ABC-CDE: Towards Approximate Bayesian Computation with Complex High-Dimensional Data and Limited Simulations

Rafael Izbicki[*], Ann B. Lee[†] and Taylor Pospisil[‡]

Abstract

Approximate Bayesian Computation (ABC) is typically used when the likelihood is either unavailable or intractable but where data can be simulated under different parameter settings using a forward model. Despite the recent interest in ABC, high-dimensional data and costly simulations still remain a bottleneck. There is also no consensus as to how to best assess the performance of such methods without knowing the true posterior. We show how a nonparametric conditional density estimation (CDE) framework, which we refer to as ABC-CDE, help address three key challenges in ABC: (i) how to efficiently estimate the posterior distribution with limited simulations and different types of data, (ii) how to tune and compare the performance of ABC and related methods in estimating the posterior itself, rather than just certain properties of the density, and (iii) how to efficiently choose among a large set of summary statistics based on a CDE surrogate loss. We provide theoretical and empirical evidence that justify ABC-CDE procedures that *directly* estimate and assess the posterior based on an initial ABC sample, and we describe settings where standard ABC and regression-based approaches are inadequate.

[*] Department of Statistics, Federal University of São Carlos, Brazil.

[†] Department of Statistics, Carnegie Mellon University, USA.

[‡] Department of Statistics, Carnegie Mellon University, USA.

Key Words: nonparametric methods, conditional density estimation, approximate Bayesian computation, likelihood-free inference

1 Introduction

For many statistical inference problems in the sciences the relationship between the parameters of interest and observable data is complicated, but it is possible to simulate realistic data according to some model; see Beaumont (2010); Estoup et al. (2012) for examples in genetics, and Cameron and Pettitt (2012); Weyant et al. (2013) for examples in astronomy. In such situations, the complexity of the data generation process often prevents the derivation of a sufficiently accurate analytical form for the likelihood function. One cannot use standard Bayesian tools as no analytical form for the posterior distribution is available. Nevertheless one can estimate $f(\theta|\mathbf{x})$, the posterior distribution of the parameters $\theta \in \Theta$ given data $\mathbf{x} \in \mathcal{X}$, by taking advantage of the fact that it is possible to forward simulate data $\mathbf{x}$ under different settings of the parameters θ. Problems of this type have motivated recent interest in methods of *likelihood-free inference*, which includes methods of *Approximate Bayesian Computation* (ABC; Marin et al. 2012)

Despite the recent surge of approximate Bayesian methods, several key challenges still remain. In this work, we present a *conditional density estimation* (CDE) framework and *a surrogate loss function for CDE* that address the following three problems:

(i) how to efficiently estimate the posterior density $f(\theta|\mathbf{x}_o)$, where $\mathbf{x}_o$ is the observed sample; in particular, in settings with complex, high-dimensional data and costly simulations,

(ii) how to choose tuning parameters and compare the performance of ABC and related methods based on simulations and observed data only; that is, without knowing the true posterior distribution, and

(iii) how to best choose summary statistics for ABC and related methods when given a very large number of candidate summary statistics.

Existing Methodology. There is an extensive literature on ABC methods; we refer the reader to Marin et al. (2012); Prangle et al. (2014) and references therein for a review. The connection between ABC and CDE has been noted by others; in fact, ABC itself can be viewed as a hybrid between nearest neighbors and kernel density estimators (Blum, 2010; Biau et al., 2015). As Biau et al. point out, the fundamental problem from a practical perspective is how to select the parameters in ABC methods in the absence of a priori information regarding the posterior $f(\theta|\mathbf{x}_o)$. Nearest neighbors and kernel density estimators are also known to perform poorly in settings with a large amount of summary statistics (Blum, 2010), and they are difficult to adapt to different data types (e.g., mixed discrete-continuous statistics and functional data). Few works attempt to use other CDE methods to estimate posterior distributions. To the best of our knowledge, the only works in this direction are Papamakarios and Murray (2016), which is based on conditional neural density estimation, Fan et al. (2013) and Li et al. (2015), which use a mixture of Gaussian copulas to estimate the likelihood function, and more recently, Raynal et al. (2017), which suggests random forests for quantile estimation.

Although the above mentioned methods utilize specific CDE models to estimate posterior distributions, they do not fully explore other advantages of a CDE framework; such as, in methods assessment, in variable selection, and in improving the final estimates with CDE as a goal (see Sections 2.2, 2.3 and 4). Summary statistics selection is indeed a key problem in likelihood-free inference: ABC methods heavily depend on the choice of statistics when comparing observed and simulated data, and using the "wrong" statistics can dramatically affect their performance. For a general review of dimension reduction methods for ABC, we refer the reader to Blum et al. (2013), who classify current approaches in three classes: (a) best subset selection approaches, (b) projection techniques, and (c) regularization techniques. Many of these approaches still face either serious computational issues or attempt to find good summary statistics for certain characteristics of the posterior rather than the *entire* posterior itself. For instance, Creel and Kristensen (2016) propose a best subset selection of summary statistics based on improving the estimate of the posterior mean $\mathbb{E}[\theta|\mathbf{x}]$. There are no guarantees however that statistics that lead to good estimates of $\mathbb{E}[\theta|\mathbf{x}]$ will also yield reasonable estimates of $f(\theta|\mathbf{x})$. As an example, if $X_1, \ldots, X_n \sim \text{Unif}(\theta, \theta+1)$, it is well known that the minimal sufficient summary statistic for θ is $(\min\{X_1, \ldots, X_n\}, \max\{X_1, \ldots, X_n\})$. The optimal statistic for estimating the posterior mean, on the other hand, is $E = \mathbb{E}[\theta|\mathbf{x}]$ (Fearnhead and Prangle, 2012; Section 2.3), which is not sufficient for θ. As a result, ABC will not converge to the true posterior distribution with such a choice of summary statistic. We will elaborate on this point in Section 2.2.

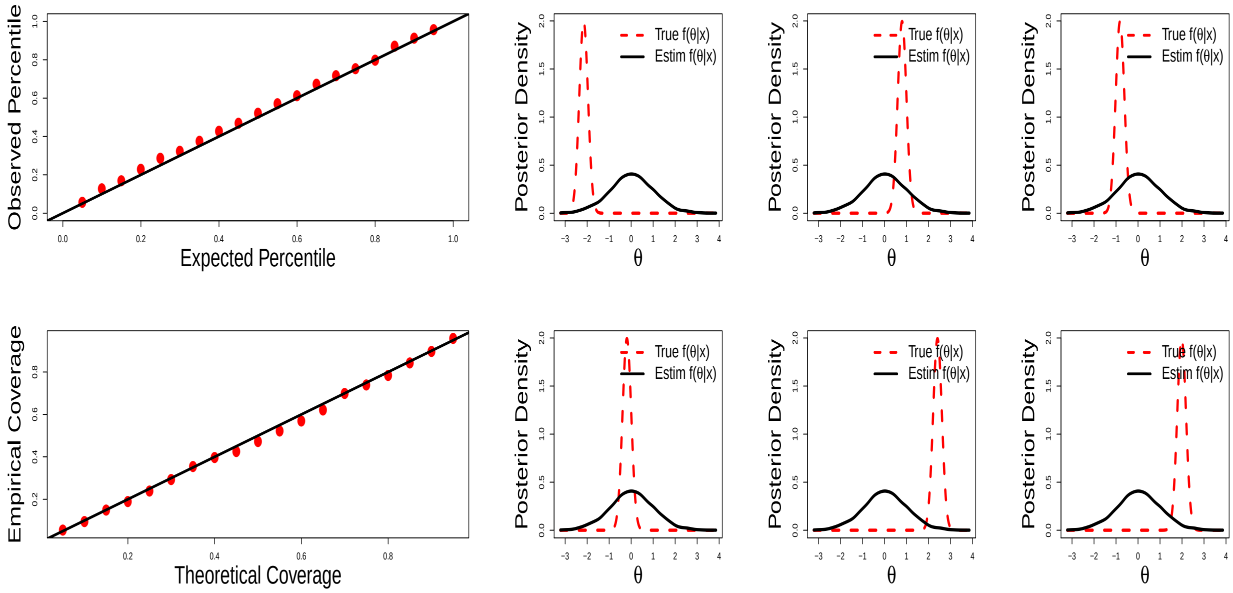


Figure 1: Limitations of diagnostic tests in conditional density estimation. The PP and coverage plots to the left indicate an excellent fit of $f(\theta|x)$ but, as indicated by the examples to the right of a few different values of x, the estimated posterior densities (solid black lines) are very far from the true densities (red dashed lines).

Moreover, the current literature on likelihood-free inference lacks methods that allow one to directly compare the performance of different posterior distribution estimators. Given a collection of estimates $\widehat{f}_1(\theta|\mathbf{x}_o), \ldots, \widehat{f}_m(\theta|\mathbf{x}_o)$ (obtained by, e.g., ABC methods with different tolerance levels, sampling techniques, and so on), an open problem is to how to select the estimate that is closest to the true posterior density $f(\theta|\mathbf{x}_o)$ for observed data $\mathbf{x}_o$. Some goodness-of-fit techniques have been proposed (for example, Prangle et al. 2014 compute the goodness of fit based on coverage properties), but although diagnostic tests are useful, they do not capture all aspects of the density estimates. Some density estimates which are not close to the true density can pass all tests (Breiman, 2001; Bickel et al., 2006), and the situation is even worse in conditional density estimation. Figure 1 shows a simple example where both probability-probability (PP) and coverage plots indicate an excellent fit, but the estimated posterior distributions are far from the true densities; here, $\theta|x \sim \text{Normal}(x, 0.3^2)$ and $X \sim \text{Normal}(0, 1)$. Indeed, standard diagnostic tests will not detect an obvious flaw in conditional density estimates $\widehat{f}(\theta|\mathbf{x})$ that, as in this example, are equal to the marginal distribution $f(\theta) = \int f(\theta|\mathbf{x}')f(\mathbf{x}')d\mathbf{x}'$, no matter what the data $\mathbf{x}$ are.

In this paper, we show how one can improve ABC methods with a novel CDE surrogate loss function (Eq. 3) that measures how well one estimates the entire posterior distribution $f(\theta|\mathbf{x}_o)$; see Section 2.2 for a discussion of its theoretical properties. Our proposed method, ABC-CDE, starts with a rough approximation from an ABC sampler and then *directly* estimates the conditional density exactly at the point $\mathbf{x} = \mathbf{x}_o$ using a nonparametric conditional density estimator. Unlike other ABC post-adjustment techniques in the literature (e.g. Beaumont et al. (2002) and Blum and François (2010)), our method is optimized for estimating posteriors, and corrects for changes in the ABC posterior sample beyond the posterior mean and variance. We also present a general framework (based on CDE) that can handle different types of data (including functional data, mixed variables, structured data, and so on) as well as a larger number of summary statistics. With, for example, FlexCode (Izbicki and Lee, 2017), one can convert any existing regression estimator to a conditional density estimator. Hence, we take a different approach to address the curse of dimensionality than in standard ABC methods. In ABC, it is essential to choose (a smaller set of) informative summary statistics to properly measure distances between observed and simulated data. The main dimension reduction in our framework is implicit in the conditional density estimation and our CDE loss function. Depending of the choice of estimator, we can adapt to different types of sparse structure in the data, and just as in high-dimensional regression, handle a large amount of covariates.

Finally, we note that ABC summary statistic selection and goodness-of-fit techniques are typically designed to estimate posterior distributions accurately for every sample $\mathbf{x}$. In reality, we often only care about estimates for the particular sample $\mathbf{x}_o$ that is observed, and even if a method produces poor estimates for some $f(\theta|\mathbf{x}')$ it can still produce good estimates for $f(\theta|\mathbf{x}_o)$. The methods we introduce in this paper take this into consideration, and directly aim at constructing, evaluating and tuning estimators for the posterior density $f(\theta|\mathbf{x}_o)$ at the *observed* value $\mathbf{x}_o$.

The organization of the paper is as follows: Section 2 describes and presents theoretical results for how a CDE framework and a surrogate loss function address issues (i)–(iii). Section 3 includes simulated experiments and examples from astronomy that demonstrate that our proposed methods work in practice. In Section 4, we revisit CDE in the context of ABC, and compare direct estimation of posteriors with CDE with existing ABC regression adjustment methods. We then end in Section 5 by discussing the significance of our work for next-generation simulations and data.

2 Methods

In this section we propose a CDE framework for (i) estimating the posterior density (Section 2.1), (ii) comparing the performance of ABC and related methods (Section 2.2), and (iii) choosing optimal summary statistics (Section 2.3).

2.1 Estimating the Posterior Density via CDE

Given a prior distribution $f(\theta)$ and a likelihood function $f(\mathbf{x}|\theta)$, our goal is to compute the posterior distribution $f(\theta|\mathbf{x}_o)$, where $\mathbf{x}_o$ is the observed sample. We assume we know how to sample from $f(\mathbf{x}|\theta)$ for a fixed value of θ.

A naive way of estimating $f(\theta|\mathbf{x}_o)$ via CDE methods is to first generate an i.i.d. sample $\mathcal{T} = \{(\theta_1, \mathbf{X}_1), \ldots, (\theta_B, \mathbf{X}_B)\}$ by sampling $\theta \sim f(\theta)$ and then $\mathbf{X} \sim f(\mathbf{x}|\theta)$ for each pair. One applies the CDE method of choice to $\mathcal{T}$, and then simply evaluates the estimated density $\widehat{f}(\theta|\mathbf{x})$ at $\mathbf{x} = \mathbf{x}_o$. The naive approach however may lead to poor results because some $\mathbf{x}$ are far from the observed data $\mathbf{x}_o$. To put it differently, standard conditional density estimators are designed to estimate $f(\theta|\mathbf{x})$ for every $\mathbf{x}$, but in ABC applications we are only interested in $\mathbf{x}_o$.

To solve this issue, one can instead estimate $f(\theta|\mathbf{x})$ using a training set $\mathcal{T}$ that only consists of sample points $\mathbf{x}$ close to $\mathbf{x}_o$. This training set is created by a simple ABC rejection sampling algorithm. More precisely: for a fixed distance function $d(\mathbf{x}, \mathbf{x}_o)$ (that could be based on summary statistics) and tolerance level ϵ, we construct a sample $\mathcal{T}$ according to Algorithm 1.

Algorithm 1 Training set for CDE via Rejection ABC

Input: Tolerance level ϵ, number of desired sample points B, distance function d, sample $\mathbf{x}_0$
Output: Training set $\mathcal{T}$ which approximates the joint distribution of $(\theta, \mathbf{X})$ in a neighborhood of $\mathbf{x}_0$

```
Let $\mathcal{T} = \{\}$
while $|\mathcal{T}| < B$ do
    Sample $\theta \sim f(\theta)$
    Sample $\mathbf{X} \sim f(\mathbf{x}|\theta)$
    If $d(\mathbf{x}, \mathbf{x}_o) < \epsilon$, let $\mathcal{T} \longleftarrow \mathcal{T} \cup \{(\theta, \mathbf{x})\}$
end while
return $\mathcal{T}$
```

To this new training set $\mathcal{T}$, we then apply our conditional density estimator, and finally evaluate the estimate at $\mathbf{x} = \mathbf{x}_o$. This procedure can be regarded as an ABC post-processing technique (Marin et al., 2012): the first (ABC) approximation to the posterior is obtained via the sample $\theta_1, \ldots, \theta_B$, which can be seen as a sample from $f(\theta|d(\mathbf{X}, \mathbf{x}_o) < \epsilon)$. That is, the standard ABC rejection sampler is implicitly performing conditional density estimation using an i.i.d. sample from the joint distribution of the data and the parameter. We take the results of the ABC sampler and estimate the conditional density exactly at the point

$\mathbf{x} = \mathbf{x}_o$ using other forms of conditional density estimation. If done correctly, the idea is that we can improve upon the original ABC approximation *even without*, as is currently the norm, simulating new data or decreasing the tolerance level ϵ.

