Inference in generative models using the Wasserstein distance

Espen Bernton*, Pierre E. Jacob*, Mathieu Gerber†, Christian P. Robert‡

Abstract

A growing range of generative statistical models are such the numerical evaluation of their likelihood functions is intractable. Approximate Bayesian computation and indirect inference have become popular approaches to overcome this issue, simulating synthetic data given parameters and comparing summaries of these simulations with the corresponding observed values. We propose to avoid these summaries and the ensuing loss of information through the use of Wasserstein distances between empirical distributions of observed and synthetic data. We describe how the approach can be used in the setting of dependent data such as time series, and how approximations of the Wasserstein distance allow the method to scale to large data sets. In particular, we propose a new approximation to the optimal assignment problem using the Hilbert space-filling curve. We provide an in-depth theoretical study, including consistency in the number of simulated data sets for a fixed number of observations and posterior concentration rates. The approach is illustrated on various examples, including a multivariate g-and-k distribution, a toggle switch model from systems biology, a queueing model, and a Lévy-driven stochastic volatility model.

1 Introduction

The likelihood function plays a central role in modern statistics. However, for many models of interest, the likelihood cannot be numerically evaluated. It might be possible to generate synthetic data sets given parameters, in which case the model is said to be generative. A popular approach to Bayesian inference in generative models is approximate Bayesian computation (ABC, Beaumont et al., 2002; Marin et al., 2012). Approximate Bayesian computation relies on simulating parameters and synthetic data sets. The parameters are then kept if the associated synthetic data sets are close enough to the observed data set, forming an approximation of the posterior distribution. Measures of similarity between data sets are often based on summary statistics, such as sample moments. In other words, data sets are considered close if some distance between their summaries is small. If, instead, one were to define a point estimator by minimizing the distance between summaries as a function of the parameters, the resulting method would be an example of indirect inference (Gouriéroux et al., 1993). Connections between these two approaches to parameter inference are discussed in Forneron and Ng (2015). The resulting ABC and indirect inference estimators have proven extremely useful, but can lead to systematic losses of information compared to the posterior distribution and maximum likelihood estimator respectively, due to the summaries being non-sufficient in many cases of interest.

We propose to view data sets as empirical distributions and to use the Wasserstein distance between synthetic and observed data sets. The Wasserstein distance, also called the Gini, Mallows or Kantorovich

*Department of Statistics, Harvard University, USA, ebernton@g.harvard.edu and pjacob@g.harvard.edu
†School of Mathematics, University of Bristol, UK
‡CEREMADE, Université Paris-Dauphine, PSL Research University, France, and Department of Statistics, University of Warwick, UK

distance, defines a metric on the space of probability distributions, and has been increasingly popular in statistics and machine learning (e.g. Cuturi, 2013; Srivastava et al., 2015; Montavon et al., 2016; Ramdas et al., 2017; Arjovsky et al., 2017; Wu and Tabak, 2017; Sommerfeld and Munk, 2017), due to its appealing computational and statistical properties. We will show that the resulting ABC posterior, which we term the Wasserstein ABC (WABC) posterior, can approximate the standard posterior distribution arbitrarily well in the limit of the number of simulations from the model, while by-passing the choice of summaries completely. Furthermore, we provide concentration rates as the number of observations goes to infinity, highlighting the impact of the dimension of the observation space and the effect of model misspecification. The WABC posterior is a particular case of a coarsened posterior, and our results are complementary to those of Miller and Dunson (2015). We will also discuss the consistency and asymptotic distributions of several point estimators derived by minimizing the Wasserstein distance between the empirical distribution and the model, or approximations thereof, extending the work of Bassetti et al. (2006).

Viewing data sets as empirical distributions raises issues in the case of dependent data. We develop two strategies to deal with time series, noting that spatial data could be treated similarly. In the first approach, which we term curve matching, each data point is augmented with the time at which it was observed. A new ground metric is defined on this extended observation space, which in turn allows for the definition of a Wasserstein distance between time series. The second approach involves transforming the time series such that its empirical distribution contains enough information for parameter estimation. We refer to such transformations as reconstructions and discuss two generic choices.

The calculation of Wasserstein distances is fast for empirical distributions in one dimension. For multivariate data sets, we can leverage the rich literature on the computation and approximation of Wasserstein distances. We propose a new distance, termed the Hilbert distance, based on the Hilbert space-filling curve (Sagan, 1994). The proposed distance can be computed orders of magnitude faster than the exact Wasserstein distance, and we provide theoretical support for its use in ABC and minimum distance estimation settings.

Our contributions are structured as follows: the proposed approaches to point estimation and Bayesian inference in generative models using the Wasserstein distance are described in Sections 2 and 3 respectively, methods to handle time series are proposed in Section 4, a theoretical study of the proposed minimum Wasserstein estimators and Wasserstein ABC posterior is detailed in Section 5, computational challenges of calculating Wasserstein distances and new solutions are described in Section 6, and numerical illustrations in Section 7. The experiments include a multivariate quantile g-and-k distribution, a toggle switch model from system biology, a M/G/1 queueing model, and a Lévy-driven stochastic volatility model. The code and tutorials are available on GitHub at github.com/pierrejacob/winference. The supplementary materials include additional theoretical results and details on the computational aspects, as referenced in the present article.

1.1 Setting and notation

Throughout this work we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with associated expectation operator $\mathbb{E}$, on which all the random variables are defined. The set of probability measures on a space $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. The data take values in $\mathcal{Y}$, a subset of $\mathbb{R}^{d_y}$ for $d_y \in \mathbb{N}$. We observe $n \in \mathbb{N}$ data points, $y_{1:n} = y_1, \ldots, y_n$, that are distributed according to $\mu_\star^{(n)} \in \mathcal{P}(\mathcal{Y}^n)$. For stationary data-generating processes, we denote by $\mu_\star$ the marginal distribution of y_1. The empirical distribution of the data $y_{1:n}$ is $\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{y_i}$, where δ_y is the Dirac distribution with mass on $y \in \mathcal{Y}$.

A model refers to a collection of distributions on $\mathcal{Y}$, denoted by $\mathcal{M} = \{\mu_\theta : \theta \in \mathcal{H}\} \subset \mathcal{P}(\mathcal{Y})$, where $\mathcal{H} \subset \mathbb{R}^{d_\theta}$ is the parameter space, endowed with a distance $\rho_\mathcal{H}$ and of dimension $d_\theta \in \mathbb{N}$. We assume that models generate i.i.d. observations from μ_θ, with joint distribution $\mu_\theta^{(n)}$, until Section 4 where we elaborate on non-i.i.d. models. A model is well-specified if there exists $\theta_\star \in \mathcal{H}$ such that $\mu_\star = \mu_{\theta_\star}$; otherwise it is misspecified. Parameters are identifiable if $\theta = \theta'$ is implied by $\mu_\theta = \mu_{\theta'}$. A sequence of measures μ_n converging weakly to μ is denoted by $\mu_n \Rightarrow \mu$.

