approach and remains stable in the presence of irrelevant summary statistics.

used the auto-correlations as the summary statistics without any additional irrelevant variables, as the 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.

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 linear LFIRE 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. Our approximate posterior provides a reasonable approximation to the exact posterior but we note that it has a larger dispersion. The results in Supplementary Material B suggest that this difference is due to the summary statistics and not the inference method.

In order to assess the performance more systematically, we next performed posterior inference for 100 observed time-series that were each generated from (24) with $\theta^0 = (\theta_1^o, \theta_2^o) = (0.3, 0.7)$. Table 1 in Supplementary Material C 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 linear LFIRE and synthetic likelihood case-by-case for the 100 different observed data sets. For this pairwise performance comparison, we computed the difference Δ_{sKL} between the symmetrised Kullback-Leibler divergences $\text{sKL}(\hat{p}(\theta|x_0)||p(\theta|x_0))$ when $\hat{p}(\theta|x_0)$ is estimated by the proposed method

and by synthetic likelihood. A value of $\Delta_{\mathrm{sKL}} < 0$ indicates a better performance of the proposed method while a value $\Delta_{\mathrm{sKL}} > 0$ 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 observed data sets.

Figure 1 in Supplementary Material D shows the distribution of Δ_{sKL} when the irrelevant 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% (irrelevant summaries absent) and 83% (irrelevant summaries present), which indicates a superior performance of the proposed method over synthetic likelihood and robustness to the perturbing irrelevant summary statistics. The p-values associated with a Wilcoxon signed-rank test ($< 10^{-10}$) demonstrate very strong evidence in favour of the LFIRE method.

Figure 2 in Supplementary Material D shows a scatter plot of the symmetrised Kullback-Leibler divergence for the LFIRE method and for the synthetic likelihood. We see that the substantial majority of simulations fall above the diagonal, indicating better performance of linear LFIRE compared to synthetic likelihood, in line with the above findings.

5.3 Comparison with a frequentist likelihood-ratio based method

Here we compare the LFIRE method with a method based on approximating likelihood ratios with calibrated discriminative classifiers ("carl", Cranmer et al., 2015), which provides an approximate maximum likelihood estimator for a parameter θ by maximising approximations of the ratio $p(x|\theta)/p(x|\theta_r)$, the ratio of the freely parametrised likelihood $p(x|\theta)$ and the likelihood evaluated at a reference value θ_r, $p(x|\theta_r)$. This is done by using a classifier to generate an approximation to the ratio, followed by further calibration by use of kernel density estimation. This corrective calibration step allows one to use a wider range of loss function to train the classifier (see the original paper for details).

The carl method relies on the choice of a reference θ_r to construct the likelihood ratio. It is possible that a choice of θ_r far from the true maximum likelihood estimate (MLE) will provide a very large value of the likelihood ratio with correspondingly high variance, making optimisation very challenging. Even in a frequentist framework, we expect the LFIRE methodology to be more robust to the choice of reference distribution, since samples from the marginal distribution $p(x)$ will be generally drawn from all regions covered by the prior $p(\theta)$. Consequently, with the exception of the unlikely situation of very narrow and mis-specified priors, the estimation of the ratio $p(x|\theta)/p(x)$ should be more stable.

We explore the behaviour of the two methods by estimating the mean of a univariate Gaussian with known variance. Fifty data observations are drawn from the true generative model with mean and variance equal to one. The LFIRE method with $n_\theta = n_m = 100$ was run with different Gaussian prior distributions on the mean parameter, with prior expectation varying between -10 and 10, and prior standard deviations taking values $[0.1, 1., 3., 5., 10.]$. The carl algorithm of Cranmer et al. (2015) was run

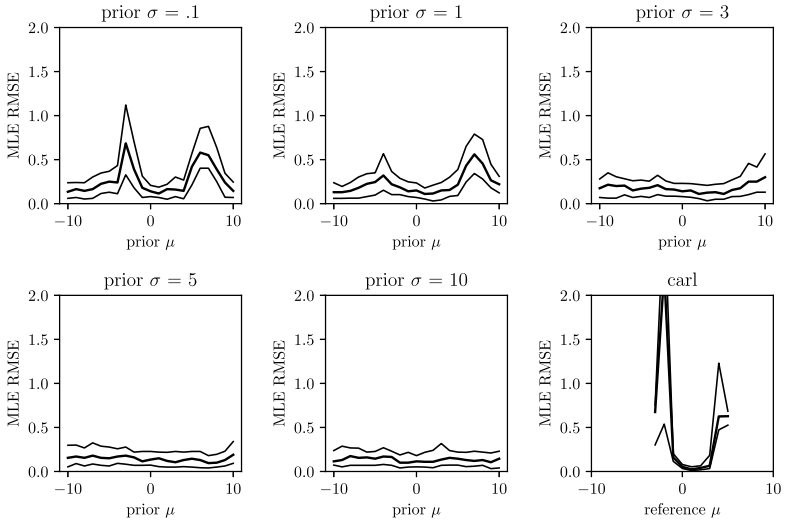


Figure 4: Estimation of the mean of a univariate Gaussian, comparison of the root mean squared error (RMSE) of the approximate MLEs for LFIRE with different priors and carl with different reference points (bottom right).

using the associated software "carl: a likelihood free inference toolbox", with reference parameter values taking integer values from -10 to 10. A multi-layer perceptron global classifier was used for the carl algorithm, trained on 10000 simulation samples and parameter values drawn from the entire parameter space. We found that increasing the number of simulations did not improve performance of the carl algorithm. Both the LFIRE and the carl simulations were repeated 50 times per setup. Approximate MLEs were obtained for both methods: for carl, we used the associated software package, for LFIRE, they were computed by maximising the approximate likelihood in (11).

Figure 4 shows the root mean squared errors (RMSE) of the obtained approximate MLEs, with the medians, 25th and 75th quantiles over the 50 repetitions plotted. The carl method (bottom right) led to small RMSEs when the reference point θ_r is well-chosen. When the reference point is further way from the true parameter value, however, the RMSE becomes larger and when too far away, the carl software failed and returned an uninformative default value. LFIRE with a mis-specified overly confident prior (top left, small standard deviation, prior mean far from the true value) produced large RMSEs. For broader and more reasonable priors, LFIRE yielded small RMSEs for a wide range of prior means, and was fairly robust to the exact choice of the prior.

Figure 5 assesses the performance in posterior density estimation in terms of the symmetrised Kullback-Leibler divergence sKL between the approximate and true posterior. The figure shows that LFIRE produced reasonable approximations unless the prior was overly narrow and mis-specified. The carl method did not provide computationally stable responses for the likelihood and hence posterior for the entire parameter range considered, so it is not included in the figure.

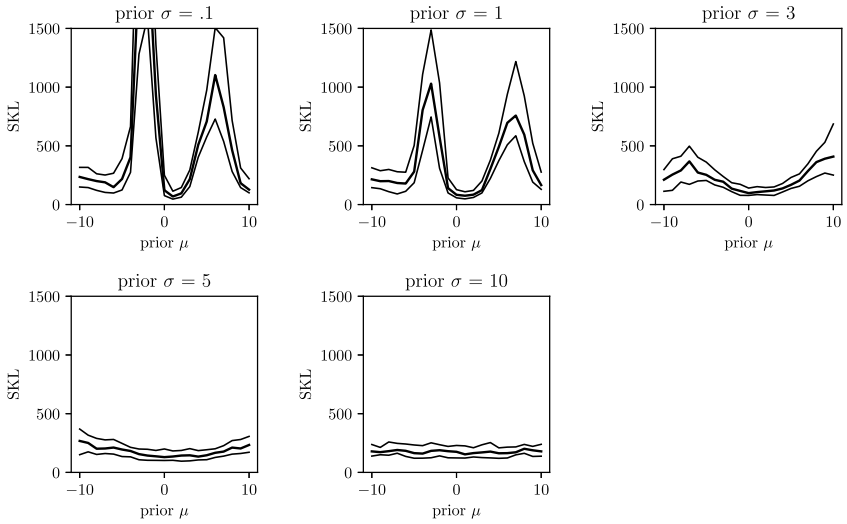


Figure 5: Setup as in Figure 4 but assessing the posterior density estimates in terms of the symmetrised Kullback-Leibler (sKL) distance to the true posteriors. The carl method did not lead to stable results and could not be included in the comparison.