Remark 1. *For simplicity, we focus on standard ABC rejection sampling, but one can use other ABC methods, such as sequential ABC (Sisson et al., 2007) or population Monte Carlo ABC (Beaumont et al., 2009), to construct $\mathcal{T}$. The data $\mathbf{x}$ can either be the original data vector, or a vector of summary statistics. We revisit the issue of summary statistic selection in Section 2.3.*

Next, we review FlexCode (Izbicki and Lee, 2017), which we use as a general-purpose methodology for estimating $f(\theta|\mathbf{x})$. However, many aspects of the paper (such as the novel approach to method selection without knowledge of the true posterior) hold for other CDE and ABC methods as well. In Section 4, for example, we use our surrogate loss to choose the tuning parameters of a nearest-neighbors estimator (Equation 15), which includes ABC as a special case.

FlexCode as a "Plug-In" CDE Method. For simplicity, assume that we are interested in estimating the posterior distribution of a single parameter $\theta \in \Re$, even if there are several parameters in the problem.[1] Similar ideas can be used if one is interested in estimating the (joint) posterior distribution for more than one parameter (see Izbicki and Lee 2017 for more details on how FlexCode can be adapted to those settings). In the context of ABC, $\mathbf{x}$ typically represents a set of statistics computed from the original data; recall Remark 1. We start by specifying an orthonormal basis $(\phi_i)_{i\in\mathbb{N}}$ in $\Re$. This basis will be used to model the density $f(\theta|\mathbf{x})$ *as a function of* θ. Note that there is a wide range of (orthogonal) bases one can choose from to capture any challenging shape of the density function of interest

[1] Most inference problems can be expressed as the computation of unidimensional quantities. Say one is interested in estimating m functions of parameters of the model $\boldsymbol{\theta}$; $g_1, \ldots, g_m$. One can then (i) use ABC to obtain a single simulation set $\mathcal{T} = \{(\boldsymbol{\theta}_1, \mathbf{X}_1), \ldots, (\boldsymbol{\theta}_B, \mathbf{X}_B)\}$, (ii) for each function g_i, compute $\mathcal{T}^{g_i} = \{(g_i(\boldsymbol{\theta}_1), \mathbf{X}_1), \ldots, (g_i(\boldsymbol{\theta}_B), \mathbf{X}_B)\}$, and then (iii) fit a (univariate) conditional density estimator to $\mathcal{T}^{g_i}$ to estimate $f(g_i(\boldsymbol{\theta})|\mathbf{x}_o)$. Note that (ii) is typically fast and (iii) can be performed in parallel; hence, the posterior distributions of all quantities of interest can be estimated with essentially no additional computational cost.

(Mallat, 1999). For instance, a natural choice for reasonably smooth functions $f(\theta|\mathbf{x})$ is the Fourier basis:

$$\phi_1(\theta) = 1; \qquad \phi_{2i+1}(\theta) = \sqrt{2}\sin(2\pi i\theta),\ i \in \mathbb{N}; \qquad \phi_{2i}(\theta) = \sqrt{2}\cos(2\pi i\theta),\ i \in \mathbb{N}$$

The key idea of FlexCode is to notice that, if $\int f^2(\theta|\mathbf{x})d\theta < \infty$ for every $\mathbf{x} \in \mathcal{X}$, then it is possible to expand $f(\theta|\mathbf{x})$ as $f(\theta|\mathbf{x}) = \sum_{i\in\mathbb{N}} \beta_i(\mathbf{x})\phi_i(\theta)$, where the expansion coefficients are given by

$$\beta_i(\mathbf{x}) = \mathbb{E}\left[\phi_i(\theta)|\mathbf{x}\right]. \tag{1}$$

That is, each $\beta_i(\mathbf{x})$ is a *regression function*. More precisely, it is the regression of a transformation of the response variable, $\phi_i(\theta)$, on the data $\mathbf{x}$. The FlexCode estimator is defined as $\widehat{f}(\theta|\mathbf{x}) = \sum_{i=1}^{I} \widehat{\beta}_i(\mathbf{x})\phi_i(\theta)$, where $\widehat{\beta}_i(\mathbf{x})$ are regression estimates. The cutoff I in the series expansion is a tuning parameter that controls the bias/variance tradeoff in the final density estimate, and which we choose via data splitting (Section 2.2).

With FlexCode, the problem of high-dimensional conditional density estimation boils down to choosing appropriate methods for estimating the regression functions $\mathbb{E}\left[\phi_i(\theta)|\mathbf{x}\right]$. The key advantage of FlexCode is that it offers more flexible CDE methods: By taking advantage of existing regression methods, which can be "plugged in" into the CDE estimator, we can adapt to the intrinsic structure of high-dimensional data (e.g., manifolds, irrelevant covariates, and different relationships between $\mathbf{x}$ and the response θ), as well as handle different data types (e.g., mixed data and functional data) and massive data sets (by using, e.g., xgboost (Chen and Guestrin, 2016)). See Izbicki and Lee (2017) and the upcoming LSST-DESC photo-z DC1 paper for examples; an implementation of FlexCode that allows for wavelet bases can be found at https://github.com/rizbicki/FlexCoDE (R) and https://github.com/tpospisi/flexcode (Python).

2.2 Method Selection: Comparing Different Estimators of the Posterior

Definition of a Surrogate Loss. Ultimately, we need to be able to decide which approach is best for approximating $f(\theta|\mathbf{x}_o)$ without knowledge of the true posterior. Ideally we would like to find an estimator $\widehat{f}(\theta|\mathbf{x}_o)$ such that the integrated squared-error (ISE) loss

$$L_{\mathbf{x}_o}(\widehat{f}, f) = \int (\widehat{f}(\theta|\mathbf{x}_o) - f(\theta|\mathbf{x}_o))^2 d\theta \tag{2}$$

is small. Unfortunately, one cannot compute $L_{\mathbf{x}_o}$ without knowing the true $f(\theta|\mathbf{x}_o)$, which is why method selection is so hard in practice. To overcome this issue, we propose the *surrogate* loss function

$$L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) = \int\int (\widehat{f}(\theta|\mathbf{x}) - f(\theta|\mathbf{x}))^2 \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x},\mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X},\mathbf{x}_o) < \epsilon)} d\theta d\mathbf{x}, \tag{3}$$

which enforces a close fit in an ϵ-neighborhood of $\mathbf{x}_o$. Here, the denominator $\mathbb{P}(d(\mathbf{X},\mathbf{x}_o) < \epsilon)$ is simply a constant that makes $\frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x},\mathbf{x}_o)<\epsilon)}{\mathbb{P}(d(\mathbf{X},\mathbf{x}_o)<\epsilon)}$ a proper density in $\mathbf{x}$.

The advantage with the above definition is that we can directly *estimate* $L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f)$ from the ABC posterior sample. Indeed, it holds that $L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f)$ can be written as

$$\begin{aligned}
&\int\int \widehat{f}^2(\theta|\mathbf{x}) \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x},\mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X},\mathbf{x}_o) < \epsilon)} d\theta d\mathbf{x} - 2\int\int \widehat{f}(\theta|\mathbf{x}) f(\theta|\mathbf{x}) \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x},\mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X},\mathbf{x}_o) < \epsilon)} d\theta d\mathbf{x} + K_f \\
&= \mathbb{E}_{\mathbf{X}'}\left[\int \widehat{f}^2(\theta|\mathbf{X}') d\theta\right] - 2\mathbb{E}_{(\theta',\mathbf{X}')}\left[\widehat{f}(\theta'|\mathbf{X}')\right] + K_f,
\end{aligned} \tag{4}$$

where $(\theta', \mathbf{X}')$ is a random vector with distribution induced by a sample generated according to the ABC rejection procedure in Algorithm 1; and K_f is a constant that does not depend on the estimator $\widehat{f}(\theta|\mathbf{x}_o)$. It follows that, given an independent validation or test sample of size B' of the ABC algorithm, $(\theta'_1, \mathbf{X}'_1), \ldots, (\theta'_B, \mathbf{X}'_B)$, we can estimate $L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f)$ (up to the constant K_f) via

$$\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) = \frac{1}{B'}\sum_{k=1}^{B'}\int \widehat{f}^2(\theta|\mathbf{x}'_k)d\theta - 2\frac{1}{B'}\sum_{k=1}^{B'}\widehat{f}(\theta'_k|\mathbf{x}'_k) \tag{5}$$

When given a set of estimators $\mathcal{F} = \{\widehat{f}_1, \ldots, \widehat{f}_m\}$, we select the method with the smallest estimated surrogate loss,

$$\widehat{f}^* := \arg\min_{\widehat{f}\in\mathcal{F}} \widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f)$$

Example 1 (Model selection based on CDE surrogate loss versus regression MSE loss)**.** Suppose we wish to estimate the posterior distribution of the mean of a Gaussian distribution with variance one. The left plot of Figure 2 shows the performance of a kernel nearest-neighbors density estimator (Equation 15) with the kernel bandwidth h and the number of nearest neighbors k chosen via (i) the estimated surrogate loss of Equation 5 versus (ii) a standard regression mean-squared-error loss.[2] The proposed surrogate loss clearly leads to better estimates of the posterior $f(\theta|\mathbf{x}_o)$ with smaller true loss (Equation 2). Indeed, as the right plot shows, if one chooses tuning parameters via the standard regression mean-squared-error loss, the estimates end up being very far from the true distribution.

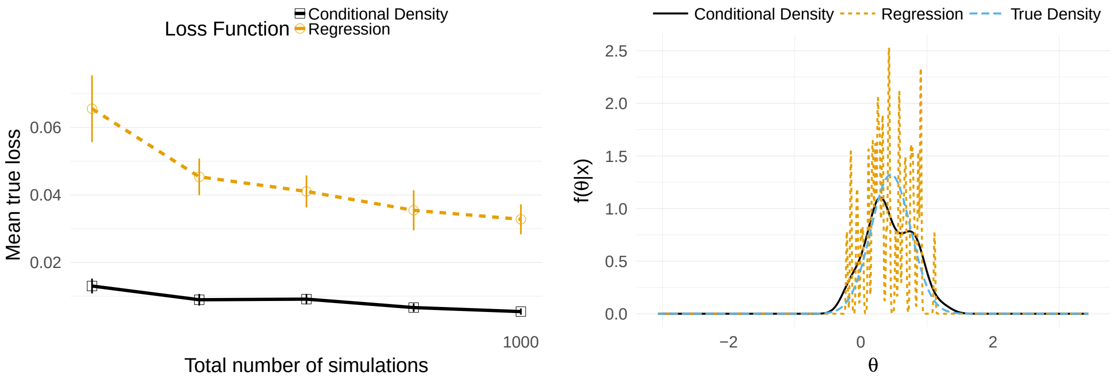


Figure 2: *Left:* Performance of a nearest-neighbors kernel density estimator with tuning parameters chosen via the surrogate loss of Equation 5 (continuous line) and via standard regression MSE loss (dashed line). *Right:* Estimated posterior distributions tuned according to both criteria after 1000 simulations. The surrogate loss of Equation 5 clearly leads to a better approximation.

[2] The data are simulated using the same gaussian model as in Section 4, but with $n = 10$, $\bar{x} = 0.5$ and at an acceptance ratio equal to 1 (that is, $\epsilon \to \infty$) and the number of simulations, B, varying.

Properties of the Surrogate Loss. Next we investigate the conditions under which the estimated surrogate loss is close to the true loss; the proofs for all results can be found in the Supplementary Material. The following theorem states that, if $(\widehat{f}(\theta|\mathbf{x}) - f(\theta|\mathbf{x}))^2$ is a smooth function of $\mathbf{x}$, then the (exact) surrogate loss $L^{\epsilon}_{\mathbf{x}_o}$ is close to $L_{\mathbf{x}_o}$ for small values of ϵ.

Theorem 1. *Assume that, for every $\theta \in \Theta$, $g_\theta(\mathbf{x}) := (\widehat{f}(\theta|\mathbf{x}) - f(\theta|\mathbf{x}))^2$ satisfies the Hölder condition of order β with a constant K_θ[3] such that $K_H := \int K_\theta d\theta < \infty$. Then $|L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) - L_{\mathbf{x}_o}(\widehat{f}, f)| \leq K_H \epsilon^\beta = O(\epsilon^\beta)$*

The next theorem shows that the estimator $\widehat{L}^{\epsilon}_{\mathbf{x}_o}$ in Equation 5 does indeed converge to the true loss $L_{\mathbf{x}_o}(\widehat{f}, f)$.

Theorem 2. *Let K_f be as in Equation 9. Under the assumptions of Theorem 4, $|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f - L_{\mathbf{x}_o}(\widehat{f}, f)| = O(\epsilon^\beta) + O_P(1/\sqrt{B'})$*

Under some additional conditions, it is also possible to guarantee that not only the estimated surrogate loss is close to the true loss, but that the result holds uniformly for a finite class of estimators of the posterior distribution. This is formally stated in the following theorem.

Theorem 3. *Let $\mathcal{F} = \{\widehat{f}_1, \ldots, \widehat{f}_m\}$ be a set of estimators of $f(\theta|\mathbf{x}_o)$. Assume that there exists M such that $|\widehat{f}_i(\theta|\mathbf{x})| \leq M$ for every $\mathbf{x}$, θ, and $i = 1, \ldots, m$.[4] Moreover, assume that for every $\theta \in \Theta$, $g_{i,\theta}(\mathbf{x}) := (\widehat{f}_i(\theta|\mathbf{x}) - f(\theta|\mathbf{x}))^2$ satisfies the Hölder condition of order β with constants K_θ such that $K_H := \int K_\theta d\theta < \infty$. Then, for every $\nu > 0$,*

$$\mathbb{P}\left(\max_{\widehat{f} \in \mathcal{F}} |\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f - L_{\mathbf{x}_o}(\widehat{f}, f)| \geq K_\epsilon \epsilon^\beta + \nu\right) \leq 2me^{-\frac{B'\nu^2}{2(M^2+2M)^2}}.$$

[3] That is, there exists a constant K_θ such that for every $\mathbf{x}, \mathbf{y} \in \Re^d$ $|g_\theta(\mathbf{x}) - g_\theta(\mathbf{y})| \leq K_\theta(d(\mathbf{x}, \mathbf{y}))^\beta$.