We consider parameter inference for purely generative models: it is possible to generate observations $z_{1:n}$ from $\mu_\theta^{(n)}$, for all $\theta \in \mathcal{H}$, but it is not possible to numerically evaluate the associated likelihood. In some cases, $z_{1:n}$ is obtained as $g_n(u, \theta)$, where g_n is a known deterministic function and u some known fixed-dimensional random variable independent of θ. Some methods require access to g_n and u (e.g. Gouriéroux et al., 1993; Prangle et al., 2017; Graham and Storkey, 2017). We will be explicit about where assumptions on the data-generating process are needed.

1.2 Wasserstein distance

We propose to use the Wasserstein distance as a discrepancy between pairs of data sets, viewing each data set as an empirical distribution. As described in the following sections, this new perspective enables a variety of methodological and theoretical improvements over the standard ABC and minimum distance approaches.

Let ρ be a distance on the observation space $\mathcal{Y}$, referred to as the ground distance. Let $\mathcal{P}_p(\mathcal{Y})$ with $p \geq 1$ (e.g. $p = 1$ or 2) be the set of distributions $\mu \in \mathcal{P}(\mathcal{Y})$ with finite p-th moment: there exists $y_0 \in \mathcal{Y}$ such that $\int_\mathcal{Y} \rho(y, y_0)^p d\mu(y) < \infty$. The space $\mathcal{P}_p(\mathcal{Y})$ is referred to as the p-Wasserstein space of distributions on $\mathcal{Y}$ (Villani, 2008). The p-Wasserstein distance is a finite metric on $\mathcal{P}_p(\mathcal{Y})$, defined by the transport problem

$$\mathfrak{W}_p(\mu, \nu)^p = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{Y}\times\mathcal{Y}} \rho(x, y)^p d\gamma(x, y), \tag{1}$$

where $\Gamma(\mu, \nu)$ is the set of probability measures on $\mathcal{Y} \times \mathcal{Y}$ with marginals μ and ν respectively; see the notes in Chapter 6 of Villani (2008) for a brief history of this distance and its central role in optimal transport. We also write $\mathfrak{W}_p(y_{1:n}, z_{1:m})$ for $\mathfrak{W}_p(\hat{\mu}_n, \hat{\nu}_m)$, where $\hat{\mu}_n$ and $\hat{\nu}_m$ stand for the empirical distributions $n^{-1}\sum_{i=1}^n \delta_{y_i}$ and $m^{-1}\sum_{i=1}^m \delta_{z_i}$. In particular, the Wasserstein distance between two empirical distributions with unweighted atoms takes the form

$$\mathfrak{W}_p(y_{1:n}, z_{1:m})^p = \inf_{\gamma \in \Gamma_{n,m}} \sum_{i=1}^n \sum_{j=1}^m \rho(y_i, z_j)^p \gamma_{ij} \tag{2}$$

where $\Gamma_{n,m}$ is the set of $n \times m$ matrices with non-negative entries, columns summing to m^{-1}, and rows summing to n^{-1}. An important special case is when $n = m$, for which it is known (see e.g. the introductory chapter in Villani, 2003) that the solution to the optimization problem $\gamma^\star$ corresponds to an assignment matrix, with only one non-zero entry per row and column, equal to n^{-1}. The Wasserstein distance can thus be represented as

$$\mathfrak{W}_p(y_{1:n}, z_{1:n})^p = \inf_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^n \rho(y_i, z_{\sigma(i)})^p, \tag{3}$$

where $\mathcal{S}_n$ is the set of permutations of $\{1, \ldots, n\}$. Computing the Wasserstein distance between two samples of the same size can therefore be thought of as an optimal matching problem; see Section 6.

2 Point estimation with the Wasserstein distance

2.1 Minimum distance estimation

Minimum distance estimation (MDE) refers to the idea of minimizing, over the parameter $\theta \in \mathcal{H}$, a distance between the empirical distribution $\hat{\mu}_n$ and the model distribution μ_θ (Wolfowitz, 1957; Basu et al., 2011). In these broad terms, it encompasses the spirit of various statistical paradigms: for instance, the maximum likelihood approach asymptotically minimizes the Kullback-Leibler (KL) divergence between $\mu_\star$ and μ_θ, defined as $\mathrm{KL}(\mu_\star|\mu_\theta) = \int \log(d\mu_\star/d\mu_\theta)d\mu_\star$. The empirical likelihood method minimizes the KL divergence between the empirical distribution and a model supported on the observed data under moment conditions (Owen, 2001). The generalized method of moments consists in minimizing a weighted Euclidean distance between moments of $\hat{\mu}_n$ and μ_θ (Hansen, 1982). Any choice of distance, or pseudo-distance measuring the similarity between two distributions, yields an associated minimum distance estimator. Denoting by $\mathfrak{D}$ a distance or divergence on $\mathcal{P}(\mathcal{Y})$, the associated minimum distance estimator can be defined as

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathfrak{D}(\hat{\mu}_n, \mu_\theta). \tag{4}$$

This raises multiple statistical and computational questions. First, the distance $\mathfrak{D}$ needs to be a meaningful notion of similarity between distributions, including empirical distributions with unequal discrete supports. This precludes some distances, such as the total variation distance. Statistically, the estimator $\hat{\theta}_n$ should preferably satisfy some desirable properties, under conditions on $\mathfrak{D}$, the data-generating distribution $\mu_\star$, and the model $\mathcal{M}$. Particularly important properties include existence and measurability of $\hat{\theta}_n$, uniqueness and consistency when $n \to \infty$. Further interesting aspects include rates of convergence, asymptotic distributions, and robustness to outliers.

The distance $\mathfrak{D}$ needs to be computable, at least up to a certain accuracy, so that we can realistically envision the above optimization program. Since the data might be multivariate ($d_y > 1$), some familiar distances such as the Kolmogorov–Smirnov distance might prove computationally inconvenient. Finally, in the context of purely generative models, it will often be more convenient to consider the alternative estimator

$$\hat{\theta}_{n,m} = \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathbb{E}\left[\mathfrak{D}(y_{1:n}, z_{1:m})\right]. \tag{5}$$

where the expectation is taken over distribution of the sample $z_{1:m} \sim \mu_\theta^{(m)}$. When n is fixed and m is large, or when $n = m$ and n is large, we expect the expectation to be close to $\mathfrak{D}(\hat{\mu}_n, \mu_\theta)$, and the two estimators to have similar properties.