These empirical results are in line with the conceptual considerations above. In particular, the use of the marginal $p(x)$ as a normalising distribution in LFIRE leads to stable estimates across a broad range of prior settings, which would be appropriate for a Bayesian approach. By contrast, using a denominator likelihood conditioned on a specific parameter value was found to be only accurate if the reference value θ_r is close to the (unknown) true parameter, but is unstable across an extended parameter range.

6 Bayesian inference for nonlinear dynamical systems

We here apply linear LFIRE in 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 a 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 out 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),

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

$$g(y_k^{(t)}, \theta) = \theta_1 + \theta_2 y_k^{(t)}, \tag{29}$$

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 (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 as in the previous sections.

6.2 Results

We used an importance sampling scheme (Ripley, 1987, IS) by sampling 10,000 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, see the method section in his paper), for the synthetic likelihood approach we used a robust variance-covariance matrix estimation scheme for a better estimation of the likelihood function. A simple approach is to add some scaled diagonal "jitter" to the covariance matrix to ensure numerical stability when computing the inverse.

Figure 6 shows example results for the Ricker model, and Figure 7 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 thus used the results from expensive rejection ABC runs for reference (threshold set to achieve approximately 2% acceptance). We assessed the results in terms of the accuracy of the posterior mean and posterior standard deviations.

The posterior mean $E_x[\hat{\theta}(x)]$ for linear LFIRE and $E_x[\hat{\theta}_{SL}(x)]$ for the synthetic likelihood approach were computed from the posterior samples. The relative errors of the proposed approach and the synthetic likelihood were computed relative to the ABC

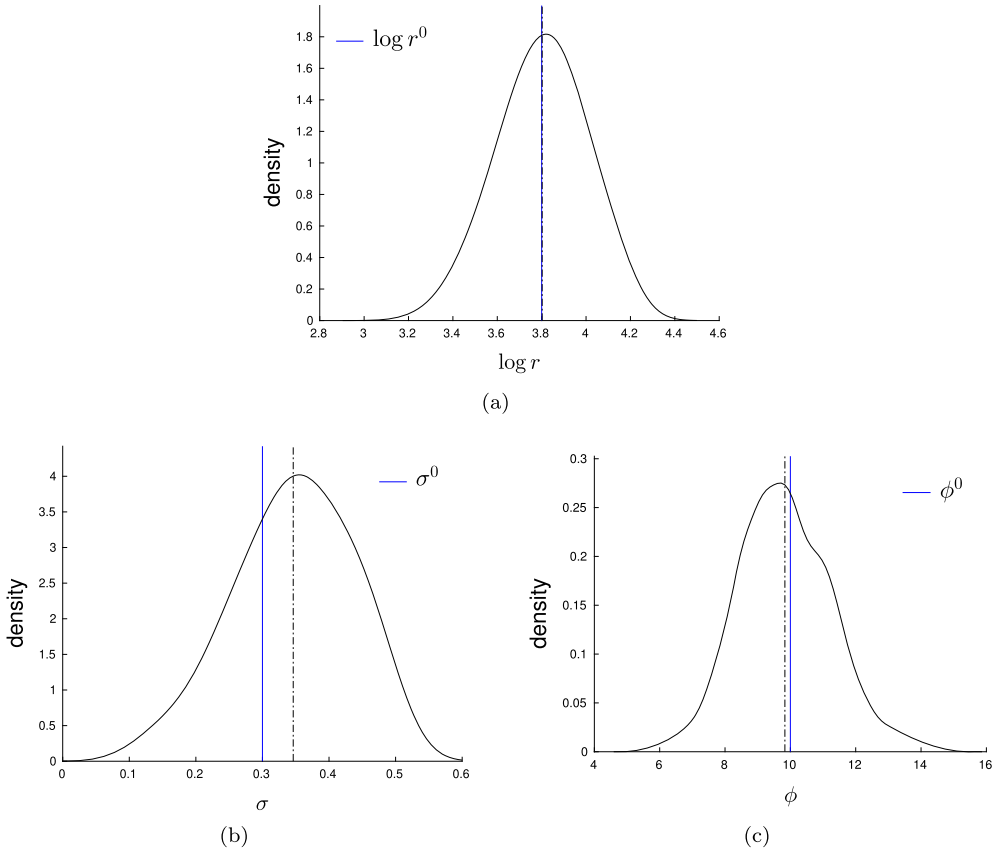


Figure 6: Ricker model: Example marginal posterior distribution of (a) $\log r$, (b) σ and (c) ϕ, estimated with linear LFIRE in Algorithm 1 using $n_\theta = n_m = 100$. The blue vertical lines show the true parameter values $(\log r^0, \sigma^0, \phi^0)$ that we used to simulate the observed data and 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.

results for each element of the parameter vector θ,

$$\mathcal{RE}(x) = \sqrt{\frac{(E_x[\hat{\theta}(x)] - E_x[\hat{\theta}_{ABC}(x)])^2}{E_x[\hat{\theta}_{ABC}(x)]^2}}, \quad \mathcal{RE}_{\mathrm{SL}}(x) = \sqrt{\frac{(E_x[\hat{\theta}_{SL}(x)] - E_x[\hat{\theta}_{ABC}(x)])^2}{E_x[\hat{\theta}_{ABC}(x)]^2}}. \tag{30}$$

The squaring and division should be understood as element-wise operations. As the relative error depends on the observed data, we computed the error for 250 different observed datasets x_0. We performed a point-wise comparison between the proposed method and synthetic likelihood by computing the difference $\Delta_{\mathrm{rel-error}}$ between the

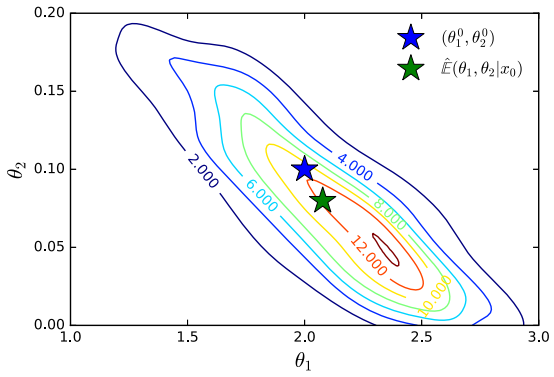


Figure 7: Lorenz model: Example posterior distribution of the closure parameters (θ_1, θ_2) estimated with linear LFIRE in Algorithm 1 using $n_\theta = n_m = 100$. The blue and green asterisk indicate the true parameter values (θ_1^0, θ_2^0) that were used to simulate the observed data and the estimated 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.

relative errors for all elements in the parameter vector θ,

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

Exactly the same procedure was used to assess the accuracy of the standard deviations.

For both the posterior means and standard deviations, 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 the other hand, indicates that the synthetic likelihood is performing better. As $\Delta_{\mathrm{rel-error}}$ is a function of x_0, we report the empirical distribution of $\Delta_{\mathrm{rel-error}}$ computed from the 250 different observed data sets x_0.

Figures 3 to 6 in Supplementary Material D show the empirical distribution of $\Delta_{\mathrm{rel-error}}$ for the posterior means and standard deviations for the Ricker and the Lorenz model. All distributions are tilted toward negative values of $\Delta_{\mathrm{rel-error}}$ for all the parameters, which indicates that the proposed method is generally performing better in both applications. As the proposed and the synthetic likelihood method use exactly the same summary statistics, we did not expect large improvements in the performance. Nevertheless, the figures show that linear LFIRE achieves better accuracy in the posterior mean for all but one parameter where the performance is roughly equal, and better accuracy in the posterior standard deviations in all cases. These results correlate well with the findings for the ARCH model (note e.g. the more accurate characterisation of the posterior uncertainty in Figure 3) and generally highlight the benefits of LFIRE taking non-Gaussian properties of the summary statistics into account.

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 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 (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, $E_x[\hat{\theta}(x)]$ and $E_x[\hat{\theta}_{SL}(x)]$ 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 250 different x_0.