[4] Such assumptions hold if the $\widehat{f}_i$'s are obtained via FlexCode with bounded basis functions (e.g., Fourier basis) or a kernel density estimator on the ABC samples.

The next corollary shows that the procedure we propose in this section, with high probability, picks an estimate of the posterior density that has a true loss that is close to the true loss of the best method in $\mathcal{F}$.

Corollary 1. *Let* $\widehat{f}^* := \arg\min_{\widehat{f}\in\mathcal{F}} \widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f)$ *be the best estimator in* $\mathcal{F}$ *according to the estimated surrogate loss, and let* $f^* = \arg\min_{\widehat{f}\in\mathcal{F}} L_{\mathbf{x}_o}(\widehat{f}, f)$ *be the best estimator in* $\mathcal{F}$ *according to the true loss. Then, under the assumptions from Theorem 6, with probability at least* $1 - 2me^{-\frac{B'\nu^2}{2(M^2+2M)^2}}$, $L_{\mathbf{x}_o}(\widehat{f}^*, f) \leq L_{\mathbf{x}_o}(f^*, f) + 2(K_H\epsilon^\beta + \nu)$.

2.3 Summary Statistics Selection

In a typical ABC setting, there are a large number of available summary statistics. Standard ABC fails if all of them are used simultaneously, especially if some statistics carry little information about the parameters of the model (Blum, 2010).

We propose to use FlexCode as a way of either (i) directly estimating $f(\theta|\mathbf{x}_o)$ when there are a large number of summary statistics,[5] or (ii) assigning an importance measure to each summary statistic, which later can be used for variable selection in ABC and related procedures.

There are two versions of FlexCode that are particularly useful for variable selection: FlexCode-SAM[6] and FlexCode-RF[7]. Izbicki and Lee (2017) show that both estimators automatically adapt to the number of relevant covariates, i.e., the number of covariates that influence the distribution of the response. In the context of ABC, this means that these methods are able to automatically detect which summary statistics are relevant in estimating the posterior distribution of θ. Corollary 1 from Izbicki and Lee (2017) implies that, *if* indeed only m out of all d summary statistics influence the distribution of θ, then the rate of convergence of these methods is $O\left(n^{-2\beta/(2\beta+m\frac{2\beta+1}{2\alpha}+1)}\right)$ instead of $O\left(n^{-2\beta/(2\beta+d\frac{2\beta+1}{2\alpha}+1)}\right)$,

[5] the dimension reduction is then implicit in the choice of (high-dimensional) regression method

[6] FlexCode with the coefficients from Equation 1 estimated via Sparse Additive Models (Ravikumar et al., 2009)

[7] FlexCode with the coefficients from Equation 1 estimated via Random Forests

where α and β are numbers associated to the smoothness of $f(\theta|\mathbf{x})$. The former rate implies a much faster convergence: if $m \ll d$, it is essentially the rate one would obtain if one knew which were the relevant statistics. In such a setting, there is no need to explicitly perform summary statistic selection prior to estimating the posterior; FlexCode-SAM or FlexCode-RF automatically remove irrelevant covariates.

More generally, one can use FlexCode to compute an importance measure for summary statistics (to be used in other procedures than FlexCode). It turns out that one can infer the relevance of the j:th summary statistic *in posterior estimation* from its relevance in estimating the I first *regression* functions in FlexCode – even if we do not use FlexCode for estimating the posterior. More precisely, assume that $\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_d)$ is a vector of summary statistics, and let $\mathbf{x}'_j = (x_1, \ldots, x_{j-1}, x'_j, x_{j+1}, \ldots, x_d)$. Define the relevance of variable j to the posterior distribution $f(\theta|\mathbf{x})$ as

$$r_j := \int\int\int (f(\theta|\mathbf{x}) - f(\theta|\mathbf{x}'_j))^2 d\mathbf{x} dx'_j d\theta,$$

and its relevance to the regression $\beta_i(\mathbf{x})$ in Equation 1 as

$$r_{i,j} := \int\int \left(\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j)\right)^2 d\mathbf{x} dx'_j.$$

Under some smoothness assumptions with respect to θ, the two metrics are related.

Assumption 1 (Smoothness in θ direction)**.** *$\forall \mathbf{x} \in \mathcal{X}$, we assume that $f(\theta|\mathbf{x}) \in W_\phi(s_\mathbf{x}, c_\mathbf{x})$, the Sobolev space of order s and radius c,*[8] *where $f(\theta|\mathbf{x})$ is viewed as a function of θ, and $s_\mathbf{x}$ and $c_\mathbf{x}$ are such that $\inf_\mathbf{x} s_\mathbf{x} \overset{def}{=} \beta > \frac{1}{2}$ and $\int c^2_\mathbf{x} d\mathbf{x} < \infty$.*

Proposition 1. *Under Assumption 2, $r_j = \sum_{i=1}^{I} r_{i,j} + O\left(I^{-2\beta}\right)$*

Now let $u_{i,j}$ denote a measure of importance of the j:th summary statistic in estimating

[8] For every $s > \frac{1}{2}$ and $0 < c < \infty$, $W_\phi(s, c) := \{f = \sum_{i \geq 1} \theta_i \phi_i : \sum_{i \geq 1} a_i^2 \theta_i^2 \leq c^2\}$, where $a_i \sim (\pi i)^s$. Notice that for the Fourier basis $(\phi_i)_i$, this is the standard definition of the Sobolev space of order s and radius c; it is the space of functions that have their s-th weak derivative bounded by c^2 and integrable in L^2.

regression i (Equation 1). For instance, for FlexCode-RF, $u_{i,j}$ may represent the mean decrease in the Mean Squared Error (Hastie et al., 2001); for FlexCode-SAM, $u_{i,j}$ may be value of the indicator function for the j:th summary statistic when estimating $\beta_i(\mathbf{x})$. Motivated by Proposition 2, we define an *importance measure* for the j:th summary statistic *in posterior estimation* according to

$$u_j := \frac{1}{I}\sum_{i=1}^{I} u_{i,j}. \tag{6}$$

We can use these values to select variables for estimating $f(\theta|\mathbf{x}_o)$ via other ABC methods. For example, one approach is to choose all summary statistics such that $u_j > t$, where the threshold value t is defined by the user. We will further explore this approach in Section 3.

In summary, our procedure has two main advantages compared to current state-of-the-art approaches for selecting summary statistics in ABC: (i) it chooses statistics that lead to good estimates of the *entire* posterior distribution $f(\theta|\mathbf{x}_o)$ rather than surrogates, such as, the regression or posterior mean $\mathbb{E}[\theta|\mathbf{x}_o]$ (Aeschbacher et al., 2012; Creel and Kristensen, 2016; Faisal et al., 2016), and (ii) it is typically faster than most other approaches; in particular, it is significantly faster than best subset selection which scales as $O(2^d)$, whereas, e.g., FlexCode-RF scales as $O(Id)$, and FlexCode-SAM scales as $O(Id^3)$.

3 Experiments

3.1 Examples with Known Posteriors

We start by analyzing examples with well-known and analytically computable posterior distributions:

1. **Mean of a Gaussian with known variance.** $X_1, \ldots, X_{20}|\mu \overset{iid}{\sim} \text{Normal}(\mu, 1)$, $\mu \sim \text{Normal}(0, \sigma_0^2)$. We repeat the experiments for σ_0 in an equally spaced grid with ten values between 0.5 and 100.

2. **Precision of a Gaussian with unknown precision.** $X_1, \ldots, X_{20}|(\mu, \tau) \overset{iid}{\sim}$ Normal$(\mu, 1/\tau)$, $(\mu, \tau) \sim$ Normal-Gamma$(\mu_0, \nu_0, \alpha_0, \beta_0)$. We set $\mu_0 = 0$, $\nu_0 = 1$, and repeat the experiments choosing α_0 and β_0 such that $\mathbb{E}[\tau] = 1$ and $\sqrt{\mathbb{V}[\tau]}$ is in an equally spaced grid with ten values between 0.1 and 5.

In Supplementary Materials we also investigate a third setting, "Mean of a Gaussian with unknown precision", with results similar to those shown here in the main manuscript.

In all examples here, *observed* data $\mathbf{x}_o$ are drawn from a Normal$(0, 1)$ distribution. We run each experiment 200 times, that is, with 200 different values of $\mathbf{x}_o$. The training set $\mathcal{T}$, which is used to build conditional density estimators, is constructed according to Algorithm 1 with $B = 10,000$ and a tolerance level ϵ that corresponds to an acceptance rate of 1%. For the distance function $d(\mathbf{x}, \mathbf{x}_o)$, we choose the Euclidean distance between minimal sufficient statistics normalized to have mean zero and variance 1; these statistics are $\bar{\mathbf{x}}$ for scenario 1 and $(\bar{\mathbf{x}}, s)$ for scenario 2. We use a Fourier basis for all FlexCode experiments in the paper, but wavelets lead to similar results.

We compare the following methods:

- ABC: rejection ABC method with the minimal sufficient statistics (that is, apply a kernel density estimator to the θ coordinate of $\mathcal{T}$, with bandwidth chosen via cross-validation),
- FlexCode_Raw-NN: FlexCode estimator with Nearest Neighbors regression,
- FlexCode_Raw-Series: FlexCode estimator with Spectral Series regression (Lee and Izbicki, 2016), and
- FlexCode_Raw-RF: FlexCode estimator with Random Forest regression.

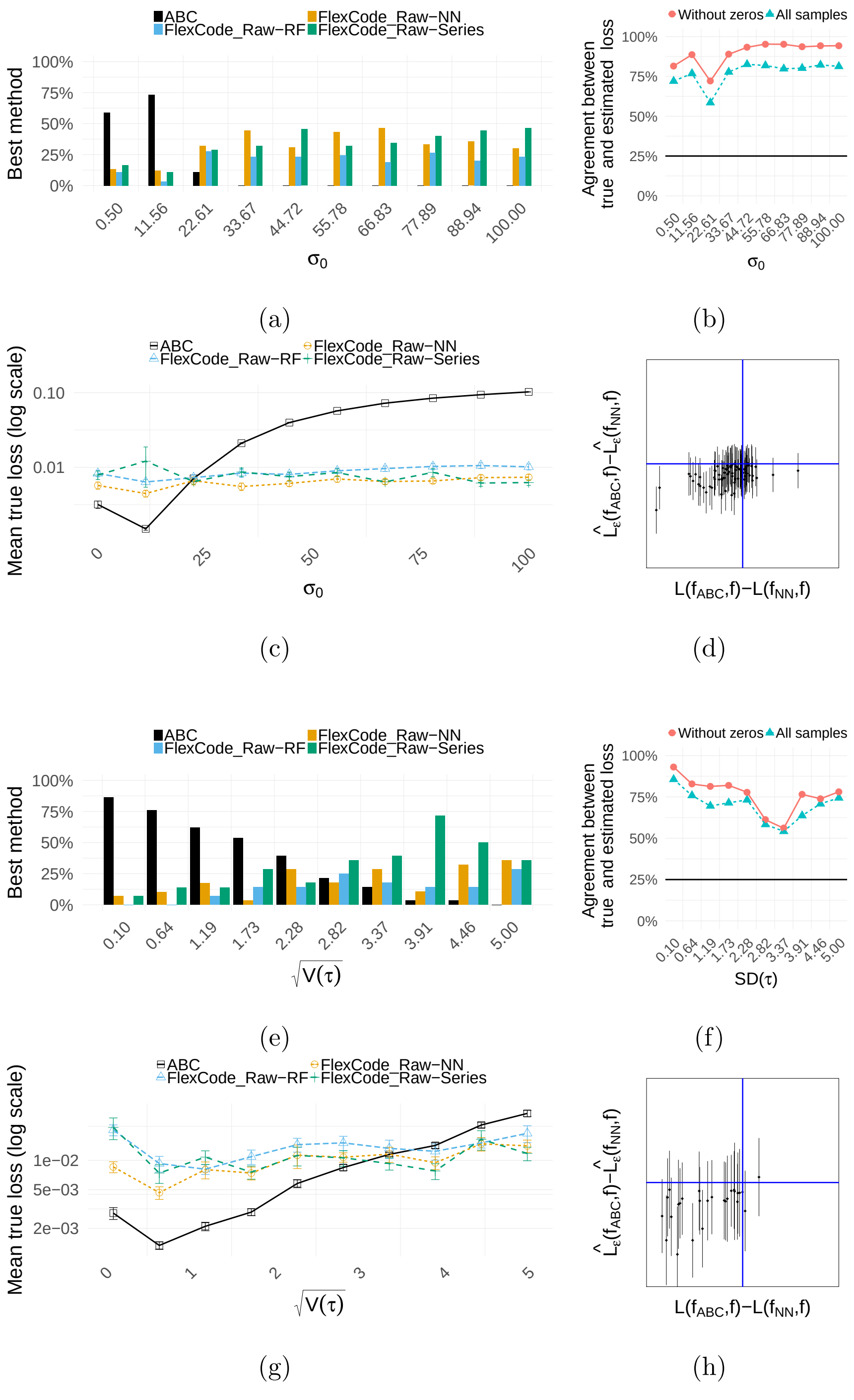
The three FlexCode estimators (denoted by "Raw") are directly applied to the sorted values of the *original* covariates $X_{(1)}, \ldots, X_{(20)}$. That is, we do not use minimal sufficient statistics or other summary statistics. To assess the performance of each method, we compute the true loss $L_{\mathbf{x}_o}$ (Equation 2) for each $\mathbf{x}_o$. In addition, we estimate the surrogate loss $L^{\epsilon}_{\mathbf{x}_o}$ according to Equation 5 using a new sample of size $B' = 10,000$ from Algorithm 1.

3.1.1 CDE and Method Selection

In this section, we investigate whether various FlexCode CDE methods improve upon standard ABC for the settings described above. We also evaluate the method selection approach in Section 2.2 by comparing decisions based on estimated surrogate losses to those made if one knew the true ISE losses.

Figure 3, left, shows how well the methods actually estimate the posterior density for Settings 1-2. Panel (a) and (e) list the proportion of times each method returns the best results (according to the true loss from Equation 2). Generally speaking, the larger the prior variance, the better FlexCode-based methods perform compared to ABC. In particular, while for small variances ABC tends to be better, for large prior variances, FlexCode with Nearest Neighbors regression tends to give the best results. FlexCode with expansion coefficients estimated via Spectral Series regression is also very competitive. Panels (c) and (g) confirms these results; here we see the average true loss of each method along with standard errors.

Figure 3, right, summarizes the performance of our method selection algorithm. Panels (b) and (f) list the proportion of times the method chosen by the true loss (Equation 2) matches the method chosen via the estimated loss (Equation 5) in all pairwise comparisons; that is, the plot tells us how often the method selection procedure proposed in Section 2.2 actually works. We present two variations of the algorithm: in the first version (see triangles), we include all the data; in the second version (see circles), we remove cases where the confidence interval for $\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}_1, f) - \widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}_2, f)$ contains zero (i.e., cases where we cannot tell whether $\widehat{f}_1$ or $\widehat{f}_2$ performs better). The baseline shows what one would expect if the method selection algorithm was totally random. The plots indicate that we, in all settings, roughly arrive at the same conclusions with the estimated surrogate loss as we would if we knew the true loss.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 3: Panels (a)-(d): CDE and method selection results for scenario 1 (mean of a Gaussian with known variance). *Left:* Panels (a) and (c) show that the rejection ABC leads to better estimates of the posterior density $f(\theta|\mathbf{x}_o)$ when the prior variance σ_0 is small, but the NN and Series versions of FlexCode yield better estimates for moderate and large values of σ_0. *Right:* Panels (b) and (d) indicate that by estimating the surrogate loss function one can tell from the data which method is better for the problem at hand. The horizontal line in panel (b) represents the behavior of a random selection. Panels (e)-(h): CDE and method selection results for scenario 2 (precision of a Gaussian with unknown precision). Conclusions are analogous.