2.2 Minimum Wasserstein estimators

By plugging $\mathfrak{W}_p$ in place of $\mathfrak{D}$ in Eqs. (4) and (5), we obtain the minimum Wasserstein estimator (MWE) and minimum expected Wasserstein estimator (MEWE) of order p, denoted $\hat{\theta}_n$ and $\hat{\theta}_{n,m}$ respectively. Some properties of the MWE have been studied in Bassetti et al. (2006), for well-specified models and i.i.d. data; we propose new results in Section 5. Intuitively, under some conditions we can expect $\hat{\mu}_n$ to converge to $\mu_\star$, in the sense that $\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \to 0$. Consequently, the minimum of $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$ might converge to the minimum of $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$, denoted by $\theta_\star$, assuming its existence and unicity. In the well-specified case, $\theta_\star$ coincides with the data-generating parameter. In the misspecified case, $\theta_\star$ is typically different from the limit of the maximum likelihood estimator (MLE), which is the minimizer of $\mathrm{KL}(\mu_\star|\mu_\theta)$. While the KL

divergence is a central notion in information theory, and is defined irrespective of the metric on the data space $\mathcal{Y}$, the Wasserstein distance is related to optimal transport theory, depends on the choice of metric ρ, and is a proper distance between probability measures.

2.3 Optimization

The exact computation of the MWE $\hat{\theta}_n$ is in general intractable, if only because of the intractability of $\mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$. We can envision the approximation of this distance based on synthetic samples generated given θ. Assume for the moment that a synthetic data set $z_{1:m}$ can be sampled from $\mu_\theta^{(m)}$ by setting $z_{1:m} = g_m(u, \theta)$, where g_m is a deterministic function of the parameter θ and a fixed-dimensional random variable u independent of θ. Given u, the approximate distance $\mathfrak{W}_p(y_{1:n}, g_m(u, \theta))$ is a deterministic function of θ which can be numerically optimized.

In order to reduce the variability of the distance approximation, one could average over $k \geq 1$ replicate datasets, effectively optimizing the approximate distance $k^{-1} \sum_{i=1}^{k} \mathfrak{W}_p(y_{1:n}, g_m(u^{(i)}, \theta))$, where $u^{(i)}$ are i.i.d. Variations of this approach were discussed already in the context of indirect inference (Section 2 of Gouriéroux et al., 1993). In the limit $k \to \infty$, $k^{-1} \sum_{i=1}^{k} \mathfrak{W}_p(y_{1:n}, g_m(u^{(i)}, \theta)) \to \mathbb{E}[\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})]$ almost surely. The resulting estimator then acts as an approximation to the MEWE when k is large. This can be made precise by viewing the above optimization as a step within a Monte Carlo Expectation-Maximization algorithm (Wei and Tanner, 1990). Convergence results for such algorithms, as both the number of iterations and the value of k go to infinity, are reviewed in Neath et al. (2013).

If m is large, the empirical distribution of $z_{1:m} = g_m(u, \theta)$ is expected to be close to μ_θ, so that we expect the MEWE to be close to the MWE with large probability. The supplementary materials contain a result showing that the MEWE $\hat{\theta}_{n,m}$ indeed converges to the MWE $\hat{\theta}_n$ as $m \to \infty$, and converges to $\theta_\star$ as $\min\{m, n\} \to \infty$. However, the rate of convergence of an empirical distribution $\hat{\nu}_m$ to its limit ν in the Wasserstein distance is known to depend adversely on the dimension of $\mathcal{Y}$, and is in general slower than $\sqrt{m}$ (see e.g. Remark 4.4 in Del Barrio and Loubes, 2017). As a consequence, it might be that the estimators discussed here are quite different from the MWE for small m. The effects of k and m on the estimators are illustrated in Section 2.4.1, Figure 2.

The incremental cost of increasing k is typically lower than that of increasing m, due in part to the potential for parallelization when calculating the distances $\mathfrak{W}_p(y_{1:n}, g_m(\theta, u^{(i)}))$ for a given θ, and in part to the algorithmic complexity in m, which might be super-linear. In Section 6, we briefly discuss alternative minimum distance estimators based on other approximations of the Wasserstein distance to reduce the computational cost.

In the spirit of Monte Carlo optimization, we can alternatively modify the sampling algorithms used for the ABC approach described in Section 3.3 to approximate the point estimator $\hat{\theta}_{n,m}$. This has the added benefit of not requiring the synthetic data to be generated via a deterministic function g_m. Related discussions can be found in Wood (2010); Rubio et al. (2013). Approximations of the MWE and MEWE are computed on toy examples in the following section, illustrating some of the theoretical properties to be proved in Section 5. However, due to the lack of a satisfactory general solution to solve the minimum distance optimization problems, we focus on the ABC method in the computational methods and numerical experiments of Sections 6 and 7, leaving an empirical study of the point estimators for future research.

Remark 2.1. The article Montavon et al. (2016) proposes an approximation of the gradient of the function that maps θ to an entropy-regularized Wasserstein distance (see Section 6.4) between $\hat{\mu}_n$ and μ_θ. Unfortunately, it is not applicable in the setting of purely generative models, as it involves point-wise evaluations of

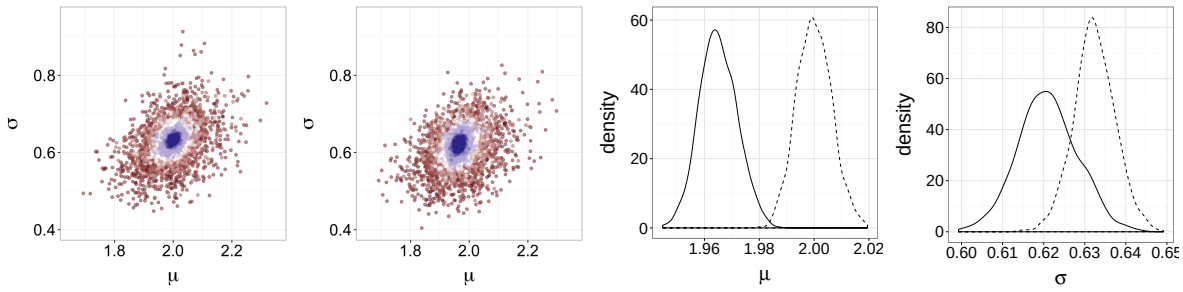
the derivative of the log-likelihood.

2.4 Illustrations

In Section 2.4.1, we compare the distribution of the MEWE with that of the MLE in a simple misspecified setting. We also investigate the effect of k and m on the distribution of the approximate MEWE. In Section 2.4.2, we consider a heavy-tailed data-generating process and highlight the robustness of the MEWE with $p = 1$.