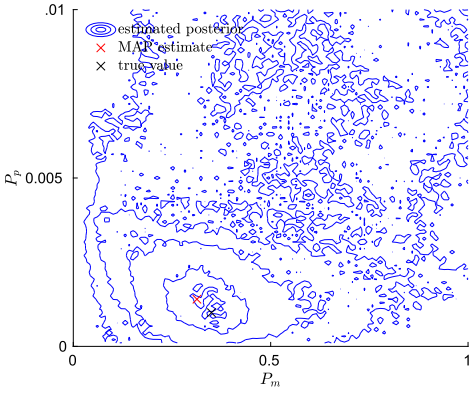
Figure 7 in Supplementary Material D shows the median, the 1/4 and the 3/4 quantile of $\zeta^{(t)}(x_0)$ for $t \in [4, 6]$ corresponding to one to ten days in the future. We achieve on average a clear improvement in prediction performance for the first days; for longer-term forecasts, the improvement becomes smaller, which is due to the inherent difficulty to make long-term predictions for chaotic time series.

7 Inference with high-dimensional summary statistics

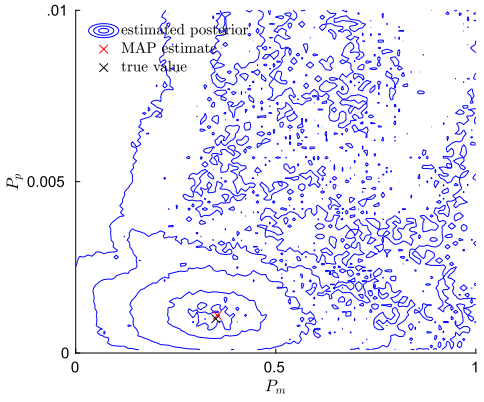
Here we present the results of the LFIRE method applied to the stochastic cell spreading model described in Price et al. (2017). This model is notable for its use of a large number of summary statistics to determine the model parameters describing motility and proliferation, P_m and P_p. The summary statistics are the total number of cells at the end of the experiment and the Hamming distances between the image grids of cell populations evaluated at each time point in the simulation, providing 145 summary statistics. This vector was then combined with its own element-wise square and a constant, resulting in a final total of 291 summary statistics.

The linear LFIRE method using a lasso-type regularisation is well-positioned to perform efficient inference for such a model, as it can select the summary statistics that are most informative for the characterisation of the posterior distribution. We performed inference with true values of $P_m = 0.35$ and $P_p = 0.001$, and varied the amount of simulated data used to train the classifier, with values of $n_\theta = n_m \in \{50, 100, 150\}$. Given the prior knowledge that the P_p would take small values, we asserted uniform priors over each model parameter between $[0, 1]$ and $[0, 0.01]$, respectively.

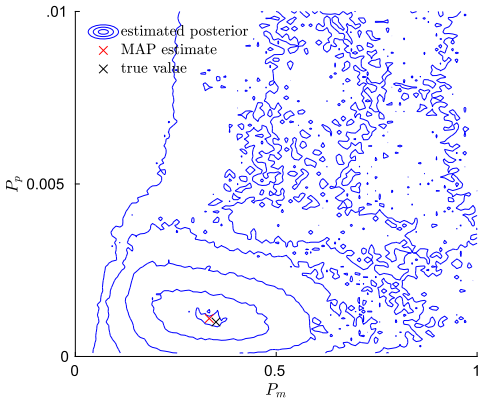
The results are presented in Figure 8. It is seen that the posterior becomes more stably characterised as $n_\theta = n_m$ increases from 50 to 150, with the MAP estimates clearly



(a) $n_\theta = n_m = 50$



(b) $n_\theta = n_m = 100$



(c) $n_\theta = n_m = 150$

Figure 8: Cell spreading model: Contour plots of the approximate posterior computed by linear LFIRE in Algorithm 1 for the parameters P_m and P_p. Each panel corresponds to different numbers of simulated data points $n_\theta = n_m$ to train the classifier. The true values and MAP estimates of the parameters are also displayed in the plots.

improving with the extra training data. The results are in line with those presented by Price et al. (2017) who used synthetic likelihood, but employing a much larger number of simulations (namely $n_\theta \in \{2500, 5000, 10000\}$). We were not able to make synthetic likelihood work with the low numbers of n_θ used in LFIRE but recent work on synthetic likelihood (Ong et al., 2018) shows that shrinkage estimation methods enable values of $n_\theta \in \{500, 1000\}$.

We further compared the number of non-zero summary statistic weights that linear LFIRE uses for each value of $n_\theta = n_m$. As the classifier is exposed to more training data, we would expect it to select more summary statistics as more evidence becomes available. This phenomenon is observed in our simulations, with the $n_\theta = n_m = 50$,

100 and 150 simulations selecting an average of 17.3, 23.9 and 30.5 summary statistics, respectively, from a total number of size 291. This demonstrates that the method is able to both select from a large pool of summary statistics and form a more complex classification model when more computational resources are made available.

8 Computational aspects

In this section we discuss the computational cost of LFIRE and compare it with the existing synthetic likelihood approach.

Both methods require generating the data set X^θ, which most often will dominate the computational cost. Linear LFIRE has the additional cost of constructing the set X^m once, and the cost of performing penalised logistic regression for each θ, including potentially computationally intensive cross-validation to establish the regularisation strength. Synthetic likelihood, on the other hand, requires inversion of the covariance matrix of the summary statistics for each θ. Cross-validation can also be used for regularisation of the synthetic likelihood with a graphical lasso for robust inference in situations with a large or poorly-conditioned covariance matrix (An et al., 2019). When the covariance matrix is well-conditioned, penalised logistic regression with cross-validation is more expensive than standard inversion of the covariance matrix, and the difference in computational cost can be seen as the price that needs to be paid for the relaxation of the Gaussianity assumption, and for feature selection through regularised inference. If, however, simulating data from the model is in the computational bottleneck, the extra cost of regularised logistic regression causes comparably little overhead.

We support these considerations with timing data that were collected for the ARCH and the cell spreading ("scratch") model. Since the absolute computational times are platform-specific, we consider the relative amount of time spent performing simulations and posterior estimation. Simulation times were averaged for robustness over one million simulations, with parameters drawn from the uniform priors used in the experiments described in Sections 5.2 and 7. Similarly, posterior estimation times were averaged over 100 evaluations of posterior proxies, including performing penalised logistic regression through cross-validation, defined over a grid across the uniform prior. The relative balance of computational times between simulations and posterior estimation considered 100 evaluations of the posterior proxies, with n_θ simulated data sets used for each evaluation. Parallel computational resources were not considered in the analysis: they could definitely affect the relative computational times, but with a heavy dependence on the local computing platform and specific inference procedure.

Since both of the ARCH and cell spreading model are computationally inexpensive, we also consider a hypothetical model for which simulations take one second, which is still rather cheap. Likelihood proxy evaluation times were assumed to be 0.09 seconds for the synthetic likelihood and 10 seconds for LFIRE as approximate midpoints of those observed for the other models.

Table 1 presents the proportion of computational time spent performing simulations for different simulator models, posterior approximation methods and values of n_θ. We see

ARCH			Scratch			Hypothetical 1s simulator		
n_θ	LFIRE	SL	n_θ	LFIRE	SL	n_θ	LFIRE	SL
100	0.0295	0.4234	50	0.1623	0.3177	50	0.8347	0.9982
500	0.0171	0.7859	100	0.1738	0.4822	100	0.9099	0.9991
			150	0.1231	0.5828	150	0.9381	0.9994

Table 1: Proportion of total compute time dedicated to simulation.

that for the very cheap ARCH simulations, the LFIRE method spends a majority of its time performing posterior estimation. For the moderately more expensive cell spreading ("scratch") simulations, posterior estimation computations are still dominant, but the relative costs become more balanced. For the hypothetical (but not unrealistic) one second simulator, we see that a comfortable majority of computational time is now spent performing simulations for both linear LFIRE and synthetic likelihood (SL).