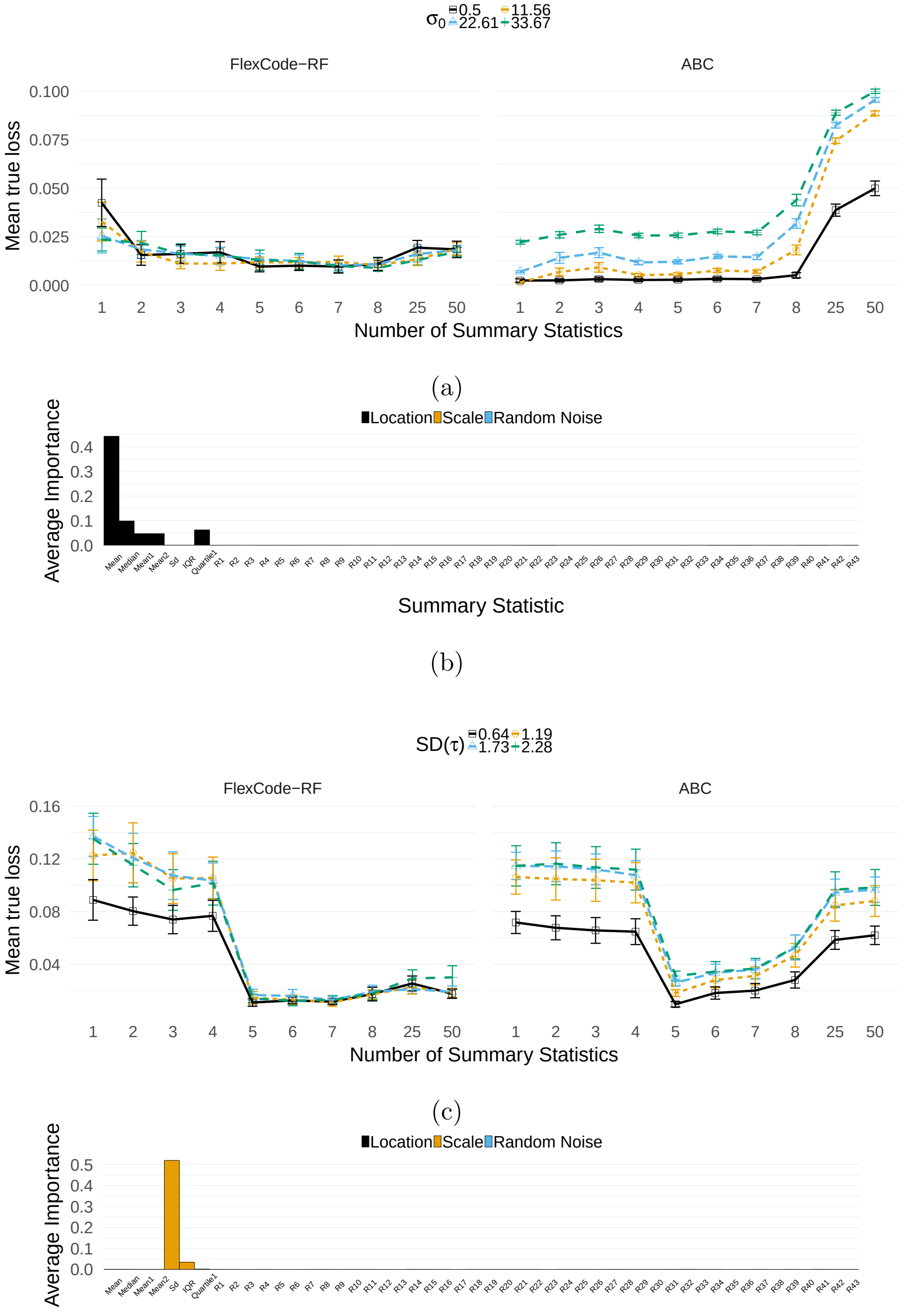
For the sake of illustration, we have also added panels (d) and (h), which show a scatterplot of differences between true losses versus the differences between the estimated losses for ABC and FlexCode_Raw-NN for the setting with $\sigma_0 = 0.5$ and $\sqrt{\mathbb{V}[\tau]} = 0.1$, respectively. The fact that most samples are either in the first or third quadrant further confirms that the estimated surrogate loss is in agreement with the true loss in terms of which method best estimates the posterior density.

3.1.2 Summary Statistic Selection

In this Section we investigate the performance of FlexCode-RF for summary statistics selection (Sec. 2.3). For this purpose, the following summary statistics were used:

1. Mean: average of the data points; $\frac{1}{n}\sum_{i=1}^{n} X_i$
2. Median: median of the data points; $\text{median}\{X_i\}_{i=1,\ldots,n}$
3. Mean 1: average of the first half of the data points; $\frac{1}{n/2}\sum_{i=1}^{n/2} X_i$
4. Mean 2: average of the second half of the data points; $\frac{n/2+1}{n}\sum_{i=n/2+1}^{n} X_i$
5. SD: standard deviation of the data points; $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{\mathbf{X}})^2}$
6. IQR: interquartile range of the data points; $\text{quantile}_{75\%}\{X_i\}_{i=1,\ldots,n} - \text{quantile}_{25\%}\{X_i\}_{i=1,\ldots,n}$
7. Quartile 1: first quantile of the data points; $\text{quantile}_{25\%}\{X_i\}_{i=1,\ldots,n}$

8–51. Independent random variables $\sim$ Normal$(0, 1)$, that is, random noise



(a)

(b)

(c)

(d)

Figure 4: Panels (a)-(b): Summary statistic selection for scenario 1 (mean of a Gaussian with known variance); Panels (c)-(d): Summary statistic selection for scenario 2 (precision of a Gaussian with unknown precision). Panels (a) and (c) show that ABC is highly sensitive to random noise (entries 8-51) with the estimates of the posteriors rapidly deteriorating with nuisance statistics. Nuisance statistics do not affect the performance of FlexCode-RF much. Furthermore, we see from panel (b) that FlexCode-RF identifies the location statistics (entries 1-5) as key variables for the first setting and assigns them a high average importance score. In the second setting, panel (d) indicates that we only need dispersion statistics (such as entry 5) to estimate the posteriors wells.

Figure 4 summarizes the results of fitting FlexCode-RF and ABC to these summary statistics for the different scenarios. Panel (a) and (c) show the true loss as we increase the number of statistics. More precisely: the values at $x = 1$ represent the true loss of ABC (left) and FlexCode-RF (right) when using only the mean (i.e., the first statistic); the points at $x = 2$ indicate the true loss of the estimates using only the mean and the median (i.e., the first and second statistics) and so on. We note that FlexCode-RF is robust to irrelevant summary statistics: the method virtually behaves as if they were not present. This is in sharp contrast with standard ABC, whose performance deteriorates quickly with added noise or nuisance statistics.

Furthermore, panels (b) and (d) show the average importance of each statistic, defined according to Equation 6, where $u_{i,j}$ is the mean decrease in the Gini index. These plots reveal that FlexCode-RF typically assigns a high score to sufficient summary statistics or to statistics that are highly correlated to sufficient statistics. For instance, in panel (b) (estimation of the mean of the distribution), measures of location are assigned a higher importance score, whereas measures of dispersion are assigned a higher score in panel (d) (estimation of the precision of the distribution). In all examples, FlexCode-RF assigns zero importance to random noise statistics. We conclude that our method for summary statistic selection indeed identifies relevant statistics for estimating the posterior $f(\theta|\mathbf{x}_o)$ well.

3.2 Application: Estimating a Galaxy's Dark Matter Density Profile

Next we consider more complex simulations. ΛCDM (Lambda cold dark matter) model is frequently referred to as the standard model of Big Bang cosmology; it is the simplest model that contains assumptions consistent with observational and theoretical knowledge of the Universe.

The ΛCDM model predicts that the dark matter profile of a galaxy in the absence of baryonic effects can be parameterized by the Navarro-French-White (NFW) model (Navarro, 1996). Given an observed galaxy, such as the Sculptor dwarf spheroidal galaxy, we wish to constrain the parameters of the NFW model. To begin we will only consider a single parameter, the critical energy E_c (Strigari et al. 2017, Equation 15), and set all other parameters at commonly accepted estimates; see Supplementary Materials for details.

The observed data $\mathbf{x}_0$ are velocities and coordinates of 200 stars in a galaxy, here simulated so as to follow the NFW model.[9] To perform ABC we define the distance function as the ℓ^2 norm between bivariate kernel density estimates of the joint distribution of the velocity and distance from the center. The same distance function will be used by FlexCode-NN and FlexCode-Series. Because the data are functional we also implement FlexCode-Functional, where the coefficients in FlexCode are estimated via functional kernel regression (Ferraty and Vieu, 2006).

To assess the performance of the CDE methods we generate 1000 test observations each with an ABC sample of 1000 accepted observations with an acceptance rate of 0.1. We use the prior $E_c \sim U(0.01, 1.0)$.

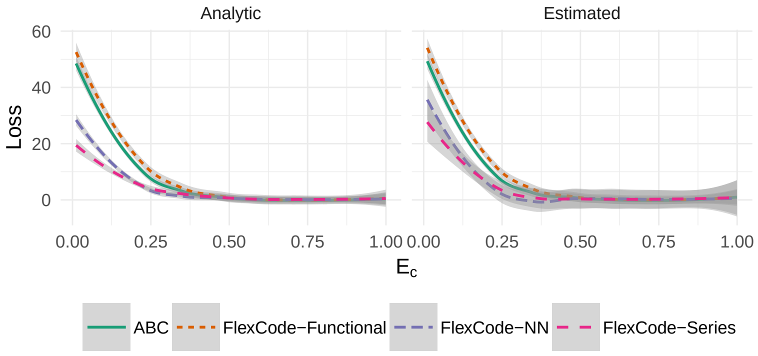


Figure 5: *Left:* True loss for each method in different parameter regions. *Right:* The estimated surrogate loss (here shifted with a constant $\int f(\theta|\mathbf{x}_o)^2 d\theta$ for easier comparison) can be used to identify when our methods improve upon the ABC estimates.

Figure 5, left, displays the true loss for each method. The plot indicates that, for at least some realizations with low true E_c, the estimates from the FlexCode estimators lead

[9] The simulations are written by Mao-Shen (Terrence) Liu and rely on an MCMC sampling scheme; the details are outlined in Liu and Walker 2018.

to better performance than ABC. Of most interest is that we reach similar conclusions with the estimated surrogate loss; see right plot. Thus the surrogate loss serves as a reasonable proxy for the true loss which in practical applications will be unavailable. Some examples of estimated posteriors can be found in the Supplementary Materials.

4 Approximate Bayesian Computation and Beyond

4.1 ABC with Fewer Simulations

As noted by (Blum, 2010; Biau et al., 2015), ABC is equivalent to a kernel-CDE. More specifically, it can be seen as a "nearest-neighbors" kernel-CDE (NN-KCDE) defined by

$$\widehat{f}_{\mathrm{nn}}(\theta \mid \mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} K_h(\rho(\theta, \theta_{s_i(\mathbf{x})})), \tag{7}$$

where $s_i(\mathbf{x})$ represents the index of the ith nearest neighbor to the target point $\mathbf{x}$ in covariate space, and we compute the conditional density of θ at $\mathbf{x}$ by applying a kernel smoother $K_h(\cdot)$ with bandwidth h to the k points closest to x.

For a given set of generated data, the above is equivalent to selecting the ABC threshold ϵ as the k/n-th quantile of the observed distances. This is commonly used in practice as it is more convenient than determining ϵ a priori. However, as pointed out by Biau et al. (2015) (Section 4; remark 1), there is currently no good methodology to select both k and h in an ABC k-nearest neighbor estimate.

Given the connection between ABC and NN-KCDE, we propose to use our surrogate loss to tune the estimator; selecting k and h such that they minimize the estimated surrogate loss in Equation 5. In this sense, we are selecting the "optimal" ABC parameters after generating some of the data: having generated 10,000 points, it may turn out that we would have preferred a smaller tolerance level ϵ and that only using the closest 1,000 points would better approximate the posterior.

Example with Normal Posterior. To demonstrate the effectiveness of our surrogate loss in reducing the number of simulations, we draw data $X_1, \ldots, X_5|\mu \overset{iid}{\sim} \text{Normal}(\mu, 0.2^2)$ where $\mu \sim \text{Normal}(1, 0.5^2)$. We examine the role of ABC thresholds by fitting the model for several values of the threshold with observed data $\mathbf{x}_0 = \{-0.5, -0.25, 0.0, 0.25, 0.5\}$. (A similar example with a two-dimensional normal distribution can be found in Supplementary Materials.) For each threshold, we perform rejection sampling until we retain $B = 1000$ ABC points. We select ABC thresholds to fix the acceptance rate of the rejection sampling; those acceptance rates are then used in place of the actual tolerance level ϵ for easier comparison.

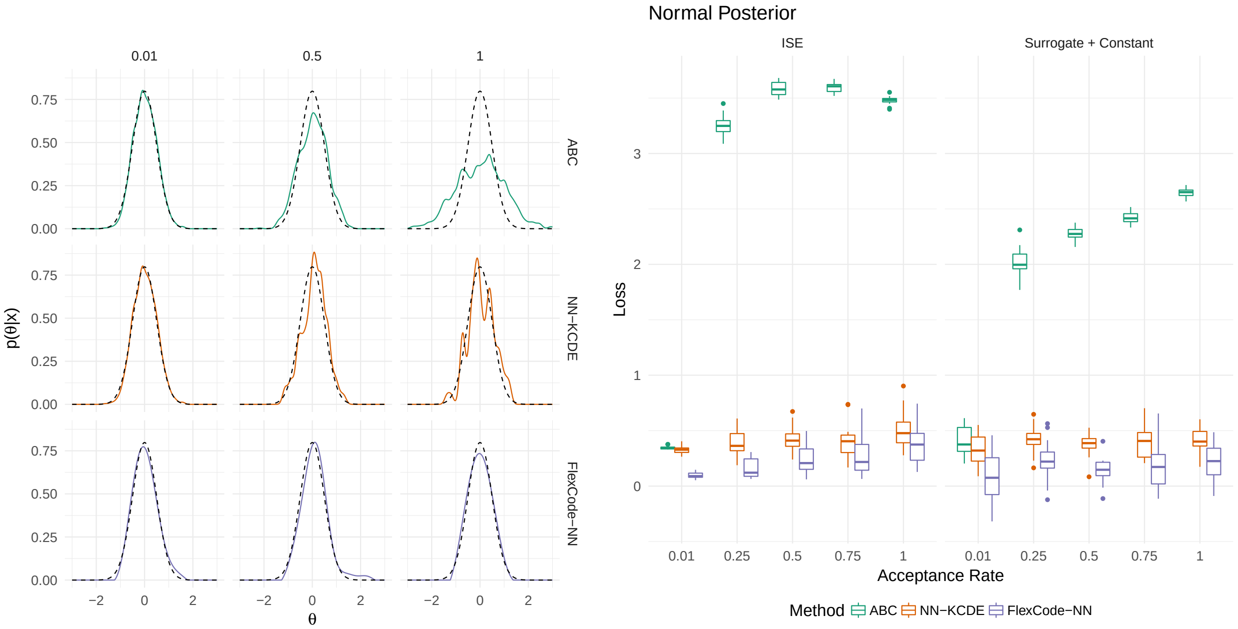


Figure 6: *Left:* Density estimates for normal posterior using ABC sample of varying acceptance rates. As the acceptance rate (or, equivalently, the ABC tolerance level) decreases, the ABC posterior approaches the true posterior. Both NN-KCDE and FlexCode-NN approximate the posterior well for all acceptance rates. *Right:* True integrated squared error (ISE) loss and estimated surrogate loss for normal posterior using ABC sample of varying acceptance rates. We need to decrease the acceptance rate considerably to attain a small loss for ABC. On the other hand, the losses for NN-KCDE and FlexCode-NN are relatively small even for large acceptance rates with the same data.