2.4.1 Gamma data fitted with a Normal model

Let $\mu_\star = \text{Gamma}(10,5)$ (parametrized by shape and rate) and $\mathcal{M} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$. Figure 1 compares the sampling distributions of the MLE and approximations of the MEWE of order 1, over $M = 1,000$ experiments. The MEWE converges at the same $\sqrt{n}$ rate as the MLE, albeit to a distribution that is centered at a different location. For the MEWE, we have used $m = 10^4$ and $k = 20$.



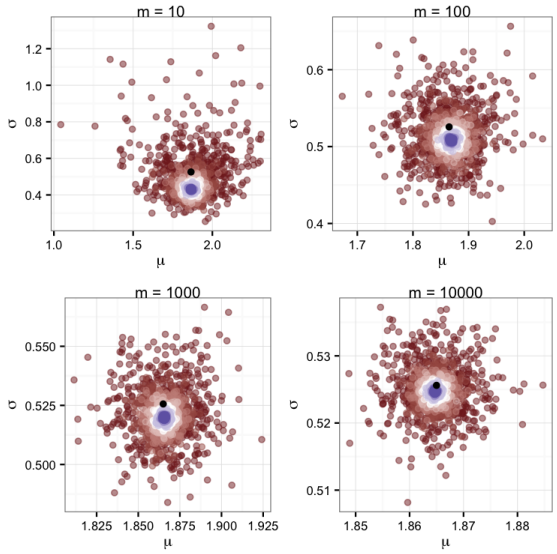
(a) MLE. (b) MEWE. (c) Estimators of μ. (d) Estimators of σ.

Figure 1: Gamma data fitted with a Normal model, as described in Section 2.4.1. Figures 1a and 1b show the sampling distributions of the MLE and MEWE of order 1 respectively, as n ranges from 50 to 10^4 (colors from red to white to blue). Figures 1c and 1d show the marginal densities of the estimators of μ and σ respectively, for $n = 10^4$; the MLEs are shown in dashed lines and the MEWE in full lines. For the MEWE, we have used $m = 10^4$ and $k = 20$.

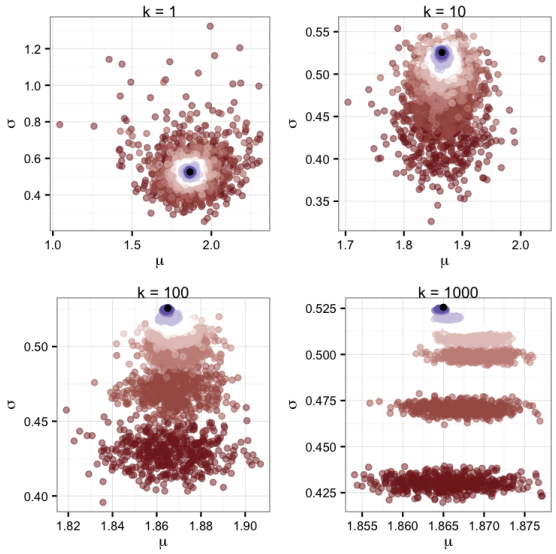
In Figure 2, we fix an observed data set of size $n = 100$, and compute $M = 500$ instances of the approximate MEWE for 8 different values of k and m, ranging from 1 to 1,000 and 10 to 10,000 respectively. In Figure 2a, we plot the estimators obtained for all the levels of k, given 4 different values of m. In Figure 2b, we plot the estimators obtained for all the levels of m, given 4 different values of k. The axis scales are different for each subplot. In both figures, black points correspond to the "true" MWE, calculated using a very large value of m ($m = 10^8$). For low values of m, the estimators might be significantly different from the MWE, as can be seen from the lower-right sub-plots of Figure 2b. When m increases, the estimators converge to the MWE. Increasing k reduces variation in the estimator. The changes in k and m had no significant impact on the number of evaluations of the objective required to locate the maximum using the `optim` function in `R` (R Core Team, 2015).

2.4.2 Cauchy data fitted with a Normal model

Let $\mu_\star$ be Cauchy with median zero and scale one, and consider the model $\mathcal{M} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$. Neither the MLE nor the MEWE of order $p \geq 2$ converges in this setting. We explore the behavior of the



(a) Approximate MEWE for increasing k (colors from red to white to blue), for different values of m.



(b) Approximate MEWE for increasing m (colors from red to white to blue), for different values of k.

Figure 2: Gamma data with $n = 100$, fitted with a Normal model, as described in Section 2.4.1. MEWEs are obtained for different values of m (from 10 to 10,000) and k (from 1 to 1,000), $M = 500$ times independently. The black dots represent the "exact" MWE computed with $m = 10^8$.

MEWE of order 1, over $M = 1,000$ repeated experiments. Figure 3 shows its sampling distributions, for n ranging from 50 to 10^4. The marginal distribution of the estimator of μ concentrates around 0, the median of $\mu_\star$. The marginal distribution of the estimator of σ also concentrates, around a value between 2 and 2.5. The concentration occurs at rate $\sqrt{n}$, as shown by the marginal densities of the rescaled estimators of μ in Figure 3b. Robustness properties of general minimum distance estimators were discussed in Parr and Schucany (1980), and of the MWE in location models in particular in Bassetti and Regazzini (2006).

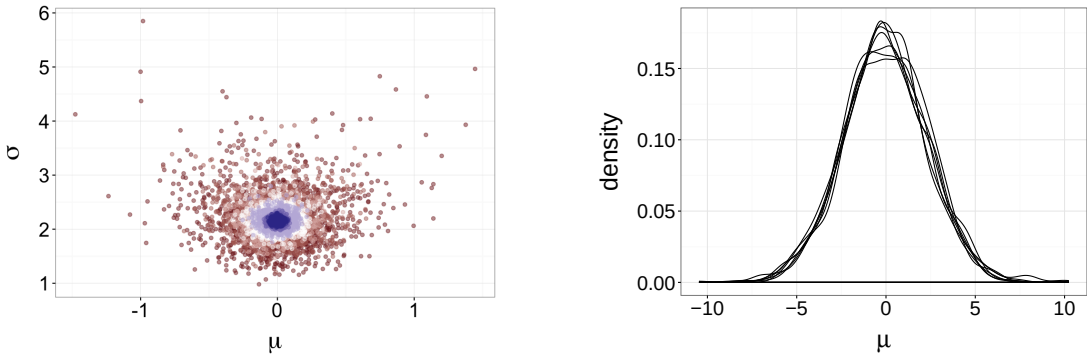
3 ABC with the Wasserstein distance

In this section we introduce Approximate Bayesian computation, first with a generic distance $\mathfrak{D}$ between data sets in Section 3.1, before focusing on the Wasserstein distance from Section 3.2 onwards.