In summary, we see that while the LFIRE method requires more time than synthetic likelihood for each posterior estimate, for simulators with non-trivial computational demands the proportion of time spent on generating data is dominant for both methods.

9 Discussion

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 and contrastive learning.

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, as well as in the related work by Pham et al. (2014); Cranmer et al. (2015), 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. As we are guaranteed to have enough data to train the classifier, the main advantage of working with two simulated data sets is that it supports posterior inference given a single observed datum only.

Our approach requires that several samples from the model are generated for the estimation of the posterior, like for synthetic likelihood (Wood, 2010). While the sampling can be performed perfectly in parallel, it constitutes the main computational cost unless the model is very cheap to simulate. There are several ways to reduce the inference cost: First, Bayesian optimisation can be used to intelligently decide where to evaluate the posterior as previously done for the synthetic likelihood, thus reducing unnecessary computations (Gutmann and Corander, 2016; Järvenpää et al., 2018). Second, rather than pointwise estimation, the inference can be amortised with respect to the parameters (Hermans et al., 2020) or one can learn the relation between the parameters

and the log-ratio from already computed parameter-ratio pairs. 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, will likely lead to computational gains. Finally, most computations can be performed offline before the observed data are seen, so that computations can be cached and recycled for newly observed data sets, which reduces the effective cost and enables amortised inference and "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.

A key feature of the proposed method is the automated selection and combination of summary statistics from a large pool of candidates. While there are several works on summary statistics selection in the framework of approximate Bayesian computation (Aeschbacher et al., 2012; Fearnhead and Prangle, 2012; Blum et al., 2013; Gutmann et al., 2014; Marin et al., 2016; Gutmann et al., 2018; Jiang et al., 2018), there is comparably little corresponding work on synthetic likelihood (Wood, 2010) with the exception of the recent work by An et al. (2019) and Ong et al. (2018) whose robust estimation techniques of the (inverse) covariance matrix are broadly related to summary statistics selection. We have shown that synthetic likelihood is a special case of the proposed approach so that our techniques for summary statistics selection could also be used there.

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 multivariate Gaussianity assumption of synthetic likelihood. Finding such a joint transformation of summaries to fulfil the multivariate Gaussianity criterion is generally challenging, as it is very difficult to constrain the resulting multivariate distribution's higher-order moments, e.g. the co-skewness and co-kurtosis, without losing information. However, it is always possible to average across multiple evaluations of summary statistics and use a central limit theorem to asymptotically approach a multivariate Gaussian distribution. This is in contrast to our approach that automatically adapts to non-Gaussianity of the summary statistics.

Our results showed that the proposed method can effectively select relevant summary statistics. We used the method to remove completely irrelevant ones but also to adaptively include more relevant ones when more computational resources become available. Moreover, the ability to automatically select data summaries substantially contributes to the interpretability of the inference procedure, and the selected summary statistics may provide additional insights into the fundamental scientific question at hand. While automated selection of summary statistics from a large pool of candidates alleviates the burden on the user to provide carefully engineered summary statistics it assumes that some of the candidates are suitable in the first place. The intrinsic connection of the proposed approach to classification facilitates the learning of summaries from raw data (Dinev and Gutmann, 2018) thereby partly addressing this point.

We have seen that the proposed approach can well handle high-dimensional summary statistics. The separate problem of likelihood-free inference for high-dimensional parameter spaces is a highly relevant question. This is a very challenging problem without a generally accepted solution: the LFIRE methodology might perform well in this context, because it does not involve the choice of an acceptance threshold or other kernel as in typical ABC, which is often a problem when inferring high-dimensional parameters. However, it was not developed with such problems specifically in mind, and such an investigation does not fall within the scope of this paper.

We have used a linear basis expansion and logistic regression to implement the proposed framework of likelihood-free inference by ratio estimation. While more general regression models and other loss functions such as Bregman divergences (Gutmann and Hirayama, 2011; Sugiyama et al., 2012) can be used, we found that already this simple instance of the framework provided a generalisation of the synthetic likelihood approach with typically more accurate estimation results.

Our findings suggest that likelihood-free inference by ratio estimation is a useful technique, and the proposed rich *framework* opens up several directions to new 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.

Supplementary Material

Likelihood-free inference by ratio estimation —Supplementary Material—
(DOI: 10.1214/20-BA1238SUPP; .pdf). Supplementary Material A contains the proof of Equation (8), Supplementary Material B an analysis of the effect of the summary statistics for the ARCH model, Supplementary Material C an additional table and D additional figures.

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. 2, 24

An, Z., South, L., Nott, D., and Drovandi, C. (2019). "Accelerating Bayesian synthetic likelihood with the graphical lasso." *Journal of Computational and Graphical Statistics*, 28(2): 471–475. MR3974895. doi: https://doi.org/10.1080/10618600.2018.1537928. 9, 22, 24

Andrieu, C. and Roberts, G. O. (2009). "The pseudo-marginal approach for efficient Monte Carlo computations." *Annals of Statistics*, 37(2): 697–725. MR2502648. doi: https://doi.org/10.1214/07-AOS574. 6

Arnold, B., Gutmann, M., Grad, Y., Sheppard, S., Corander, J., Lipsitch, M., and Hanage, W. (2018). "Weak Epistasis May Drive Adaptation in Recombining Bacteria." *Genetics*, 208(3): 1247–1260. 2

Beaumont, M. A. (2010). "Approximate Bayesian Computation in Evolution and Ecology." *Annual Review of Ecology, Evolution, and Systematics*, 41(1): 379–406. MR3939526. doi: https://doi.org/10.1146/annurev-statistics-030718-105212. 2

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). "Approximate Bayesian Computation in Population Genetics." *Genetics*, 162(4): 2025–2035. 2

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, 81–88. New York, NY, USA: ACM. 4

Blum, M. G. B. (2010). "Approximate Bayesian Computation: A Nonparametric Perspective." *Journal of the American Statistical Association*, 105(491): 1178–1187. MR2752613. doi: https://doi.org/10.1198/jasa.2010.tm09448. 2

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. MR3112405. doi: https://doi.org/10.1214/12-sts406. 2, 24

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. MR2109057. doi: https://doi.org/10.1198/106186004X12803. 17

Carnahan, B., Luther, H. A., and Wilkes, J. O. (1969). *Applied Numerical Methods*. New York: Wiley. 16

Chen, Y. and Gutmann, M. U. (2019). "Adaptive Gaussian Copula ABC." In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, 1584–1592. PMLR. 2

Cheng, K. and Chu, C. (2004). "Semiparametric density estimation under a two-sample density ratio model." *Bernoulli*, 10(4): 583–604. MR2076064. doi: https://doi.org/10.3150/bj/1093265631. 4

Corander, J., Fraser, C., Gutmann, M., Arnold, B., Hanage, W., Bentley, S., Lipsitch, M., and Croucher, N. (2017). "Frequency-dependent selection in vaccine-associated pneumococcal population dynamics." *Nature Ecology & Evolution*, 1: 1950–1960. 2

Cranmer, K., Pavez, J., and Louppe, G. (2015). "Approximating Likelihood Ratios with Calibrated Discriminative Classifiers." *ArXiv:1506.02169*. 3, 4, 13, 23

Del Moral, P., Doucet, A., and Jasra, A. (2006). "Sequential Monte Carlo samplers." *Royal Statistical Society: Series B (Statistical Methodology)*, 68(3): 411–436. MR2278333. doi: https://doi.org/10.1111/j.1467-9868.2006.00553.x. 17

Dinev, T. and Gutmann, M. (2018). "Dynamic Likelihood-free Inference via Ratio Estimation (DIRE)." *arXiv:1810.09899*. MR3747571. doi: https://doi.org/10.1007/s11222-017-9738-6. 3, 6, 24

Durkan, C., Murray, I., and Papamakarios, G. (2020). "On Contrastive Learning for

Likelihood-free Inference." In *Proceedings of the thirty-seventh International Conference on Machine Learning (ICML)*. 3

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. MR2975813. 9

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. MR2925370. doi: https://doi.org/10.1111/j.1467-9868.2011.01010.x. 2, 24