The left panel of Figure 6 shows examples of posterior densities for varying acceptance rates. For the highest acceptance rate of 1 (corresponding to the ABC tolerance level $\epsilon \to \infty$), the ABC posterior (top left) is the prior distribution and thus a poor estimate. In contrast, the two ABC-CDE methods (FlexCode-NN and NN-KCDE) have a decent performance even at an acceptance rate of 1; more generally, they perform well at a higher acceptance rate than standard ABC.

To corroborate this qualitative look, we examine the loss for each method. The right panel of Figure 6 plots the true and surrogate losses against the acceptance rate for the given methods. As seen in Section 3.2, the surrogate loss provides the same conclusion as the (unavailable in practice) true loss. As the acceptance rate decreases, the ABC realizations more closely approximate the true posterior and the ABC estimate of the posterior improves. The main result is that NN-KCDE and FlexCode-NN have roughly constant performance over all values of the acceptance rate. As such, we could generate *only* 1,000 realizations of the ABC sample at an acceptance rate of 1 and achieve *similar* result as standard ABC generating 100,000 values at an acceptance rate of 0.01.

There are two different sources of improvement: the first exhibited by NN-KCDE amounts to selecting the "optimal" ABC parameters k and h using surrogate loss. However, as FlexCode-NN performs slightly better than kernel-NN for the same sample, there is an additional improvement in using CDE methods other than kernel CDE; this difference becomes more pronounced for high-dimensional and complex data (see Izbicki and Lee 2017 for examples of when traditional kernel smoothers fail).

4.2 Comparison with ABC Post-Processing Methods

We can view the connection between ABC and NN-KCDE in yet another way: NN-KCDE provides a post-processing step for improving the density estimate obtained from ABC. There are several other post-processing procedures in the ABC literature, most notably the regression adjustment methods of Beaumont et al. (2002) and Blum and François (2010); see Li and Fearnhead (2017) for asymptotic results. These two methods use regression adjustment to correct for the impact of the conditional distribution changing with $\mathbf{x}$, modeling the response as

$$\theta_i = m(\mathbf{x}_i) + \sigma(\mathbf{x}_i) \times \epsilon_i,$$

where $m(\mathbf{x}_i)$ is the conditional expectation, and $\sigma(\mathbf{x}_i)$ is the conditional standard deviation. Assuming this is the true model, the sample points can be transformed to

$$\widetilde{\theta}_i = m(\mathbf{x}_o) + (\theta_i - m(\mathbf{x}_i)) \times \frac{\sigma(\mathbf{x}_o)}{\sigma(\mathbf{x}_i)}, \tag{8}$$

which scales the sample to have the same mean and standard deviation as the fitted distribution around $\mathbf{x}_o$.

A visual representation of this procedure can be seen in Figure 7. The data are drawn from the conjugate normal posterior described in Section 4.1. In the joint distribution we see a clear linear relationship between the summary statistic $\bar{\mathbf{x}}$ and parameter θ. The ABC joint sample we obtain from restricting our data to a neighborhoood around $\bar{\mathbf{x}}_{\text{obs}}$ is skewed, as seen in the ABC kernel density estimate. With access to the true $m(\mathbf{x})$ and $\sigma(\mathbf{x})$, however, one could transform the ABC joint sample with Equation 4.2 to remove the trend (see regression-adjusted joint distribution) and achieve a better fit (see regression-adjusted kernel density estimate). Beaumont et al. (2002) and Blum and François (2010) use local-linear and neural-net regression respectively to estimate $\widehat{m}(\mathbf{x})$ and $\widehat{\sigma}(\mathbf{x})$ with similar effect.

In Figure 8, we calculate the ABC density loss and CDE loss for the unimodal example, and see that regression-adjusted methods achieve similar performance to ABC-CDE (FlexCode-NN and NN-KCDE) here. We use the abc package (Csillery et al., 2012) to fit both the methods of Beaumont et al. (2002) and Blum and François (2010).

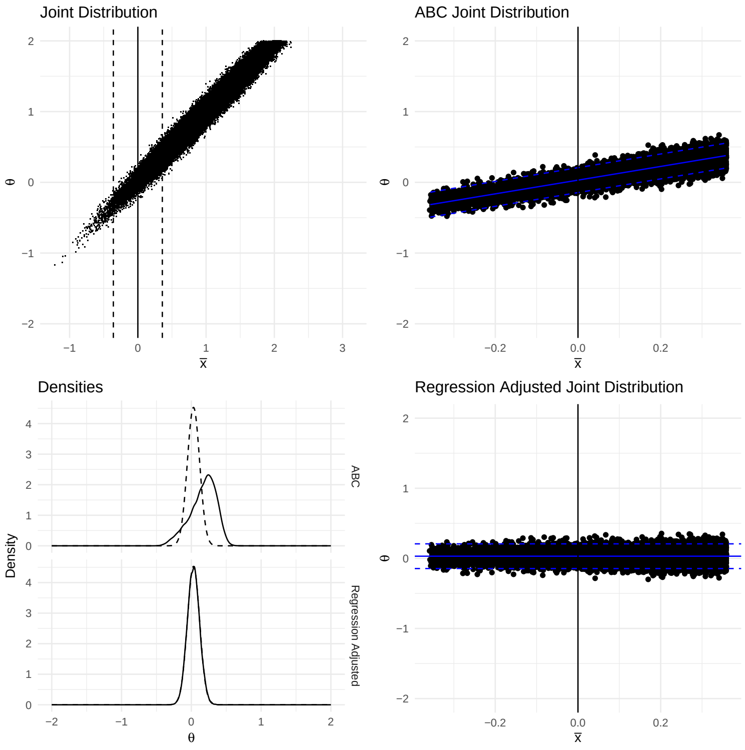


Figure 7: Regression adjustment in ABC for unimodal example assuming we know the true $m(\mathbf{x})$ and $\sigma(\mathbf{x})$. See text for details.

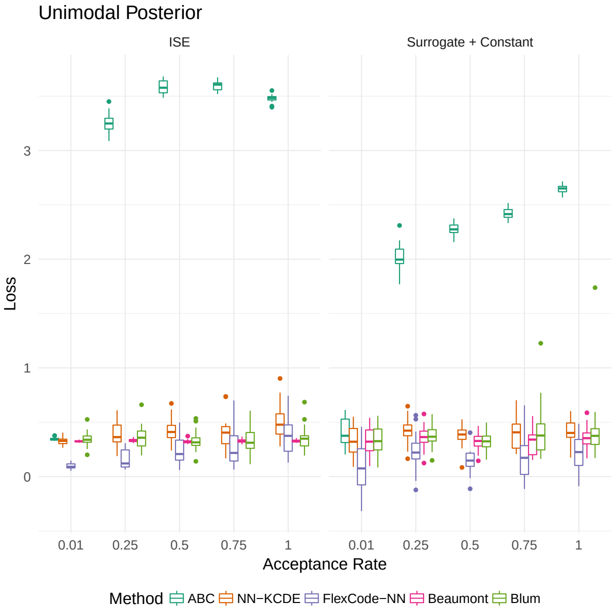


Figure 8: True integrated squared error loss (*left*) and estimated surrogate loss (*right*) for unimodal example using ABC sample of varying acceptance rates. Both regression adjustment and ABC-CDE methods improve upon standard ABC.

However, the regression-based methods rely upon the assumption that the distributions of ϵ_i at different $\mathbf{x}$ are similar up to a translation and scaling. To illustrate where this assumption can break down consider the case of a multimodal posterior. Given the mixture prior $\mu \sim \sum_i w_i \, \text{Normal}(\mu_i, \sigma_i^2)$ and likelihood $X_i \sim \text{Normal}(\mu, \sigma_x^2)$, we obtain the conjugate mixture model posterior

$$\mu \mid X \sim \sum_i w_i^* \, \text{Normal}(\mu_i^*, \sigma_i^{*2}),$$

where μ_i^* and σ_i^* are the parameters of the conjugate posterior for that particular normal prior. The mixing weights $w_i^* \propto w_i P_i(X)$ where $P_i(X)$ is the marginal likelihood under the i-th mixture component.

We follow the same procedure as Figure 7 for this setting resulting in Figure 9. Here the regression is misleading; the error distribution at $\bar{\mathbf{x}}_{\text{obs}}$ is bimodal whereas the error distribution away from $\bar{\mathbf{x}}_{\text{obs}}$ is unimodal. When the regression adjustment is made, the bimodality is lost and a single peak is fit. When we calculate the losses for this simulation (Figure 10), we see that the regression methods perform worse than ABC whereas the ABC-CDE methods still improve upon ABC.

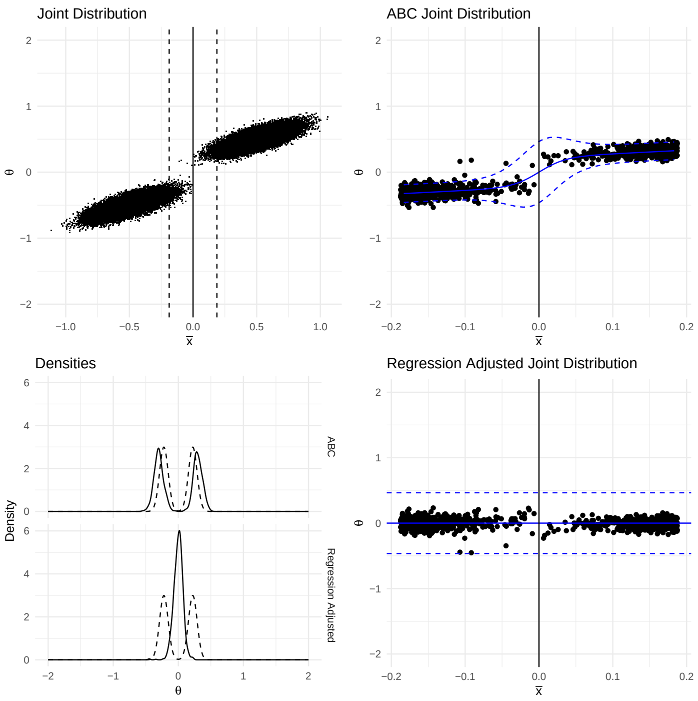


Figure 9: For the multimodal example, a regression of the ABC joint sample is misleading, as we cannot simply adjust for the change in the distribution of $\theta|\mathbf{x}$ around $\mathbf{x}_{\text{obs}}$ by shifting and rescaling the sample by the conditional mean $m(\mathbf{x})$ and the conditional variance $\sigma(\mathbf{x})$, respectively.

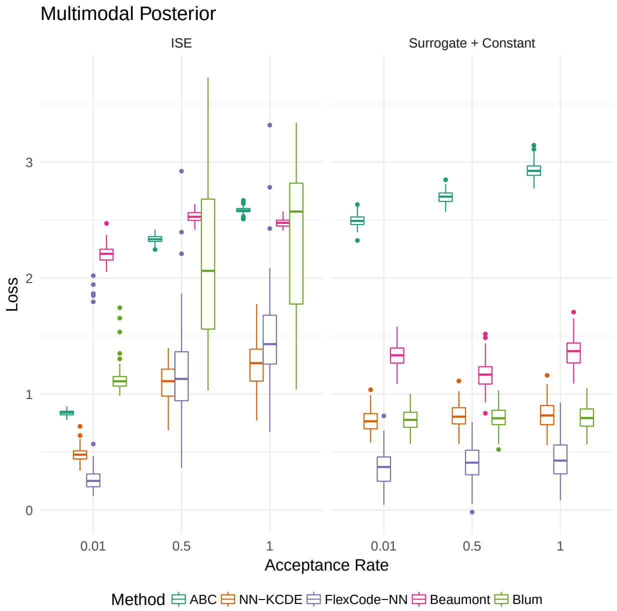


Figure 10: True integrated squared error loss (*left*) and estimated surrogate loss (*right*) for multimodal example using ABC sample of varying acceptance rates. ABC-CDE methods (e.g, NN-KCDE and FlexCode-NN) improve upon standard ABC, whereas regression adjustment methods (e.g., Beaumont and Blum) can lead to worse results.

4.3 Application: Cosmological Parameter Inference via Weak Lensing

We end by considering the problem of cosmological parameter inference via weak gravitational lensing. Gravitational lensing causes distortion in images of distance galaxies; this is called cosmic shear. Because the universe has varying matter densities, these create tidal gravitational fields which cause light to deflect differentially. The size and direction of distortion is directly related to the size and shape of the matter along that line of sight. We can use shear correlation functions to study the properties and evolution of the large scale structure and geometry of the Universe. In particular we can constrain parameters of the ΛCDM cosmological model such as the dark matter density Ω_M and matter power spectrum normalization σ_8. For further background see Hoekstra and Jain (2008), Munshi et al. (2008) and Mandelbaum (2017).

We can perform ABC rejection sampling using the Euclidean distance between binned shear correlation functions as our summary statistic. We use the `lenstools` package (Petri, 2016) to generate power spectra given parameter realizations. We then use the `GalSim` toolkit (Rowe et al., 2015) to generate simplified galaxy shears distributed according to a Gaussian random field determined by (Ω_M, σ_8).

For our prior distribution we assume a uniform distribution: $\Omega_M \sim U(0.1, 0.8)$ and $\sigma_8 \sim U(0.5, 1.0)$. Other parameters are fixed to $h = 0.7$, $\Omega_b = 0.045$, $z = 0.7$.

One result that is apparent from Figure 11 is that kernel-NN tuned with the surrogate loss quickly converges to the degeneracy curve $\Omega_M^\alpha \sigma_8$ on which observable data are indistinguishable.