3.1 Approximate Bayesian computation

Introduce a prior distribution π on the parameter θ. Consider the following algorithm, where $\varepsilon > 0$ is referred to as the threshold, and $\mathfrak{D}$ denotes a discrepancy measure between two data sets $y_{1:n}$ and $z_{1:n}$, taking non-negative values.

1. Draw a parameter θ from the prior distribution π, and a synthetic dataset $z_{1:n} \sim \mu_\theta^{(n)}$.
2. If $\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \varepsilon$, keep θ, otherwise reject it.



(a) Estimators of (μ, σ).

(b) $\sqrt{n}$-rescaled estimators of μ.

Figure 3: Cauchy data fitted with a Normal model, as described in Section 2.4.2. Sampling distributions of MEWE as n ranges from 50 to 10^4 (colors from red to white to blue, left). The marginals of MEWE of μ, rescaled by $\sqrt{n}$, are shown on the right.

The accepted samples are drawn from the ABC posterior distribution

$$\pi^{\varepsilon}(d\theta|y_{1:n}) = \frac{\pi(d\theta)\int_{\mathcal{Y}^n} \mathbb{1}\left(\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \varepsilon\right) \mu_\theta^{(n)}(dz_{1:n})}{\int_{\mathcal{H}} \pi(d\theta)\int_{\mathcal{Y}^n} \mathbb{1}\left(\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \varepsilon\right) \mu_\theta^{(n)}(dz_{1:n})}, \tag{6}$$

where $\mathbb{1}$ is the indicator function. A more sophisticated algorithm to approximate ABC posteriors is described in Section 3.3, and will be used throughout the numerical experiments of Section 7.

Suppose that $\mathfrak{D}$ is chosen as the Euclidean distance between the vectors $y_{1:n}$ and $z_{1:n}$. Then, the resulting ABC posterior can be shown to converge to the standard posterior as $\varepsilon \to 0$ (Prangle et al., 2017, see also Proposition 5.1). However, the approach tends to be impractical due to the large variation of $\mathfrak{D}(y_{1:n}, z_{1:n})$ over repeated samples from $\mu_\theta^{(n)}$. An example of practical use of ABC with the Euclidean distance is given in Sousa et al. (2009). A large proportion of the ABC literature is devoted to studying ABC posteriors in the setting where $\mathfrak{D}$ is the Euclidean distance between summaries, i.e. $\mathfrak{D}(y_{1:n}, z_{1:n}) = \|\eta(y_{1:n}) - \eta(z_{1:n})\|$, where $\eta : \mathcal{Y}^n \to \mathbb{R}^{d_\eta}$ for some small d_η. Using summaries can lead to a loss of information: the resulting ABC posterior converges, at best, to the conditional distribution of θ given $\eta(y_{1:n})$, as $\varepsilon \to 0$. A trade-off ensues, where using more summaries reduces the information loss, but increases the variation in the distance over repeated model simulations (Fearnhead and Prangle, 2012).

3.2 Wasserstein ABC

The distribution $\pi^{\varepsilon}(d\theta|y_{1:n})$ of Eq. (6), with $\mathfrak{D}$ replaced by $\mathfrak{W}_p$, is referred to as the Wasserstein ABC (WABC) posterior.

In some cases, the WABC posterior coincides with more familiar ABC posterior distributions. For instance, consider the case where $\mathcal{Y} \subset \mathbb{R}$. Then, as discussed in Section 6, the WABC posterior corresponds to the ABC posterior based on order statistics. Using order statistics as a choice of summary within ABC has been suggested multiple times in the literature, see e.g. Sousa et al. (2009); Fearnhead and Prangle (2012), without explicitly making the link to the Wasserstein distance. The connection to the Wasserstein distance justifies that choice and leads to methodological extensions in multivariate and dependent data settings.

In Section 5.2, we will show that, in some generality, the WABC posterior converges to the standard

posterior as $\varepsilon \to 0$. We will also consider the asymptotic behavior of $\pi^{\varepsilon}(d\theta|y_{1:n})$ when both $n \to \infty$ and $\varepsilon \to \varepsilon_\star$, for some minimal value $\varepsilon_\star$, and study its concentration around $\theta_\star = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$, assuming it is well-defined. The WABC posterior is also a special case of the coarsened posterior, as introduced in Miller and Dunson (2015), and, as such, might benefit from robustness to model misspecification; see Section 5.2.2.

3.3 Sampling sequentially from the WABC posterior

Instead of the rejection sampler of Section 3.1, we will target the WABC posterior using a sequential Monte Carlo (SMC) approach, with N particles exploring the parameter space (Del Moral et al., 2012). The algorithm starts with a threshold $\varepsilon_0 = +\infty$, for which the WABC posterior is the prior. Given the Monte Carlo approximation of the WABC posterior for ε_{t-1}, the next value ε_t is chosen so as to maintain a number of unique particles of at least αN, with α set to 50% by default. Upon choosing ε_t, resampling and rejuvenation steps are triggered and the algorithm proceeds. In the experiments we will run the algorithm until a fixed budget of model simulations is reached. At the end of the run, the algorithm provides N parameter samples and synthetic data sets, associated with a threshold ε_T.

The algorithm is parallelizable over the N particles, and thus over equally many model simulations and distance calculations. Any choice of MCMC kernel can be used within the rejuvenation steps. In particular, we use the r-hit kernel of Lee (2012), shown to be advantageous compared to standard ABC-MCMC kernels in Lee and Łatuszyński (2014). We choose the number of hits to be 2 by default. For the proposals of the MCMC steps, we use a mixture of multivariate Normal distributions, with 5 components by default. These default tuning parameters are used throughout all the experiments of this article. Full details on the SMC algorithm are given in the supplementary materials.

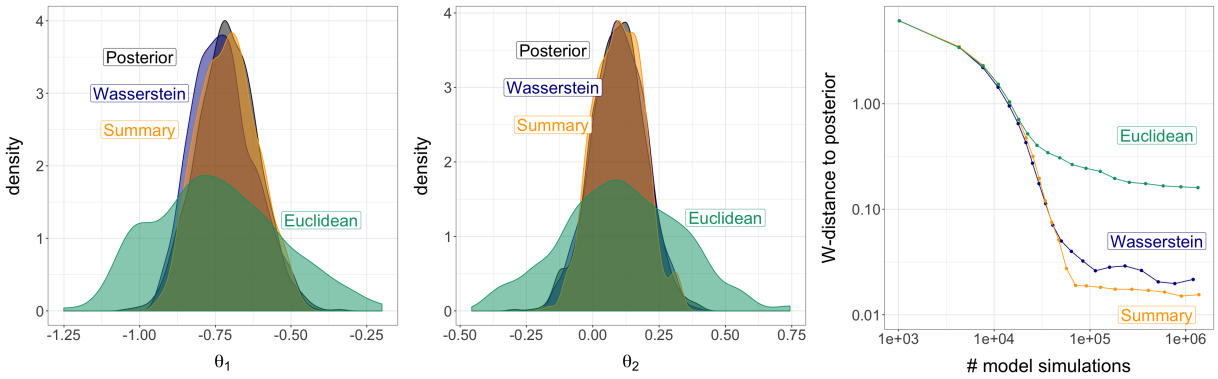
3.4 Illustration on a Normal location model

In a simple numerical experiment, we compare the WABC posterior with two other methods: ABC using the Euclidean distance between the data sets, and ABC using sufficient statistics. The three ABC posteriors converge to the standard posterior as $\varepsilon \to 0$.