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. 8

Geyer, C. J. (1994). "Estimating normalizing constants and reweighting mixtures." Technical Report 568, School of Statistics University of Minnesota. 4

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). "Generative Adversarial Nets." In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, 2672–2680. Curran Associates, Inc. 4

Greenberg, D., Nonnenmacher, M., and Macke, J. (2019). "Automatic Posterior Transformation for Likelihood-Free Inference." In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2404–2414. 4

Gutmann, M., Dutta, R., Kaski, S., and Corander, J. (2014). "Likelihood-free inference via classification." *arXiv:1407.4981*. MR3747571. doi: https://doi.org/10.1007/s11222-017-9738-6. 2, 4, 24

Gutmann, M., Dutta, R., Kaski, S., and Corander, J. (2018). "Likelihood-free inference via classification." *Statistics and Computing*, 28(2): 411–425. MR3747571. doi: https://doi.org/10.1007/s11222-017-9738-6. 2, 4, 10, 24

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)*, 283–290. Corvallis, Oregon: AUAI Press. 4, 25

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. MR3555016. 2, 6, 23

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. MR2913702. 4, 23

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. 17

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. 2

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. MR2722294. doi: https://doi.org/10.1007/978-0-387-84858-7. 5, 7, 9

Hermans, J., Begy, V., and Louppe, G. (2020). "Likelihood-free MCMC with Amortized Approximate Ratio Estimators." In *Proceedings of the thirty-seventh International Conference on Machine Learning (ICML)*. 3, 4, 6, 23

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*, 420–429. 4

Järvenpää, M., Gutmann, M., Vehtari, A., and Marttinen, P. (2018). "Efficient acquisition rules for model-based approximate Bayesian computation." *Bayesian Analysis*, in press. MR3934099. doi: https://doi.org/10.1214/18-BA1121. 23

Jiang, B., Wu, T.-y., Zheng, C., and Wong, W. H. (2018). "Learning Summary Statistic for Approximate Bayesian Computation via Deep Neural Network." *Statistica Sinica*. MR3701500. 24

Leuenberger, C. and Wegmann, D. (2010). "Bayesian Computation and Model Selection Without Likelihoods." *Genetics*, 184(1): 243–252. 7

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. 2

Lorenz, E. (1995). "Predictability: a problem partly solved." In *Proceedings of the Seminar on Predictability, 4-8 September 1995*, volume 1, 1–18. European Center on Medium Range Weather Forecasting, Shinfield Park, Reading: European Center on Medium Range Weather Forecasting. 15, 16

Marin, J.-M., Pudlo, P., Robert, C., and Ryder, R. (2012). "Approximate Bayesian computational methods." *Statistics and Computing*, 22(6): 1167–1180. MR2992292. doi: https://doi.org/10.1007/s11222-011-9288-2. 2

Marin, L., J. M. ans Raynal, Pudlo, P., Ribatet, M., and Robert, C. (2016). "ABC Random Forests for Bayesian Parameter Inference." *arXiv:1605.05537*. 24

Meeds, E. and Welling, M. (2014). "GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation." In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, 593–602. AUAI Press. 2

Mohamed, S. and Lakshminarayanan, B. (2016). "Learning in Implicit Generative Models." *arXiv:1610.03483*. 4

Ong, V., Nott, D., Tran, M.-N., Sisson, S., and Drovandi, C. (2018). "Likelihood-free inference in high dimensions with synthetic likelihood." *Computational Statistics & Data Analysis*, 128: 271–291. MR3850637. doi: https://doi.org/10.1016/j.csda.2018.07.008. 21, 24

Papamakarios, G. and Murray, I. (2016). "Fast epsilon-free Inference of Simulation Models with Bayesian Conditional Density Estimation." In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, 1028–1036. Curran Associates, Inc. 2

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). "Masked autoregressive flow for density estimation." In *Advances in Neural Information Processing Systems*, 2338–2347. 2

Papamakarios, G., Sterratt, D., and Murray, I. (2019). "Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows." In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89, 837–848. Proceedings of Machine Learning Research: PMLR. 2

Pham, K. C., Nott, D. J., and Chaudhuri, S. (2014). "A note on approximating ABC-MCMC using flexible classifiers." *Stat*, 3(1): 218–227. MR4027338. doi: https://doi.org/10.1002/sta4.56. 3, 4, 23

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)*. 4

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2017). "Bayesian Synthetic Likelihood." *Journal of Computational and Graphical Statistics*, 1–11. MR3788296. doi: https://doi.org/10.1080/10618600.2017.1302882. 2, 20, 21

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. 2

Qin, J. (1998). "Inferences for case-control and semiparametric two-sample density ratio models." *Biometrika*, 85(3): 619–630. MR1665814. doi: https://doi.org/10.1093/biomet/85.3.619. 4

Ricker, W. E. (1954). "Stock and Recruitment." *Journal of the Fisheries Research Board of Canada*, 11(5): 559–623. 15, 16

Ripley, B. D. (1987). *Stochastic simulation*. New York, USA: John Wiley & Sons Inc. MR0875224. doi: https://doi.org/10.1002/9780470316726. 17

Rogers-Smith, C., Pesonen, H., and Kaski, S. (2018). "Approximate Bayesian Computation via Population Monte Carlo and Classification." *arXiv:1810.12233*. 3

Sirén, J., Lens, L., Cousseau, L., and Ovaskainen, O. (2018). "Assessing the dynamics of natural populations by fitting individual-based models with approximate Bayesian computation." *Methods Ecol Evol*, 9(5): 1286–1295. 2

Sisson, S., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation.*, chapter Overview of Approximate Bayesian Computation. Chapman and Hall/CRC Press. MR3889278. 2

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. New York, NY, USA: Cambridge University Press, 1st edition. MR2895762. doi: https://doi.org/10.1017/CBO9781139035613. 4, 25

Tanaka, M. M., Francis, A. R., Luciani, F., and Sisson, S. A. (2006). "Using Approximate Bayesian Computation to Estimate Tuberculosis Transmission Parameters From Genotype Data." *Genetics*, 173(3): 1511–1520. 2

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). "Inferring Coalescence Times From DNA Sequence Data." *Genetics*, 145(2): 505–518. 2

Thomas, O., Dutta, R., Corander, J., Kaski, S., and Gutmann, M. (2016). "Likelihood-Free Inference by Ratio Estimation." *arXiv:1611.10242*. 2, 3

Thomas, O., Dutta, R., Corander, J., Kaski, S., and Gutmann, M. (2020). "Supplementary Material of "Likelihood-Free Inference by Ratio Estimation"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1238SUPP. 6

Tibshirani, R. (1994). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 58: 267–288. MR1379242. 8, 9

Tibshirani, R. J. and Taylor, J. (2012). "Degrees of freedom in lasso problems." *Annals of Statistics*, 40(2): 1198–1232. MR2985948. doi: https://doi.org/10.1214/12-AOS1003. 9

Wang, H. and Leng, C. (2007). "Unified LASSO Estimation by Least Squares Approximation." *Journal of the American Statistical Association*, 102(479): 1039–1048. MR2411663. doi: https://doi.org/10.1198/016214507000000509. 9

Wilkinson, R. D. (2014). "Accelerating ABC methods using Gaussian processes." In *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33. 2, 6

Wilks, D. S. (2005). "Effects of stochastic parametrizations in the Lorenz '96 system." *Quarterly Journal of the Royal Meteorological Society*, 131(606): 389–407. 15, 16

Wood, S. N. (2010). "Statistical inference for noisy nonlinear ecological dynamic systems." *Nature*, 466(7310): 1102–1104. 2, 7, 10, 15, 16, 17, 23, 24

Zou, H., Hastie, T., and Tibshirani, R. (2007). "On the "degrees of freedom" of the lasso." *Annals of Statistics*, 35(5): 2173–2192. MR2363967. doi: https://doi.org/10.1214/009053607000000127. 9

Acknowledgments

The work was partially done when OT, RD and MUG were at the Department of Biostatistics, University of Oslo, 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 on an earlier version of the paper and gratefully acknowledge the computational resources provided by the Aalto Science-IT project. RD was supported by Swiss National Science Foundation grant no. 105218_163196. JC and OT were supported by ERC grant no. 742158.