As we in the future analyze more complex simulation mechanisms and higher-dimensional data (with, for example, galaxies divided into time-space bins and measurements from different probes), the dimension of the data and the simulation time will eventually make standard ABC intractable. The above results are very encouraging as our CDE and feature selection

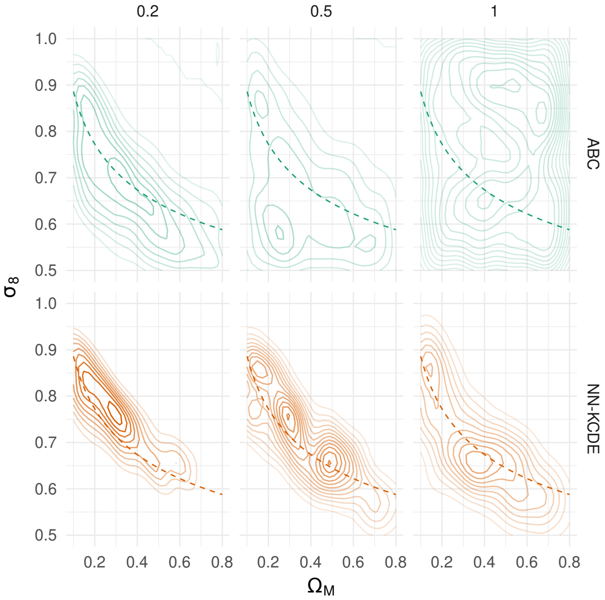


Figure 11: Estimated posteriors of cosmological parameters for weak lensing mock data. The dashed line represents the parameter degeneracy curve on which the data are indistinguishable.

approaches offer flexible ways of estimating posteriors for complex data with limited number of simulations.

5 Conclusions

In this work, we have demonstrated three ways in which our conditional estimation framework can improve upon approximate Bayesian computational methods for next-generation complex data and simulations.

First, realistic simulation models are often such that the computational cost of generating a single sample is large, making lower acceptance ratios unrealistic. For example, recent analyses in cosmological analysis (e.g., Abbott et al., 2016; Hildebrandt et al., 2017) have employed ∼1000 expensive N-body simulations, which altogether can take months to run on thousands of CPUs (Sato et al., 2011; Harnois-Déraps and van Waerbeke, 2015). Often there are also several parameters, which results in prior distributions that are typically not concentrated around the true value of θ. Our work indicates that these are exactly the settings where nonparametric conditional density estimators of $f(\theta|\mathbf{x})$ will lead to better performance than ABC.

Secondly, our CDE framework allows one to compare ABC and related methods in a principled way, making it possible to pick the best method for a given data set without knowing the true posterior. Our approach is based on a surrogate loss function and data splitting. We note that a related cross-validation procedure to choose the tolerance level ϵ in ABC has been proposed by Csillery et al. (2012), albeit using a loss function that is appropriate for point estimation only.

Finally, when dealing with complex models, it is often difficult to know exactly what summary statistics would be appropriate for ABC. Nevertheless, the practitioner can usually make up a list of a large but redundant number of candidate statistics, including statistics generated with automatic methods. As our results show, FlexCode-RF (unlike ABC) is robust to irrelevant statistics. Moreover, FlexCode, in combination with RF for regression, offers a way of evaluating the importance of each summary statistic in estimating the full posterior distribution; hence, these importance scores could be used to choose relevant summary statistics for ABC and any other method used to estimate posteriors.

In brief, there are really two estimation problems in ABC-CDE: The first is that of estimating $f(\theta|\mathbf{x}_o)$; ABC-CDE starts with a rough approximation from an ABC sampler and then directly estimates the conditional density exactly at the point $\mathbf{x} = \mathbf{x}_o$ using a nonparametric conditional density estimator. The second is that of estimating the integrated squared error loss (Eq. 2). Here we propose a surrogate loss that weights all points in the ABC posterior sample equally, but a weighted surrogate loss could potentially return more accurate estimates of the ISE. For example, Figure 10 (left) and Figure 5 from the Appendix show that NN-KCDE perform better than both "Beaumont" and "Blum" for the multimodal example. The current estimated loss however cannot identify a difference in ISE loss between NN-KCDE and "Blum", because of the rapidly shifting posterior in the vicinity of $\mathbf{x} = \mathbf{x}_0$.

In future works we will explore other ways in which one can use non-parametric CDE to directly estimate seemingly intractable likelihoods and posteriors. For instance, the expansion $f(\theta|\mathbf{x}) = \sum_{i \in \mathbb{N}} \beta_i(\mathbf{x})\phi_i(\theta)$ suggests that $(\beta_1(\mathbf{x}), \ldots, \beta_I(\mathbf{x}))$ are good summary statistics to use. Although $\beta_i(\mathbf{x})$'s cannot be directly computed, they may be estimated using the simulations from the model in a similar fashion as Fearnhead and Prangle (2012). Alternatively, one could choose the coefficients $\beta_i(\mathbf{x})$ so as to minimize our surrogate loss using other optimization methods. This will be discussed in a separate paper.

Acknowledgments. We are grateful to Terrence Liu, Brendan McVeigh, and Matt Walker for the NFW simulation code, and Michael Vespe for help with the weak lensing simulations. We would also like to thank Rafael Stern for his insightful comments on the manuscript. This work was partially supported by NSF DMS-1520786 and Fundação de Amparo à Pesquisa do Estado de São Paulo (2017/03363-8).

Links to Nonparametric CDE Software:

- FlexCode: `https://github.com/rizbicki/FlexCoDE`, `https://github.com/tpospisi/flexcode`
- NN-KCDE: `https://github.com/tpospisi/NNKCDE`
- RF-CDE: `https://github.com/tpospisi/rfcde` (Pospisil and Lee, 2018)

References

Abbott, T., F. B. Abdalla, S. Allam, et al. (2016, July). Cosmology from cosmic shear with Dark Energy Survey Science Verification data. *Physical Review D 94*(2), 022001.

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

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

Beaumont, M. A., J. Cornuet, J. Marin, and C. P. Robert (2009). Adaptive approximate bayesian computation. *Biometrika*, asp052.

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

Biau, G., F. Cérou, and A. Guyader (2015). New insights into approximate bayesian computation. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Volume 51, pp. 376–403. Institut Henri Poincaré.

Bickel, P. J., Y. Ritov, and T. M. Stoker (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, 721–741.

Blum, M. G. (2010). Approximate bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association 105*(491), 1178–1187.

Blum, M. G. B. and O. François (2010). Non-linear regression models for approximate bayesian computation. *Statistics and Computing 20*(1), 63–73.

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

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science 16*(3), 199–231.

Cameron, E. and A. Pettitt (2012). Approximate bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society 425*(1), 44–65.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. ACM.

Creel, M. and D. Kristensen (2016). On selection of statistics for approximate bayesian computing (or the method of simulated moments). *Computational Statistics & Data Analysis 100*, 99–114.

Csillery, K., O. Francois, and M. G. B. Blum (2012). abc: an r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution.*

Estoup, A., E. Lombaert, J. Marin, et al. (2012). Estimation of demo-genetic model probabilities with approximate bayesian computation using linear discriminant analysis on summary statistics. *Molecular Ecology Resources 12*(5), 846–855.

Faisal, M., A. Futschik, I. Hussain, and M. Abd-el. Moemen (2016). Choosing summary statistics by least angle regression for approximate bayesian computation. *Journal of Applied Statistics*, 1–12.

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

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

Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis: theory and practice.* Springer Science & Business Media.

Harnois-Déraps, J. and L. van Waerbeke (2015, July). Simulations of weak gravitational lensing - II. Including finite support effects in cosmic shear covariance matrices. *Monthly Notices of the Royal Astronomical Society 450*, 2857–2873.

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The elements of statistical learning: data mining, inference, and prediction.* New York: Springer-Verlag.

Hildebrandt, H., M. Viola, C. Heymans, et al. (2017, February). KiDS-450: cosmological parameter constraints from tomographic weak gravitational lensing. *Monthly Notices of the Royal Astronomical Society 465*, 1454–1498.

Hoekstra, H. and B. Jain (2008). Weak gravitational lensing and its cosmological applications. *Annual Review of Nuclear and Particle Science 58*, 99–123.

Izbicki, R. and A. Lee (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Eletronic Journal of Statistics 11*, 2800–2831.

Lee, A. B. and R. Izbicki (2016). A spectral series approach to high-dimensional nonparametric regression. *Electronic Journal of Statistics 10*(1), 423–463.

Li, J., D. J. Nott, Y. Fan, and S. A. Sisson (2015). Extending approximate bayesian computation methods to high dimensions via gaussian copula. *arXiv preprint arXiv:1504.04093*.

Li, W. and P. Fearnhead (2017). Convergence of regression adjusted approximate bayesian computation. *Biometrika*.

Liu, T. and M. Walker (2018). In preparation.

Mallat, S. (1999). *A wavelet tour of signal processing*. Academic press.

Mandelbaum, R. (2017). Weak lensing for precision cosmology. *arXiv preprint:1710.03235*.

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

Munshi, D., P. Valageas, L. Van Waerbeke, and A. Heavens (2008). Cosmology with weak lensing surveys. *Physics Reports 462*(3), 67–121.

Navarro, J. F. (1996). The structure of cold dark matter halos. In *Symposium-international astronomical union*, Volume 171, pp. 255–258. Cambridge University Press.

Papamakarios, G. and I. Murray (2016). Fast ϵ-free inference of simulation models with bayesian conditional density estimation. *arXiv preprint arXiv:1605.06376*.

Petri, A. (2016). Mocking the weak lensing universe: The lenstools python computing package. *Astronomy and Computing 17*, 73–79.

Pospisil, T. and A. B. Lee (2018). RFCDE: Random forests for conditional density estimation. *arXiv preprint:1804.05753*.

Prangle, D., M. G. B. Blum, G. Popovic, and S. A. Sisson (2014). Diagnostic tools for approximate bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics 56*(4), 309–329.

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B 71*(5), 1009–1030.

Raynal, L., J.-M. Marin, , et al. (2017). ABC random forests for Bayesian parameter inference. *arXiv preprint arXiv:1605.05537*.

Rowe, B., M. Jarvis, R. Mandelbaum, et al. (2015). Galsim: The modular galaxy image simulation toolkit. *Astronomy and Computing 10*, 121–150.

Sato, M., M. Takada, T. Hamana, and T. Matsubara (2011, June). Simulations of Wide-field Weak-lensing Surveys. II. Covariance Matrix of Real-space Correlation Functions. *The Astrophysical Journal 734*, 76.

Sisson, S. A., Y. Fan, and M. M. Tanaka (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences 104*(6), 1760–1765.

Strigari, L. E., C. S. Frenk, and S. D. M. White (2017). Dynamical models for the Sculptor dwarf spheroidal in a λCDM universe. *The Astrophysical Journal 838*(2), 123.

Weyant, A., C. Schafer, and W. M. Wood-Vasey (2013). Likelihood-free cosmological inference with type ia supernovae: Approximate bayesian computation for a complete treatment of uncertainty. *The Astrophysical Journal 764*(2), 116.

A Proofs

A.1 Results on the surrogate loss

Theorem 4. *Assume that, for every $\theta \in \Theta$, $g_\theta(\mathbf{x}) := (\widehat{f}(\theta|\mathbf{x}) - f(\theta|\mathbf{x}))^2$ satisfies the Hölder condition of order β with a constant K_θ*[10] *such that $K_H := \int K_\theta d\theta < \infty$. Then*

$$|L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) - L_{\mathbf{x}_o}(\widehat{f}, f)| = K_H \epsilon^\beta = O(\epsilon^\beta)$$

Proof. First, notice that

$$L_{\mathbf{x}_o}(\widehat{f}, f) = \int g_\theta(\mathbf{x}_o) d\theta \int \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x}, \mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X}, \mathbf{x}_o) < \epsilon)} d\mathbf{x} = \int\int g_\theta(\mathbf{x}_o) \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x}, \mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X}, \mathbf{x}_o) < \epsilon)} d\mathbf{x} d\theta.$$

It follows that

$$\begin{aligned}
|L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) - L_{\mathbf{x}_o}(\widehat{f}, f)| &= \left| \int \left(\int g_\theta(\mathbf{x}) \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x}, \mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X}, \mathbf{x}_o) < \epsilon)} d\mathbf{x} - \int g_\theta(\mathbf{x}_o) \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x}, \mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X}, \mathbf{x}_o) < \epsilon)} d\mathbf{x} \right) d\theta \right| \\
&= \left| \int \left(\int (g_\theta(\mathbf{x}) - g_\theta(\mathbf{x}_o)) \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x}, \mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X}, \mathbf{x}_o) < \epsilon)} d\mathbf{x} \right) d\theta \right| \\
&\leq \int \left(\int |g_\theta(\mathbf{x}) - g_\theta(\mathbf{x}_o)| \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x}, \mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X}, \mathbf{x}_o) < \epsilon)} d\mathbf{x} \right) d\theta \\
&\leq \int \left(\int K_\theta d(\mathbf{x}, \mathbf{x}_o)^\beta \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x}, \mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X}, \mathbf{x}_o) < \epsilon)} d\mathbf{x} \right) d\theta \\
&\leq \int K_\theta \epsilon^\beta \left(\int \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x}, \mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X}, \mathbf{x}_o) < \epsilon)} d\mathbf{x} \right) d\theta \\
&= \epsilon^\beta \int K_\theta 1 d\theta = K_H \epsilon^\beta
\end{aligned}$$

□

[10] That is, there exists a constant K_θ such that for every $\mathbf{x}, \mathbf{y} \in \Re^d$ $|g_\theta(\mathbf{x}) - g_\theta(\mathbf{y})| \leq K_\theta(d(\mathbf{x}, \mathbf{y}))^\beta$.

$$
\begin{aligned}
&L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) = \\
&\qquad \int\!\!\int \widehat{f}^2(\theta|\mathbf{x}) \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x},\mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X},\mathbf{x}_o) < \epsilon)} d\theta d\mathbf{x} - 2\int\!\!\int \widehat{f}(\theta|\mathbf{x}) f(\theta|\mathbf{x}) \frac{f(\mathbf{x})\mathbb{I}(d(\mathbf{x},\mathbf{x}_o) < \epsilon)}{\mathbb{P}(d(\mathbf{X},\mathbf{x}_o) < \epsilon)} d\theta d\mathbf{x} + K_f \\
&\qquad = \mathbb{E}_{\mathbf{X}'}\left[\int \widehat{f}^2(\theta|\mathbf{X}) d\theta\right] - 2\mathbb{E}_{(\theta',\mathbf{X}')}\left[\widehat{f}(\theta|\mathbf{X})\right] + K_f, \qquad (9)
\end{aligned}
$$

Theorem 5. *Let K_f be as in Equation 9. Under the assumptions of Theorem 4,*

$$|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f - L_{\mathbf{x}_o}(\widehat{f}, f)| = O(\epsilon^{\beta}) + O_P(1/\sqrt{B'})$$