Consider 100 observations generated from a bivariate Normal distribution. The mean components are drawn from a standard Normal distribution, and the generated values are approximately -0.59 and 0.03. The covariance is equal to 1 on the diagonal and 0.5 off the diagonal. The parameter θ is the mean vector, and is assigned a centered Normal prior with variance 25 on each component. All methods are run for a budget of 10^6 model simulations with the SMC approach of Section 3.3, using $N = 1,024$ particles. Approximations of the marginal posterior distributions of both parameters are given in Figures 4a and 4b, illustrating that ABC methods with Wasserstein and with summaries both approximate the posterior accurately.

To quantify the difference between the obtained ABC samples and posterior samples, we use, again, the Wasserstein distance. That is, we sample $1,024$ times independently from the posterior distribution, and compute the Wasserstein distance between these samples and the WABC samples produced by SMC using $N = 1,024$ particles. We plot the resulting distances against the number of model simulations in Figure 4c, in log-log scale. As expected, ABC with sufficient statistics converges fastest to the posterior. The proposed WABC approach requires more model simulations to yield comparable results. Finally, the ABC approach with the Euclidean distance struggles to approximate the posterior accurately. Extrapolating from the plot, it would seemingly take billions of model simulations for the latter ABC approach to approximate

the posterior as accurately as the other two methods. On the other hand, computing Euclidean distances between data sets is faster than computing Wasserstein distances; see Section 6.



(a) Posteriors of θ_1.

(b) Posteriors of θ_2.

(c) $\mathfrak{W}_1$-distance to posterior, versus number of model simulations (log-log scale).

Figure 4: ABC in the bivariate Normal location model of Section 3.4. ABC approximations of the posterior after 10^6 model simulations (left and middle), for three different distances. On the right, the Wasserstein distance between ABC posterior samples and exact posterior samples is plotted against the number of model simulations (in log-log scale). In principle, the three ABC approximations converge to the posterior as $\varepsilon \to 0$. Yet, for a given number of model simulations, the quality of the ABC approximation is sensitive to the choice of distance.

4 Time series

Viewing data sets as empirical distributions requires some additional care in the case of dependent data, which are common in settings where ABC and indirect inference methods are useful. A naïve approach consists in ignoring dependencies and computing distances between marginal empirical distributions. This might be enough to estimate all parameters in some cases, as illustrated in Section 7.3. However, in general, ignoring dependencies might prevent some parameters from being identifiable, as illustrated in Example 4.1.

Example 4.1. *Consider an autoregressive process of order 1, written AR(1), where* $y_1 \sim \mathcal{N}(0, \sigma^2/(1-\phi^2))$*, for some* $\sigma > 0$ *and* $\phi \in (-1, 1)$*. For each* $t \geq 2$*, let* $y_t = \phi y_{t-1} + \sigma w_t$*, where* $w_t \sim \mathcal{N}(0, 1)$ *are independent. The marginal distribution of each* y_t *is* $\mathcal{N}(0, \sigma^2/(1-\phi^2))$*. Furthermore, by an ergodic theorem, the empirical distribution* $\hat{\mu}_n$ *of the time series converges to this marginal distribution. The two parameters* (ϕ, σ^2) *are not identifiable from the limit* $\mathcal{N}(0, \sigma^2/(1-\phi^2))$*. Figure 6a shows WABC posterior samples derived while ignoring time dependence, obtained for decreasing values of* ε*. The prior is uniform on* $[-1, 1]$ *for* ϕ*, and standard Normal on* $\log(\sigma)$*. The data are generated using* $\phi = 0.7, \log(\sigma) = 0.9$ *and* $n = 1,000$*. The WABC posteriors concentrate on a ridge of values with constant* $\sigma^2/(1-\phi^2)$*.*

We propose two main routes to extend the proposed methodology to time series.

4.1 Curve matching

Visually, we might consider two time series to be similar if their curves are similar, in a trace plot of the series in the vertical axis against the time indices on the horizontal axis. The Euclidean vector distance

between curves sums the vertical differences between pairs of points with identical time indices. We can instead introduce the points $\tilde{y}_t = (t, y_t)$ and $\tilde{z}_t = (t, z_t)$ for all $t \in 1 : n$, viewing the trace plot as a scatter plot. The distance between two points, (t, y_t) and (s, z_s), can be measured by a weighted distance $\rho_\lambda((t, y_t), (s, z_s)) = \|y_t - z_s\| + \lambda|t - s|$, where λ is a non-negative weight, and $\|y - z\|$ refers to the Euclidean distance between y and z. Intuitively, the distance ρ_λ takes into account both vertical and horizontal differences between points of the curves, λ tuning the relative importance of horizontal to vertical differences. We can then define the Wasserstein distance between two empirical measures supported by $\tilde{y}_{1:n}$ and $\tilde{z}_{1:n}$, with ρ_λ as a ground distance on the observation space $\{1, \ldots, n\} \times \mathcal{Y}$. Since computing the Wasserstein distance can be thought of as solving an assignment problem, a large value of λ implies that y_t will be assigned to z_t, for all t. The transport cost will then be $n^{-1} \sum_{t=1}^{n} \|y_t - z_t\|$, corresponding to the Euclidean distance (up to a scaling factor). If λ is smaller, (t, y_t) is assigned to some (s, z_s), for some s possibly different than t. If λ goes to zero, the distance coincides with the Wasserstein distance between the marginal empirical distributions of $y_{1:n}$ and $z_{1:n}$, where the time element is entirely ignored.