(September 2020)

Likelihood-free inference by ratio estimation —Supplementary Material—

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

$$\mathcal{J}(h) = \frac{1}{n+m}\left\{\sum_{i=1}^{n}\log\left[1+\nu\exp(-h(x_i))\right] + \sum_{i=1}^{m}\log\left[1+\frac{1}{\nu}\exp(h(y_i))\right]\right\}. \tag{S1}$$

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

$$\begin{aligned}
\mathcal{J}(h) &= \frac{n}{n+m}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\nu\exp(-h(x_i))\right] + \frac{1}{n}\sum_{i=1}^{m}\log\left[1+\frac{1}{\nu}\exp(h(y_i))\right]\right\} &\text{(S2)}\\
&= \frac{n}{n+m}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\nu\exp(-h(x_i))\right] + \nu\frac{1}{m}\sum_{i=1}^{m}\log\left[1+\frac{1}{\nu}\exp(h(y_i))\right]\right\} &\text{(S3)}\\
&= \frac{1}{1+\nu}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\nu\exp(-h(x_i))\right] + \nu\frac{1}{m}\sum_{i=1}^{m}\log\left[1+\frac{1}{\nu}\exp(h(y_i))\right]\right\} &\text{(S4)}
\end{aligned}$$

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

$$\tilde{\mathcal{J}}(h) = \mathbb{E}_x \log\left[1+\nu\exp(-h(x))\right] + \nu\mathbb{E}_y\log\left[1+\frac{1}{\nu}\exp(h(y))\right]. \tag{S5}$$

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

$$\log\left(1+\frac{1}{\nu}\exp h\right) = \log(\nu\exp(-h)+1) - \log(\nu\exp(-h)) \tag{S6}$$

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

$$\begin{aligned}
\tilde{\mathcal{J}}(h) =& \mathbb{E}_x\log(1+\nu\exp(-h(x))) + \nu\mathbb{E}_y\log(\nu\exp(-h(y))+1)\\
&- \nu\mathbb{E}_y\log(\nu\exp(-h(y))) &\text{(S7)}\\
=& \mathbb{E}_x\log(1+\nu\exp(-h(x))) + \nu\mathbb{E}_y\log\left(1+\nu\exp(-h(y))\right)\\
&- \nu\log\nu + \nu\mathbb{E}_y h(y) &\text{(S8)}
\end{aligned}$$

$$= \int p_x(u) \log(1 + \nu \exp(-h(u)))du + \nu \int p_y(u) \log(1 + \nu \exp(-h(u)))du$$
$$- \nu \log \nu + \nu \int p_y(u)h(u)du \tag{S9}$$
$$= \int (p_x(u) + \nu p_y(u)) \log(1 + \nu \exp(-h(u)))du$$
$$- \nu \log \nu + \nu \int p_y(u)h(u)du. \tag{S10}$$

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

$$\begin{aligned}\log(1 + \nu \exp(-h(u) - \epsilon q(u))) = &\log(1 + \nu \exp(-h(u)) \\ &- \epsilon q(u) \frac{\nu \exp(-h(u))}{1 + \nu \exp(-h(u))} \\ &+ \frac{\epsilon^2 q(u)^2}{2} \frac{\nu \exp(-h(u))}{1 + \nu \exp(-h(u))} \frac{1}{1 + \nu \exp(-h(u))} \\ &+ O(\epsilon^3)\end{aligned} \tag{S11}$$

we have

$$\begin{aligned}\tilde{\mathcal{J}}(h + \epsilon q) = &\int (p_x(u) + \nu p_y(u)) \log(1 + \nu \exp(-h(u))du \\ &- \int (p_x(u) + \nu p_y(u)) \epsilon q(u) \frac{\nu \exp(-h(u))}{1 + \nu \exp(-h(u))} du \\ &+ \int (p_x(u) + \nu p_y(u)) \frac{\epsilon^2 q(u)^2}{2} \frac{\nu \exp(-h(u))}{1 + \nu \exp(-h(u))} \frac{1}{1 + \nu \exp(-h(u))} du \\ &- \nu \log \nu + \nu \int p_y(u)h(u)du + \nu \int p_y(u)\epsilon q(u)du + O(\epsilon^3).\end{aligned} \tag{S12}$$

Collecting terms gives

$$\begin{aligned}\tilde{\mathcal{J}}(h + \epsilon q) = &\tilde{\mathcal{J}}(h) - \epsilon \int q(u) \left((p_x(u) + \nu p_y(u)) \frac{\nu \exp(-h(u))}{1 + \nu \exp(-h(u))} - \nu p_y(u) \right) du \\ &+ \frac{\epsilon^2}{2} \int q(u)^2 (p_x(u) + \nu p_y(u)) \frac{\nu \exp(-h(u))}{1 + \nu \exp(-h(u))} \frac{1}{1 + \nu \exp(-h(u))} du \\ &+ O(\epsilon^3).\end{aligned} \tag{S13}$$

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

$$\nu p_y(u) = \frac{p_x(u) + \nu p_y(u)}{1 + \frac{1}{\nu} \exp(h^*(u))} \quad \Leftrightarrow \quad \nu p_y(u) + p_y(u) \exp(h^*(u)) = p_x(u) + \nu p_y(u) \tag{S14}$$

that is, if and only if

$$\exp(h^*(u)) = \frac{p_x(u)}{p_y(u)}, \tag{S15}$$

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. Note that the same ratio is obtained for $p_x(u) = p(x|\theta)f(\theta)$ and $p_y(u) = p(x)f(\theta)$ because $f(\theta)$ cancels out. Here, $f(\theta)$ can be any density with support on the parameter space where we want to evaluate the ratio or posterior, and the classification would be performed in the joint θ, x space.

B ARCH model: effect of summary statistics

Unless the summary statistics are sufficient, the posteriors conditioned on the observed data and the posteriors conditioned on the observed summary statistics are different. In the main text, we performed an overall comparison between the approximate and exact posteriors. This is valuable because it measures what we ultimately care about. But it confounds the effect of the summary statistics and the effect of the ratio estimation approach. In order to separate the two effects, we here present an additional comparison using a "gold-standard" rejection ABC algorithm with a small ($2.94 \cdot 10^{-4}$) rejection threshold, drawing 1000 samples for each of the 100 simulated data sets, which provide samples from the posterior conditioned on the summary statistics. We then directly compared the posterior means and standard deviations of the ABC, linear LFIRE, and the exact posteriors.

Averaged over observed data sets, the ABC algorithm yielded posterior means of $\theta_{ABC} = [0.2723, 0.6345]$, and average posterior standard deviations took values of 0.1728 and 0.1783 for each parameter. Performing quadrature over the grid of parameter values for the LFIRE simulations, then averaging over all 100 observed data sets gave mean estimates of $[0.3038, 0.6159]$ and standard deviation estimates of $[0.1494, 0.1928]$, and a similar quadrature approach for the true posterior conditioned on the whole data set gave values of $[0.2924, 0.6779]$ with average standard deviations of $[0.0921, 0.1510]$. We observe that the posterior means of all these distributions are similar, but the standard deviations of θ_1 differ by a factor of approximately two between the ABC samples and the true posterior, in line with the broader posterior reported in the main text. The summary statistics thus broaden the posterior, which also means, because the posteriors integrate to one, that the estimated ratio $\hat{r}(x,\theta)$ is typically smaller than the true ratio.

C Supplementary Table

Method	$n_s = 100$	$n_s = 500$	$n_s = 1000$
Synthetic likelihood	1.82	1.80	2.25
Linear LFIRE	2.04	1.57	1.48
Linear LFIRE with irrelevant summaries	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.