Proof. Using the triangle inequality,

$$
\begin{aligned}
|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f - L_{\mathbf{x}_o}(\widehat{f}, f)| &\leq |\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f - L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f)| + |L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) - L_{\mathbf{x}_o}(\widehat{f}, f)| \\
&= O(\epsilon^{\beta}) + O_P(1/\sqrt{B'}),
\end{aligned}
$$

where the last inequality follows from Theorem 4 and the fact that $\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f$ is an average of B' iid random variables. □

Lemma 1. *Assume there exists M such that $|\widehat{f}(\theta|\mathbf{x})| \leq M$ for every $\mathbf{x}$ and θ. Then*

$$\mathbb{P}\left(|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f - L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f)| \geq \nu\right) \leq 2e^{-\frac{B'\nu^2}{2(M^2+2M)^2}}$$

Proof. Notice that

$$\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f - L^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) = \frac{1}{B'}\sum_{k=1}^{B'} W_k - \mathbb{E}[W_1],$$

where $W_k = \int \widehat{f}^2(\theta|\mathbf{X}'_k)d\theta - 2\widehat{f}(\Theta'_k|\mathbf{X}'_k)$, with $W_1, \ldots, W_{B'}$ iid. The conclusion follows from Hoeffding's inequality and the fact that $|W_k| \leq |\int \widehat{f}^2(\theta|\mathbf{X}'_k)d\theta - 2\widehat{f}(\Theta'_k|\mathbf{X}'_k)| \leq M^2 + 2M$. □

Lemma 2. *Under the assumptions of Lemma 1 and if $g_\theta(\mathbf{x}) := (\widehat{f}(\theta|\mathbf{x}) - f(\theta|\mathbf{x}))^2$ satisfies the Hölder condition of order β with constants K_θ such that $K_H := \int K_\theta d\theta < \infty$,*

$$\mathbb{P}\left(|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f - L_{\mathbf{x}_o}(\widehat{f}, f)| \geq K_H\epsilon^{\beta} + \nu\right) \leq 2e^{-\frac{B'\nu^2}{2(M^2+2M)^2}},$$

Proof. Notice that

$$\begin{aligned}
&|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)+K_f-L_{\mathbf{x}_o}(\widehat{f},f)|-K_H\epsilon^{\beta}\\
&=|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)+K_f-L^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)+L^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)-L_{\mathbf{x}_o}(\widehat{f},f)|-K_H\epsilon^{\beta}\\
&\leq|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)+K_f-L^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)|+|L^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)-L_{\mathbf{x}_o}(\widehat{f},f)|-K_H\epsilon^{\beta}\\
&\leq|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)+K_f-L^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)|,
\end{aligned}$$

where the last line follows from Theorem 4. It follows that

$$|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)+K_f-L_{\mathbf{x}_o}(\widehat{f},f)|\geq K_H\epsilon^{\beta}+\nu\Rightarrow|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)+K_f-L^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)|\geq\nu.$$

The conclusion follows from Lemma 1. □

Theorem 6. *Let $\mathcal{F}=\{\widehat{f}_1,\ldots,\widehat{f}_m\}$ be a set of estimators of $f(\theta|\mathbf{x}_o)$. Assume there exists M such that $|\widehat{f}_i(\theta|\mathbf{x})|\leq M$ for every $\mathbf{x}$, θ, and $i=1,\ldots,m$.* [11] *Moroever, assume that for every $\theta\in\Theta$, $g_{i,\theta}(\mathbf{x}):=(\widehat{f}_i(\theta|\mathbf{x})-f(\theta|\mathbf{x}))^2$ satisfies the Hölder condition of order β with constants K_θ such that $K_H:=\int K_\theta d\theta<\infty$. Then,*

$$\mathbb{P}\left(\max_{\widehat{f}\in\mathcal{F}}|\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)+K_f-L_{\mathbf{x}_o}(\widehat{f},f)|\geq K_\epsilon\epsilon^{\beta}+\nu\right)\leq 2me^{-\frac{B'\nu^2}{2(M^2+2M)^2}}.$$

Proof. The theorem follows from Lemma 2 and the union bound. □

Corollary 2. *Let $\widehat{f}^*:=\arg\min_{\widehat{f}\in\mathcal{F}}\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f},f)$ be the best estimator in $\mathcal{F}$ according to the estimated surrogate loss, and let $f^*=\arg\min_{\widehat{f}\in\mathcal{F}}L_{\mathbf{x}_o}(\widehat{f},f)$ be the best estimator in $\mathcal{F}$ according to the true loss. Then, under the assumptions from Theorem 6, with probability at least $1-2me^{-\frac{B'\nu^2}{2(M^2+2M)^2}}$,*

$$L_{\mathbf{x}_o}(\widehat{f}^*,f)\leq L_{\mathbf{x}_o}(f^*,f)+2(K_H\epsilon^{\beta}+\nu).$$

[11] Such assumptions hold if the $\widehat{f}_i$'s are obtained via FlexCode with bounded basis functions (e.g., Fourier basis) or a kernel density estimator on the ABC samples.

Proof. From Theorem 6, with probability at least $1 - 2me^{-\frac{B'\nu^2}{2(M^2+2M)^2}}$

$$\begin{aligned} L_{\mathbf{x}_o}(\widehat{f}^*, f) - L_{\mathbf{x}_o}(f^*, f) = & \\ & L_{\mathbf{x}_o}(\widehat{f}^*, f) - (\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}^*, f) + K_f) \\ & + (\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}^*, f) + K_f) - (\widehat{L}^{\epsilon}_{\mathbf{x}_o}(f^*, f) + K_f) \\ & + (\widehat{L}^{\epsilon}_{\mathbf{x}_o}(f^*, f) + K_f) - L_{\mathbf{x}_o}(f^*, f) \\ & \leq 2(K_H \epsilon^{\beta} + \nu), \end{aligned}$$

where the inequality follows from the fact that, by definition, $(\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}^*, f) + K_f) - (\widehat{L}^{\epsilon}_{\mathbf{x}_o}(f^*, f) + K_f) < 0$ and $L_{\mathbf{x}_o}(\widehat{f}, f) - (\widehat{L}^{\epsilon}_{\mathbf{x}_o}(\widehat{f}, f) + K_f) \leq K_H \epsilon^{\beta} + \nu$ for every $\widehat{f} \in \mathcal{F}$. □

A.2 Results on summary statistics selection

Assumption 2 (Smoothness in θ direction)**.** *$\forall \mathbf{x} \in \mathcal{X}$, $f(\theta|\mathbf{x}) \in W_{\phi}(s_{\mathbf{x}}, c_{\mathbf{x}})$, where $f(\theta|\mathbf{x})$ is viewed as a function of θ, and $s_{\mathbf{x}}$ and $c_{\mathbf{x}}$ are such that $\inf_{\mathbf{x}} s_{\mathbf{x}} \overset{def}{=} \beta > \frac{1}{2}$ and $\int c^2_{\mathbf{x}} d\mathbf{x} < \infty$.*

Lemma 3. *Let $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{x}' = (x_1, \ldots, x'_j, \ldots, x_d)$. Then, for every $\mathbf{x}$ and x_j, $g_{\mathbf{x},x_j}(\theta) := f(\theta|\mathbf{x}) - f(\theta|\mathbf{x}'_j) \in W_{\phi}(\beta, c^2_{\mathbf{x}} + c^2_{\mathbf{x}'_j} + 2\sqrt{c^2_{\mathbf{x}} c^2_{\mathbf{x}'_j}})$.*

Proof. First we expand $f(\theta|\mathbf{x})$ and $f(\theta|\mathbf{x}'_j)$ in the basis $(\phi_i)_i$. We have that

$$g_{\mathbf{x},x_j}(\theta) = f(\theta|\mathbf{x}) - f(\theta|\mathbf{x}'_j) = \sum_{i \geq 0} (\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j))\phi_i(\theta).$$

Now, using Cauchy-Schwarz inequality, the expansion coefficients satisfy

$$\begin{aligned}
&\sum_{i\geq 1} i^{2\beta}(\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j))^2 \\
&\qquad = \sum_{i\geq 1} i^{2\beta}(\beta_i(\mathbf{x}))^2 + \sum_{i\geq 1} i^{2\beta}(\beta_i(\mathbf{x}'_j))^2 + 2\sum_{i\geq 1} i^{2\beta}\beta_i(\mathbf{x})\beta_i(\mathbf{x}'_j) \\
&\qquad \leq c_{\mathbf{x}}^2 + c_{\mathbf{x}'_j}^2 + 2\sqrt{\left(\sum_{i\geq 1} i^{2\beta}(\beta_i(\mathbf{x}))^2\right)\left(\sum_{i\geq 1} i^{2\beta}(\beta_i(\mathbf{x}'_j))^2\right)} \\
&\qquad \leq c_{\mathbf{x}}^2 + c_{\mathbf{x}'_j}^2 + 2\sqrt{c_{\mathbf{x}}^2 c_{\mathbf{x}'_j}^2},
\end{aligned}$$

where the last inequality follows from Assumption 2. □

Proposition 2. *Under Assumption 2,*

$$r_j = \sum_{i=1}^{I} r_{i,j} + O\left(I^{-2\beta}\right)$$

Proof. Because $\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j)$ are the expansion coefficients of $f(\theta|\mathbf{x}) - f(\theta|\mathbf{x}'_j)$ on the basis $(\phi_i)_i$, it follows from Lemma 3 (see appendix) that

$$\sum_{i\geq I} I^{2\beta}\left(\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j)\right)^2 \leq \sum_{i\geq I} i^{2\beta}\left(\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j)\right)^2 \leq c_{\mathbf{x}}^2 + c_{\mathbf{x}'_j}^2 + 2\sqrt{c_{\mathbf{x}}^2 c_{\mathbf{x}'_j}^2}.$$

Hence,

$$\sum_{i\geq I} r_{i,j} = \sum_{i\geq I} \int\int \left(\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j)\right)^2 d\mathbf{x}dx'_j \leq \frac{K}{I^{2\beta}} = O(I^{-2\beta}). \tag{10}$$

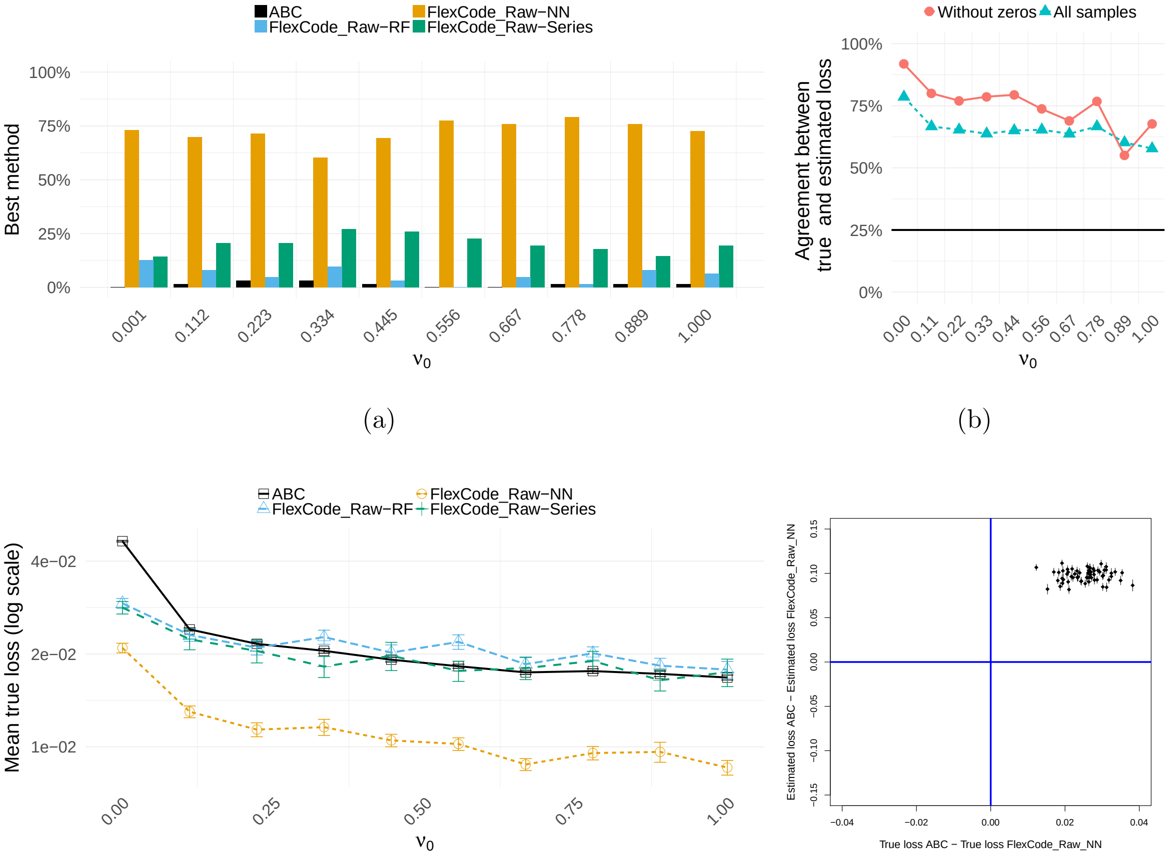
Because $f(\theta|\mathbf{x}) - f(\theta|\mathbf{x}'_j) = \sum_{i\geq 0}(\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j))\phi_i(\theta)$ and the basis $(\phi_i)_i$ is orthonormal, we have that

$$r_j = \int\int \sum_{i\geq 0}(\beta_i(\mathbf{x}) - \beta_i(\mathbf{x}'_j))^2 d\mathbf{x}dx'_j = \sum_{i\geq 0} r_{i,j}. \tag{11}$$

The final result follows from putting Equations 10 and 11 together. □

B Mean of a Gaussian with unknown precision

In this section, we repeat the experiments of Section 3.1 of the paper, but in the case $X_1, \ldots, X_{20}|(\mu, \tau) \overset{iid}{\sim} \mathrm{N}(\mu, 1/\tau)$, $(\mu, \tau) \sim$ Normal-Gamma$(\mu_0, \nu_0, \alpha_0, \beta_0)$. We set $\mu_0 = 0$, $\alpha_0 = 2$, $\beta_0 = 50$ and repeat the experiments for ν_0 in an equally spaced grid with ten values between 0.001 and 1.