For any $\lambda > 0$, we will discuss in Section 5.2 how the WABC posterior converges to the standard posterior distribution as $\varepsilon \to 0$. The choice of λ is open, but a simple heuristic for univariate time series goes as follows. Consider the aspect ratio of the trace plot of the time series (y_t), with horizontal axis spanning from 1 to t, and vertical axis from $\min_{t \in 1:n} y_t$ to $\max_{t \in 1:n} y_t$. For an aspect ratio of $H : V$, one can choose λ as $((\max_{t \in 1:n} y_t - \min_{t \in 1:n} y_t)/V) \times (H/n)$. This corresponds to the Euclidean distance in a rectangular plot with the given aspect ratio.

The proposed curve matching distance shares similarities with dynamic time warping (Berndt and Clifford, 1994), with the Skorokhod distance between curves (Majumdar and Prabhu, 2015) and with the Fréchet distance between polygons (Buchin et al., 2008), in which y_t would be compared to $z_{r(t)}$, where r is a retiming function to be optimized.

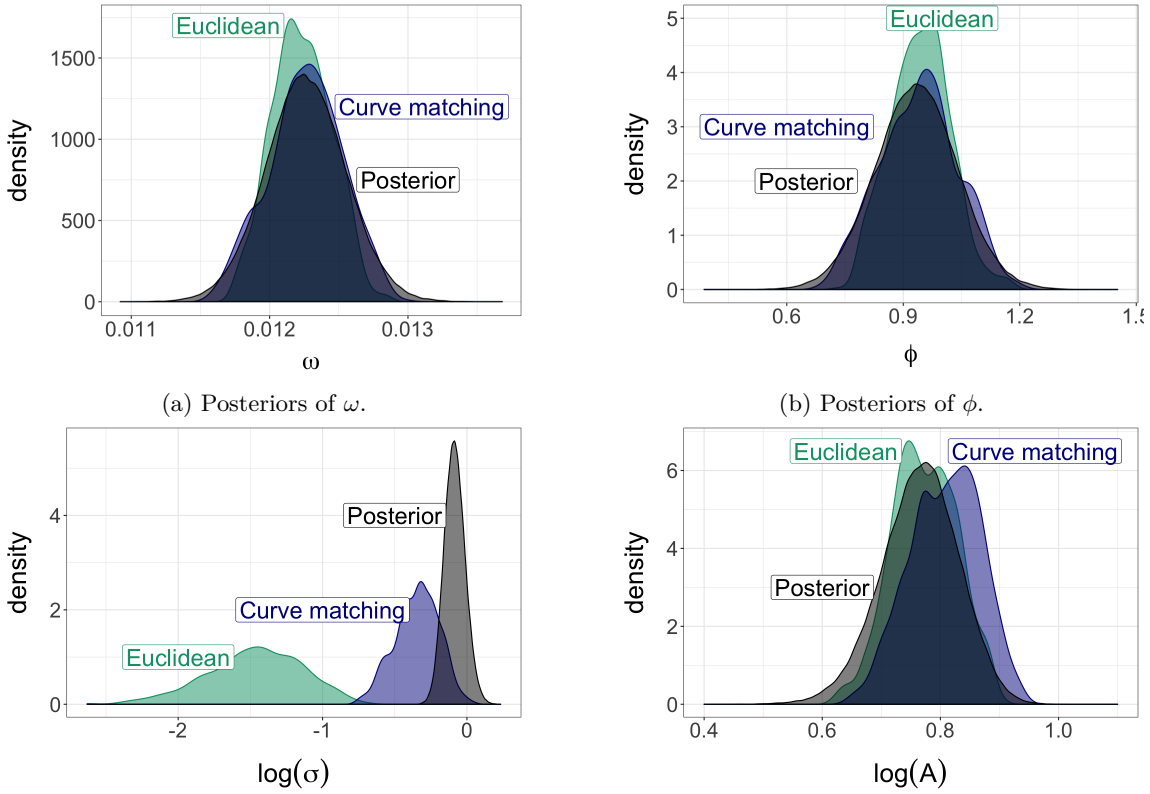
Example 4.2. *Consider a cosine model where* $y_t = A\cos(2\pi\omega t + \phi) + \sigma w_t$, *where* $w_t \sim \mathcal{N}(0, 1)$, *for all* $t \geq 1$, *are independent. Information about* ω *and* ϕ *is mostly lost when considering the marginal empirical distribution of* $y_{1:n}$. *In Figure 5, we compare the ABC posteriors obtained either with the Euclidean distance between the series, or with curve matching, with an aspect ratio of one; in both cases the algorithm is run for* 10^6 *model simulations. The figure also shows an approximation of the exact posterior distribution, obtained via Metropolis–Hastings. The prior distributions are uniform on* $[0, 1/10]$ *and* $[0, 2\pi]$ *for* ω *and* ϕ *respectively, and standard Normal on* $\log(\sigma)$ *and* $\log(A)$. *The data are generated using* $\omega = 1/80$, $\phi = \pi/4$, $\log(\sigma) = 0$ *and* $\log(A) = \log(2)$, *with* $n = 100$. *We see that curve matching yields a more satisfactory estimation of* σ *in Figure 5c, and a similar approximation for the other parameters. By contrast, an ABC approach based on the marginal distribution of* $y_{1:n}$ *would fail to identify* ϕ.

4.2 Reconstructions

Our second approach consists in transforming the time series to define an empirical distribution $\tilde{\mu}_n$ from which parameters can be estimated.

4.2.1 Delay reconstruction

In time series analysis, the lag-plot is a scatter plot of the pairs $(y_t, y_{t-k})_{t=k+1}^{n}$, for some lag $k \in \mathbb{N}$, from which one can inspect the dependencies between lagged values of the series. Similarly, delay reconstructions can be defined as $\tilde{y}_t = (y_t, y_{t-\tau_1}, \ldots, y_{t-\tau_k})$ for some integers $\tau_1, \ldots, \tau_k$. The sequence, denoted $\tilde{y}_{1:n}$ after



(a) Posteriors of ω.

(b) Posteriors of ϕ.

(c) Posteriors of $\log(\sigma)$.

(d) Posteriors of $\log(A)$.

Figure 5: ABC posterior samples in the cosine model of Example 4.2, using either the Euclidean distance or curve matching, with an aspect ratio of one, after 10^6 model simulations; and posterior distribution, obtained by Metropolis–Hastings. The standard deviation of the noise σ is better estimated with curve matching than with the Euclidean distance between time series.

relabelling and redefining n, inherits many properties from the original series, such as stationarity. Therefore, the empirical distribution of $\tilde{y}_{1:n}$, denoted by $\tilde{\mu}_n$, might converge to a limit $\tilde{\mu}_\star$. In turn, $\tilde{\mu}_\star$ might capture enough of the dependency structure of original series for the model parameters to be identified. Delay reconstructions (or embeddings) play a central role in dynamical systems (Kantz and Schreiber, 1997), for instance in Takens' theorem and variants thereof (Stark et al., 2003). The Wasserstein distance between the empirical distributions of delay reconstructions has previously been proposed as a way of measuring distance between time series (Moeckel and Murray, 1997; Muskulus and Verduyn-Lunel, 2011), but not as a device for parameter inference. In the ABC and MDE settings, we propose to construct the delay reconstructions of each synthetic time series, and to compute the Wasserstein distance between their empirical distribution and the empirical distribution of $\tilde{y}_{1:n}$. We refer to this approach as WABC and MWE with delay reconstruction respectively.