D Supplementary Figures

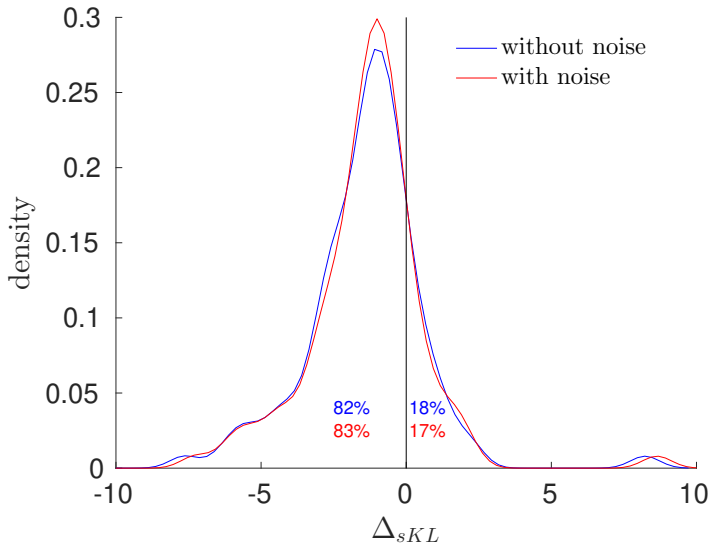


Figure 1: ARCH(1): Estimated density of the difference Δ_{sKL} between the symmetrised Kullback-Leibler divergence for linear LFIRE and synthetic likelihood with $n_\theta = n_m = 1000$, averaged over 100 simulated data sets. A negative value of Δ_{sKL} indicates that the proposed LFIRE method has a smaller divergence and thus is performing better. Depending on whether irrelevant summary statistics are absent (blue) or present (red) in the proposed method, it performs better than synthetic likelihood for 82% or 83% of the simulations. These results correspond to p-values from a Wilcoxon signed-rank test for pairwise median comparison of $1.06 \cdot 10^{-11}$ and $1.42 \cdot 10^{-11}$, respectively. The densities were estimated with a Gaussian kernel density estimator with bandwidth 0.5.

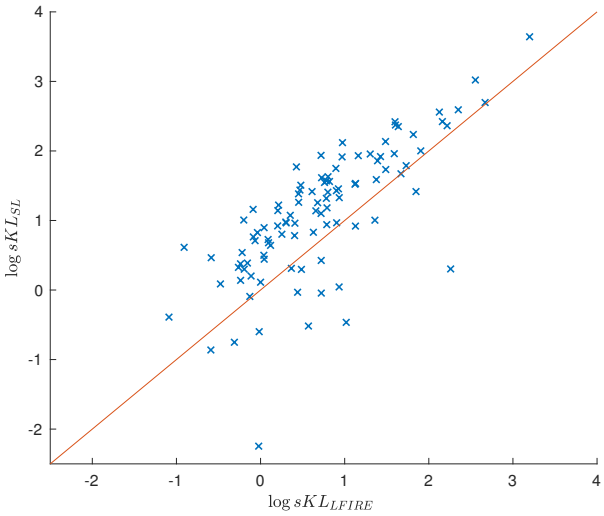
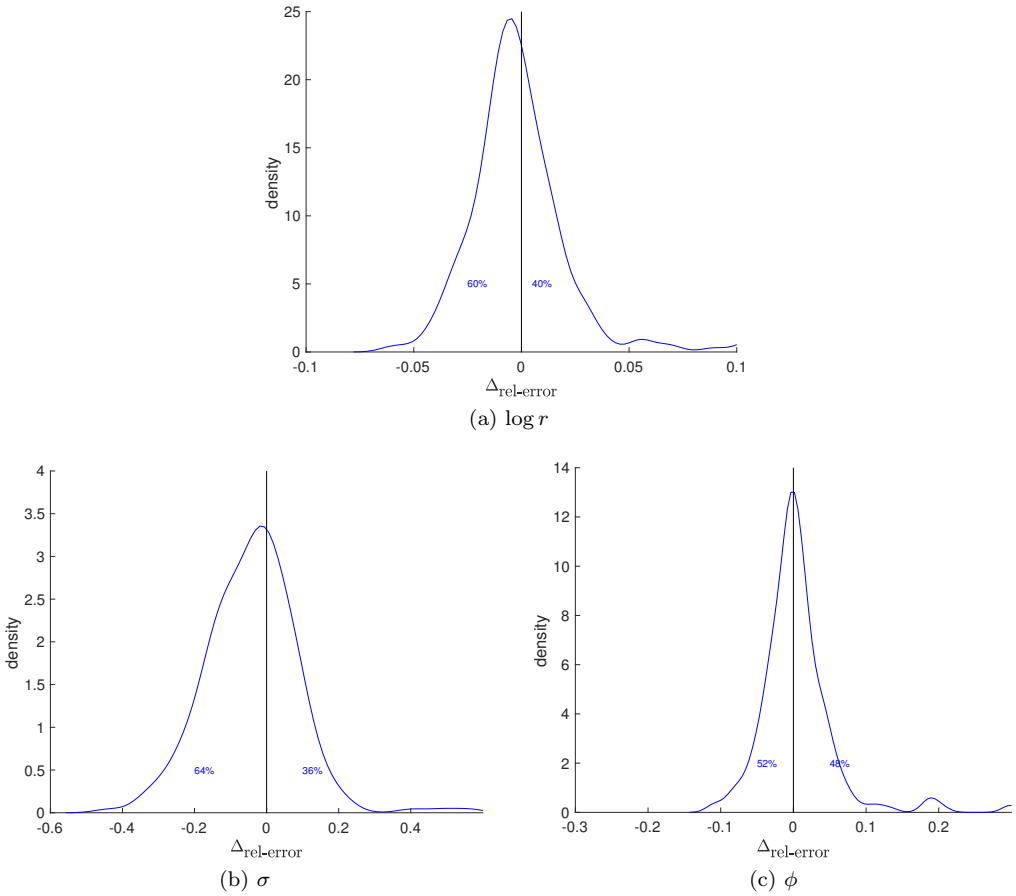
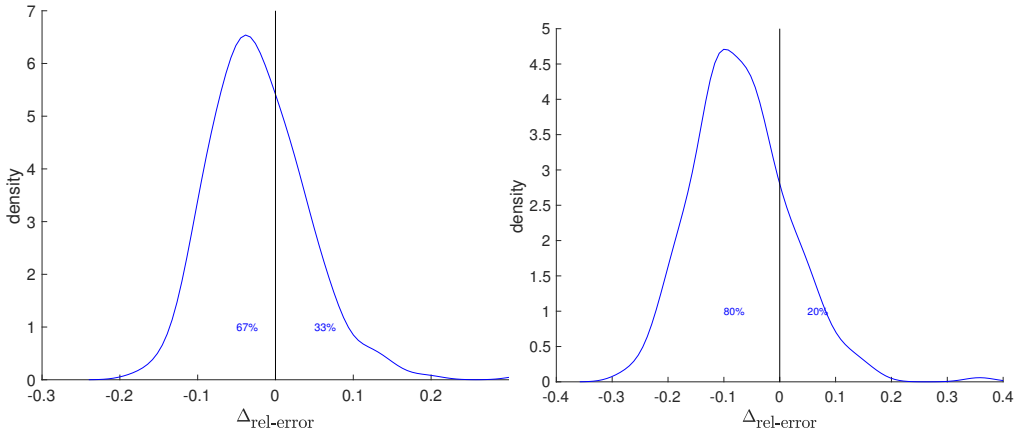


Figure 2: ARCH(1): A scatter plot of the logarithm of the symmetrised Kullback-Leibler divergence (sKL) of the proposed method and synthetic likelihood for $n_\theta = n_m = 1000$, evaluated over 100 simulated data sets. The red line represents hypothetical equal performance of the two methods: we see that a substantial majority of simulations fall above this line, indicating better performance of the LFIRE method.