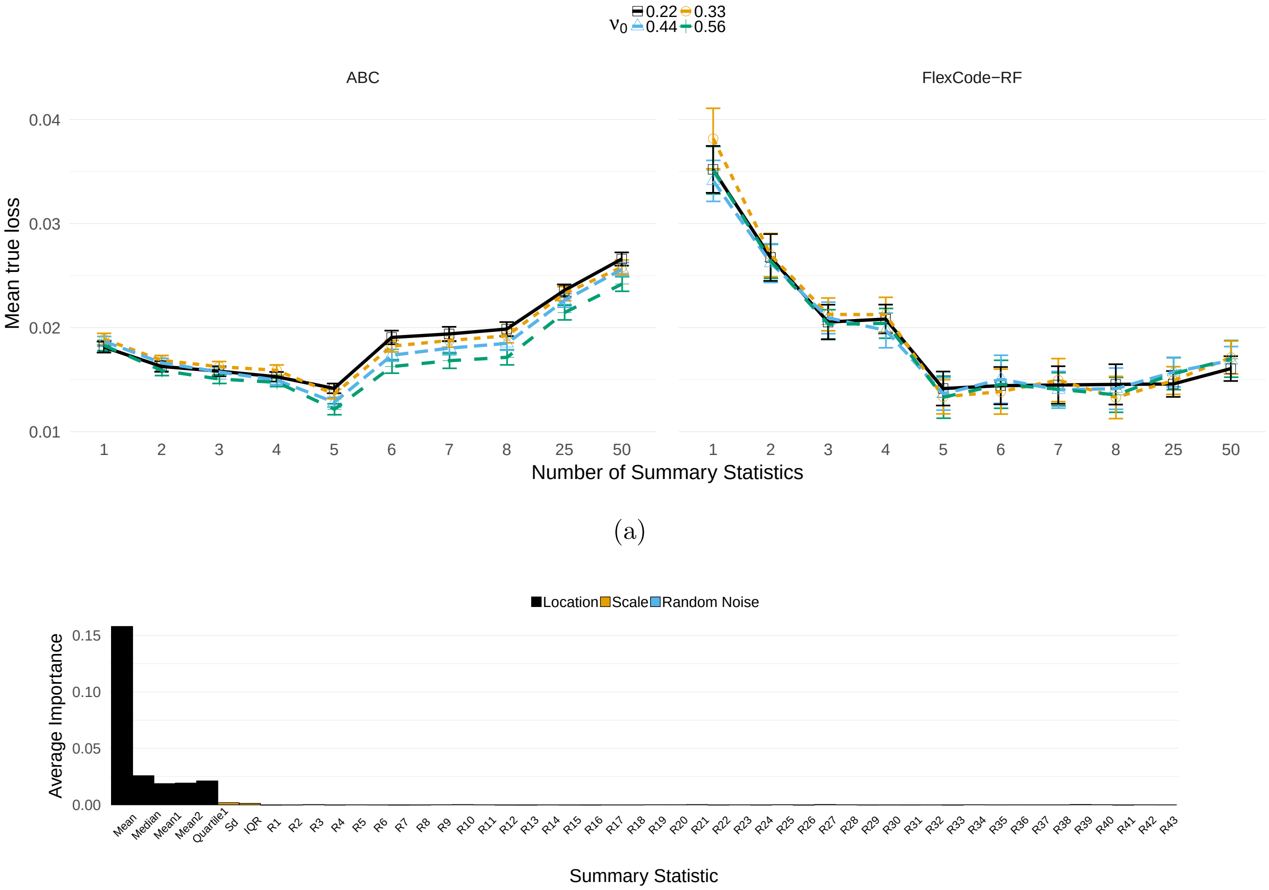
(a)

(b)

(c)

(d)

Figure 12: CDE and method selection results for scenario 3 (mean of a Gaussian with unknown precision). *Left:* Panels (a) and (c) show that the NN version of FlexCode yield better estimates of the posterior density $f(\theta|\mathbf{x}_o)$ than the competing methods. *Right:* Panels (b) and (d) indicate that one by estimating the surrogate loss function can tell from the data which method is better for the problem at hand. The horizontal line in panel (b) represents the behavior of a random selection.



(a)

(b)

Figure 13: Summary statistic selection for scenario 3 (mean of a Gaussian with unknown precision). Panel (a) shows the performance of ABC and FlexCode-RF using different sets of summary statistics. The ABC estimates of the posteriors rapidly deteriorate when adding other statistics than the location statistics 1-5 (*top left*), whereas nuisance statistics do not decrease the performance of FlexCode-RF significantly (*top right*). Panel (b), furthermore, shows that FlexCode-RF identifies the location statistics (entries 1-5) as key variables and assigns them a high average importance score.

C Additional details on the Galaxy's Dark Matter Density Profile Model

Under the NFW model the joint likelihood for the specific angular momentum J and specific energy E factorizes independently

$$f(E, J) \propto g(J)h(E) \tag{12}$$

with

$$g(J) = \begin{cases} [1 + (J/J_\beta)^{-b}]^{-1} & b \leq 0 \\ 1 + (J/J_\beta)^{b} & b > 0 \end{cases} \tag{13}$$

and

$$h(E) = \begin{cases} E^{\alpha}(E^{q} + E_c^{q})^{d/q}(\Phi_{lim} - E)^{e} & E < \Phi_{lim} \\ 0 & E \geq \Phi_{lim} \end{cases} \tag{14}$$

We can relate E and J to the observed values of position r and velocity v as follows

$$\begin{aligned} E &= \frac{1}{2}v^2 + \Phi_s(1 - \frac{\log(1 + \frac{r}{r_s})}{\frac{r}{r_s}}) \\ J &= vr\sin(\theta) \end{aligned}$$

We set the following constants at commonly accepted values

Parameter	Value
α	2.0
d	-5.3
e	2.5
v_{max}	21
r_{max}	1.5
ϕ_s	$(v_{max}/0.465)^2$
r_s	$r_{max}/2.16$
r_{lim}	1.5
ϕ_{lim}	$\phi_s(1 - \frac{\log(1+\frac{r_{lim}}{r_s})}{\frac{r_{lim}}{r_s}}$
b	-9.0
q	6.9
J_β	0.086

and focus only on estimating E_c.

Figure 14 displays examples of estimated posterior; each plot consists of an observed sample generated using a different true of E_c.

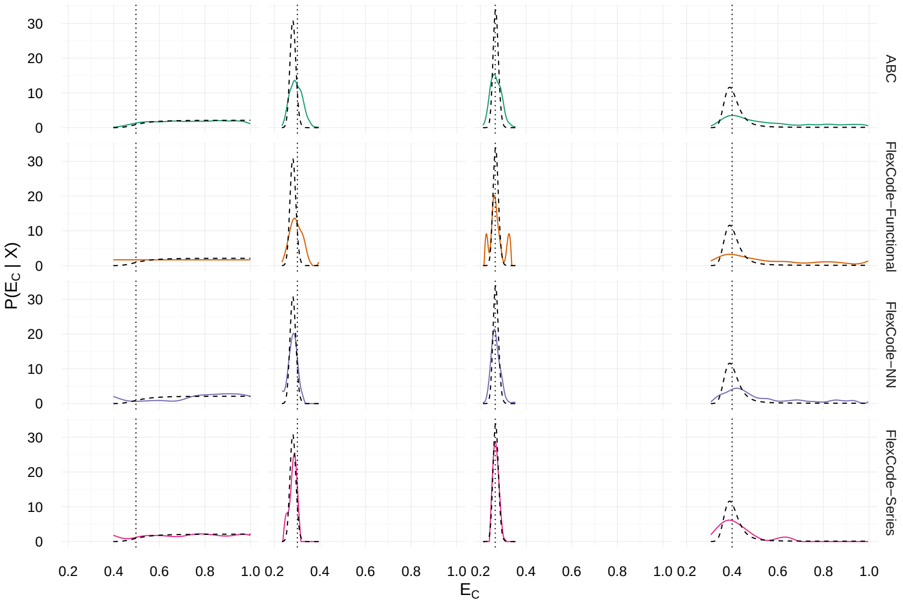


Figure 14: Sample posterior densities for simulated galaxy data; the dashed curve is the true posterior, and the vertical line indicates the true parameter value.

D Two-Dimensional Normal.

We can extend the normal example of Section 4 of the paper to multiple dimensions with similar results. Given a two-dimensional multivariate normal with fixed covariance $\Sigma_X = I_2$, we put a normal conjugate prior on the mean $\mu \sim N(\mu_0 = 0, \Sigma_0 = I_2)$. This results in the true posterior

$$\mu \mid X \sim N(\mu_n, \Sigma_n),$$

where

$$\begin{aligned}\mu_n &= \Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma_X)^{-1}\bar{x} + \frac{1}{n}\Sigma_X(\Sigma_0 + \frac{1}{n}\Sigma_X)^{-1}\mu_0 \\ \Sigma_n &= \frac{1}{n}\Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma_X)^{-1}\Sigma_X\end{aligned}$$

As before we use the sufficient statistic of the sample mean as our statistic and the Euclidean norm as the distance function.

Figure 15 shows density estimates for ABC and kernel-NN for different values of the acceptance rate. At higher acceptance rates the ABC density estimate performs poorly, reflecting the prior distribution rather than the posterior. Eventually, with a suitably low acceptance rate the ABC density approaches the posterior. Once again kernel-NN achieves similar performance as standard ABC with 100000 simulations (at acceptance ratio 0.01) but using only 1000 ABC realizations (at acceptance ratio 1).

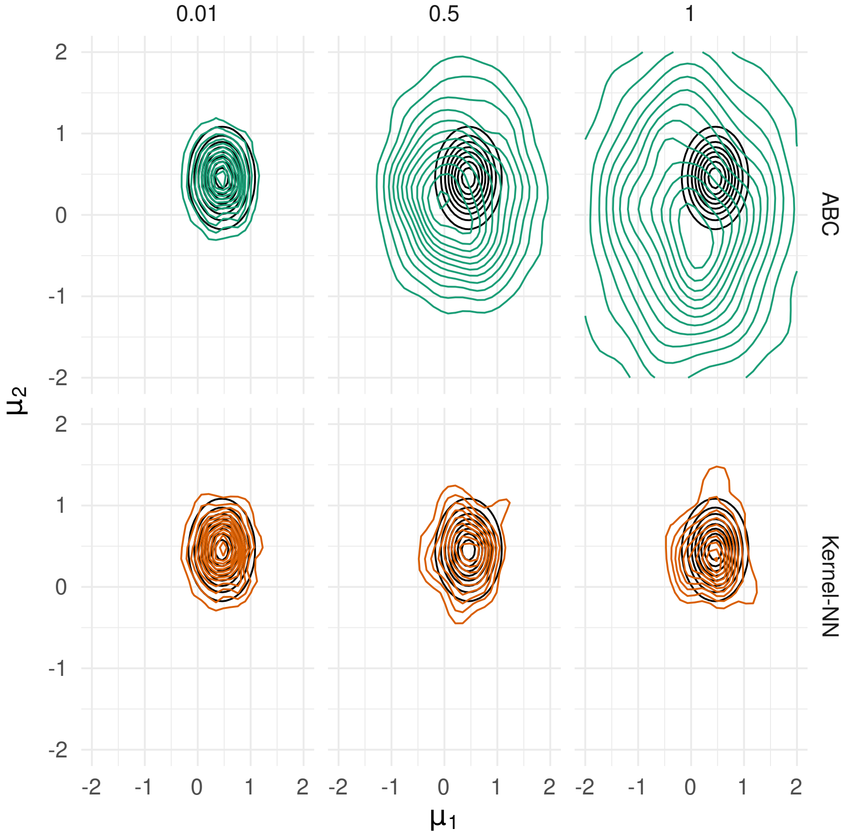


Figure 15: Contours for density estimates of two-dimensional normal posterior; black lines are the contours of the true posterior.

E Fast Implementation of NN-KCDE

NN-KCDE (or nearest-neighbors kernel CDE) is the usual kernel density estimate using only the points closest in covariate space to the target point $\mathbf{x}$:

$$\widehat{f}_{\text{nn}}(\theta \mid \mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} K_h(\rho(\theta, \theta_{s_i(\mathbf{x})})), \tag{15}$$

where $s_i(\mathbf{x})$ represents the index of the ith nearest neighbor to $\mathbf{x}$. As mentioned in Section 4.1, NN-KCDE has a close connection with ABC in that, for every choice of k for a data set, there is a choice of ϵ for the accept-reject ABC algorithm that produces an equivalent estimate of the posterior. With the CDE loss, we can choose k to minimize the loss and improve upon the naive pre-selected ϵ approach. For large ϵ we expect k to be small to avoid bias in the posterior estimate. As ϵ shrinks, larger values of k will be selected to reduce the variance of the estimate.

To make this model selection procedure computationally feasible, we need to be able to efficiently calculate the surrogate loss function. We examine the two terms (Section 2.2, Equation 5) separately: The second term

$$\sum_i \widehat{f}(\theta_i \mid \mathbf{x}_i)$$

poses no difficulties as we simply plug in the kernel density estimate. The first term

$$\sum_i \int \widehat{f}^2(\theta \mid \mathbf{x}_i) dz$$

is more difficult. Numerically integrating this integral is infeasible especially as the number of validation samples increases.

Fortunately there is an analytic solution. We can express the integral in terms of convolutions of the kernel function:

$$\int \widehat{f}^2(\theta \mid \mathbf{x}) dz = \frac{1}{k^2 h} \sum_{i \in N_k(\mathbf{x})} \sum_{j \in N_k(\mathbf{x})} \int K(t) K\left(t - \frac{d_{i,j}}{h}\right) dt$$

with $d_{i,j}$ representing the pairwise distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$. In the Gaussian case we have the analytic solution

$$\int K(t)K(t-d)dt = \frac{1}{2\sqrt{\pi}} \exp\left(\frac{-d^2}{4}\right)$$

For other kernels we can work out the analytic solution as well, or, if that proves intractable, we can approximate the function using numerical integration.

For both terms we have nested calculations, in that we can reuse computations for $k = k_1 < k_2$ when calculating the estimated loss for $k = k_2$. In this way, there is little additional computational time in considering all settings for k as opposed to trying only a large value of k.

An implementation of this method is available at https://github.com/tpospisi/NNKCDE.

F Densities for Regression Adjustment Example

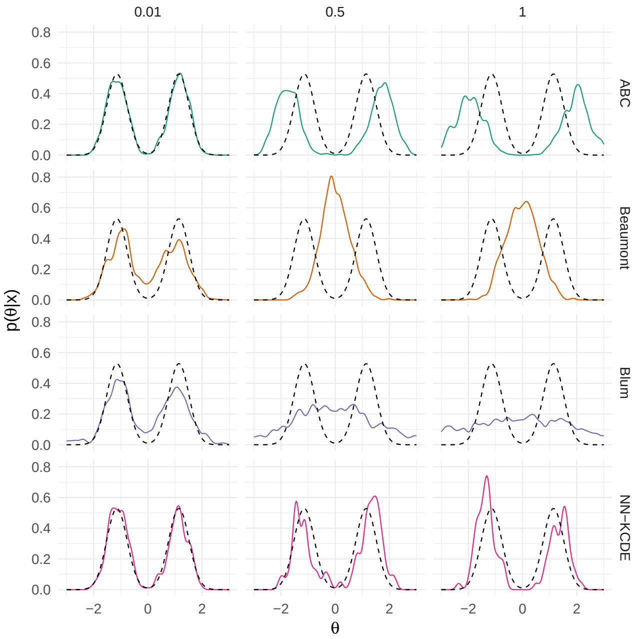


Figure 16: The regression-adjustment methods adjust for the change in the distribution of $\theta \mid \mathbf{x}$ around $\mathbf{x}_{\text{obs}}$ by shifting and rescaling the sample by the conditional mean $m(\mathbf{x})$ and the conditional variance $\sigma(\mathbf{x})$, respectively. However, the change in the distribution from unimodal to multimodal cannot be expressed by shifting or rescaling which results in misleading posteriors for the regression-adjustment methods for larger values of the ABC tolerance while NN-KCDE performs well.