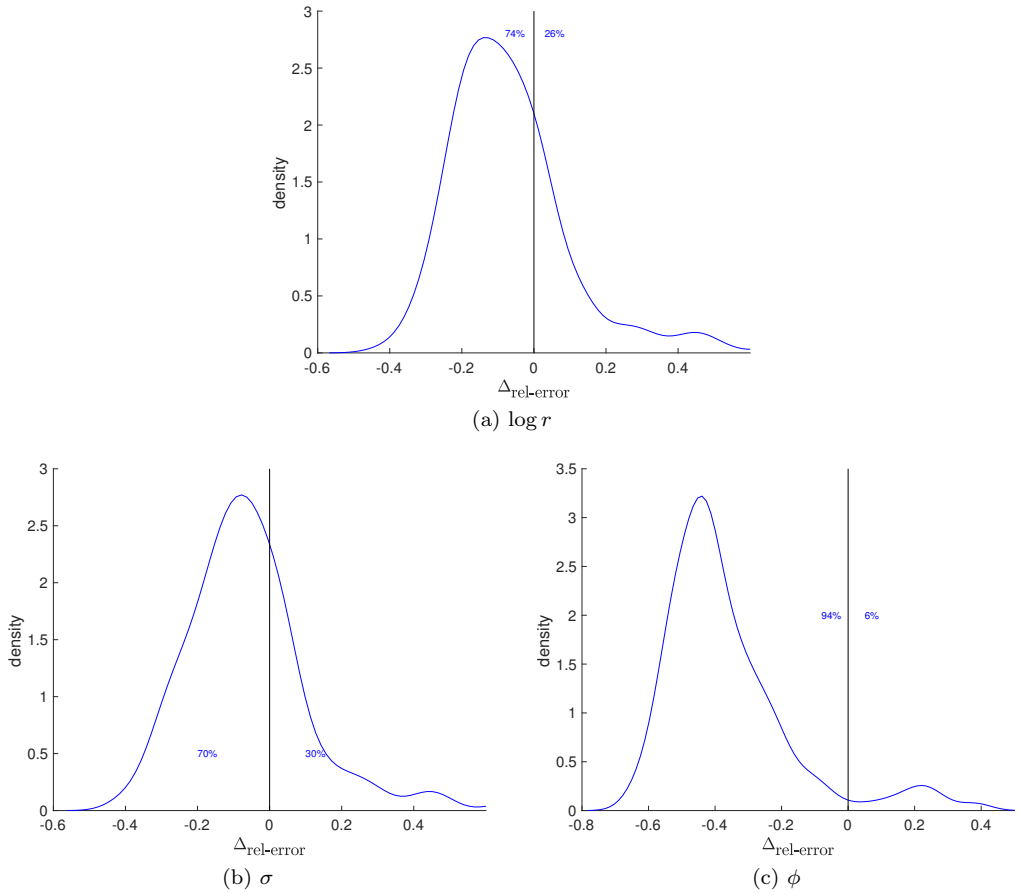
(a) $\log r$ (b) σ (c) ϕ

Figure 3: Ricker model: Empirical pdf of $\Delta_{\text{rel}-\text{error}}$ for the posterior mean of the parameters (a) $\log r$, (b) σ and (c) ϕ, compared against the mean from a rejection ABC algorithm, which drew $10,000$ samples at an acceptance rate of 0.02. 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 linear LFIRE in Algorithm 1 and synthetic likelihood with $n_\theta = n_m = 100$ to estimate the posterior pdf for 250 simulated observed data sets. The densities in (a–c) were estimated using a Gaussian kernel density estimator with bandwidth 0.01, 0.07 and 0.02, respectively. Using a nonparametric Wilcoxon signed-rank test for pairwise median comparison, these results correspond to p-values of 0.0074, $4.99 \cdot 10^{-8}$ and 0.7748, respectively. The plots thus show that linear LFIRE is more accurate than synthetic likelihood in estimating the posterior mean of $\log r$ and σ while the performance is similar for ϕ.



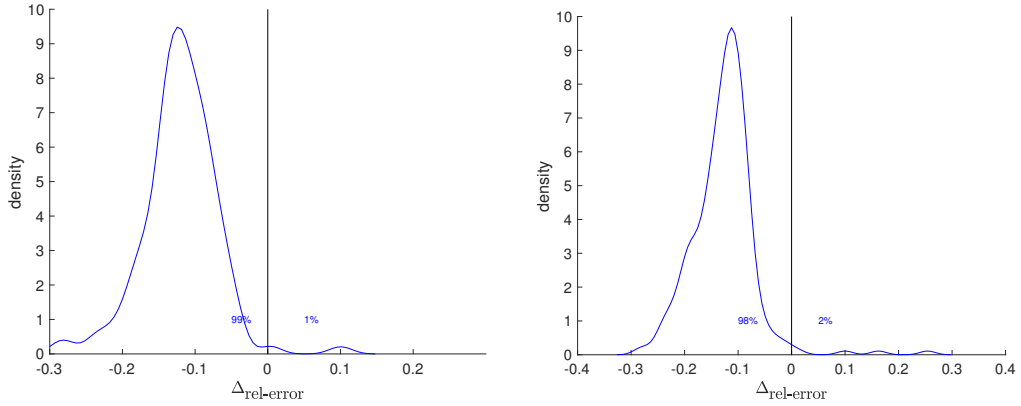
(a) θ_1 (b) θ_2

Figure 4: Lorenz model: Empirical pdf of $\Delta_{\text{rel}-\text{error}}$ for the posterior mean of the parameters (a) θ_1 and (b) θ_2, compared against the mean from a rejection ABC algorithm, which drew $48,000$ samples at an acceptance rate of 0.016. More area under the curve on the negative side of the x-axis indicates a better performance of the proposed method. We used linear LFIRE in Algorithm 1 and synthetic likelihood with $n_\theta = n_m = 100$ to estimate the posterior pdf for 250 simulated observed data sets. The densities in (a–b) were estimated using a Gaussian kernel density estimator with bandwidth 0.025 and 0.037, respectively. Using a nonparametric Wilcoxon signed-rank test for pairwise median comparison, these results correspond to p-values of $4.59 \cdot 10^{-9}$ and $6.92 \cdot 10^{-25}$, respectively. The plots thus show that linear LFIRE is more accurate than synthetic likelihood in estimating the posterior mean of the parameters of the Lorenz model.



(a) $\log r$ (b) σ (c) ϕ

Figure 5: Ricker model: Empirical pdf of $\Delta_{\text{rel}-\text{error}}$ for the posterior standard deviations of the parameters (a) $\log r$, (b) σ and (c) ϕ. Setup is as in Figure 3. Using a nonparametric Wilcoxon signed-rank test for pairwise median comparison, these results correspond to p-values of $2.48 \cdot 10^{-14}, 1.76 \cdot 10^{-11}, 5.98 \cdot 10^{-40}$, respectively. The plots thus show that linear LFIRE is more accurate than synthetic likelihood in estimating the posterior standard deviation (uncertainty) of the parameters of the Ricker model.



(a) θ_1 (b) θ_2

Figure 6: Lorenz model: Empirical pdf of $\Delta_{\text{rel}-\text{error}}$ for the posterior standard deviations of the parameters (a) θ_1 and (b) θ_2. Setup is as in Figure 4. Using a nonparametric Wilcoxon signed-rank test for pairwise median comparison, these results correspond to p-values of $6.29 \cdot 10^{-42}$ and $3.8 \cdot 10^{-40}$, respectively. The plots thus show that linear LFIRE is more accurate than synthetic likelihood in estimating the posterior standard deviation (uncertainty) of the parameters of the Lorenz model.

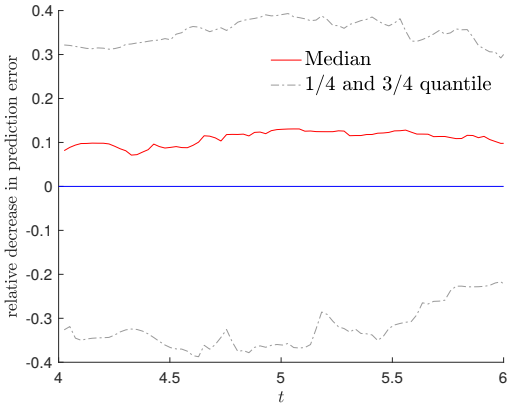


Figure 7: Lorenz Model: Median, 1/4 and 3/4 quantile of the relative decrease in the prediction error $\zeta^{(t)}$ for $t \in [4, 6]$ corresponding to 1 to 10 days in the future. We used linear LFIRE in Algorithm 1 and synthetic likelihood with $n_\theta = n_m = 100$ to estimate the posterior pdf. As the median is always positive, the proposed method obtains, on average, a smaller prediction error than synthetic likelihood.