Denote by $\tilde{\mu}_\theta$ the marginal distribution of delay reconstructions under the model distribution given θ, assuming that the model generates strictly stationary time series. Denote by $\tilde{\mu}_{\theta,m}$ the empirical distribution of $\tilde{z}_{1:m}$. Then, provided that the empirical distribution $\tilde{\mu}_{\theta,m}$ converges to $\tilde{\mu}_\theta$, we are back in a setting where we can study the behavior of the MWE and the WABC posterior. Define $\tilde{\theta}_\star = \operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\tilde{\mu}_\star, \tilde{\mu}_\theta)$, assuming existence and uniqueness. In well-specified settings, if $\tilde{\theta}_\star$ is unique, then it must correspond to the

data-generating parameters.

When the entries of the vectors $y_{1:n}$ and $z_{1:n}$ are all unique, which happens with probability one when $\mu_\star^{(n)}$ and $\mu_\theta^{(n)}$ are continuous distributions, then $\mathfrak{W}_p(\tilde{y}_{1:n}, \tilde{z}_{1:n}) = 0$ if and only if $y_{1:n} = z_{1:n}$. To see this, consider the setting where $\tilde{y}_t = (y_t, y_{t-1})$, and $\tilde{z}_t = (z_t, z_{t-1})$. For the empirical distributions of $\tilde{y}_{1:n}$ and $\tilde{z}_{1:n}$ to be equal, we require that for every t there exists a unique s such that $\tilde{y}_t = \tilde{z}_s$. However, since the values in $y_{1:n}$ and $z_{1:n}$ are unique, the values y_1 and z_1 appear only as the second coordinates of $\tilde{y}_2$ and $\tilde{z}_2$ respectively. It therefore has to be that $y_1 = z_1$ and $\tilde{y}_2 = \tilde{z}_2$. In turn, this implies that $y_2 = z_2$, and inductively, $y_t = z_t$ for all $t \in 1:n$. A similar reasoning can be done for any $k \geq 2$ and $1 \leq \tau_1 < \ldots < \tau_k$. This property is important in establishing that the WABC posterior based on delay reconstruction converges to the true posterior as $\varepsilon \to 0$, which will be proved in Section 5.2.

In practice, for a non-zero value of ε, the obtained ABC posteriors might be different from the posterior, but still identify the parameter $\tilde{\theta}_\star$ with a reasonable accuracy, as illustrated in Example 4.3. Similar reasoning holds for approximations of the MWE. Since the order of the original data is only partly reflected in delay reconstructions, some model parameters might be difficult to estimate with delay reconstruction, such as the phase shift ϕ in Example 4.2.

Example 4.3 (Example 4.1 continued)**.** *Using $k = 1$, we consider $\tilde{y}_t = (y_t, y_{t-1})$ for $t \geq 2$. The reconstructions are then sub-sampled to* 500 *values, $\tilde{y}_2 = (y_2, y_1), \tilde{y}_4 = (y_4, y_3), \ldots, \tilde{y}_{1000} = (y_{1000}, y_{999})$; similar results were obtained with the* 999 *reconstructed values, but sub-sampling leads to computational gains in the Wasserstein distance calculations; see Section 6. The stationary distribution of $\tilde{y}_t$ is given by*

$$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \frac{\sigma^2}{1-\phi^2}\begin{pmatrix}1 & \phi\\ \phi & 1\end{pmatrix}\right). \tag{7}$$

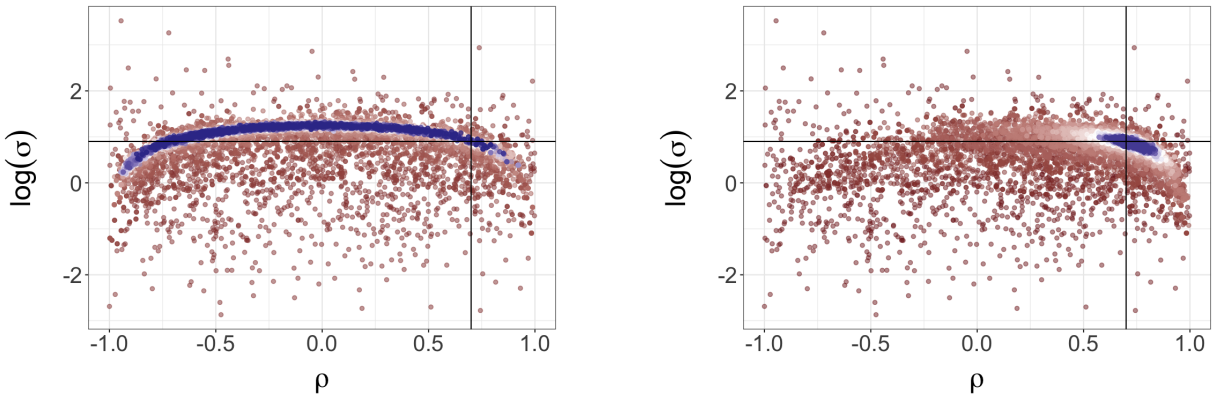
Both parameters σ^2 and ϕ can be identified from a sample approximating the above distribution. Figure 6b shows the WABC posteriors obtained with delay reconstruction, concentrating around the data-generating values as ε decreases.

4.2.2 Residual reconstruction

Another approach to handle dependent data is advocated in Mengersen et al. (2013), in the context of ABC via empirical likelihood. In various time series models, the observations are modeled as transformations of some parameter θ and residual variables $w_1, \ldots, w_n$. Then, given a parameter θ, one might be able to reconstruct the residuals corresponding to the observations. In Example 4.2, one can define $w_t = (y_t - A\cos(2\pi\omega t + \phi))/\sigma$. In Example 4.1, one can define $w_t = (y_t - \phi y_{t-1})/\sigma$; other examples are given in Mengersen et al. (2013). Once the residuals have been reconstructed, their empirical distribution can be compared to the distribution that they would follow under the model, e.g. a standard Normal in Examples 4.1 and 4.2.

5 Theoretical properties

For the point estimators based on the Wasserstein distance, we establish conditions for the existence, measurability and consistency. Bassetti et al. (2006) study these questions for estimators derived from minimizing a class of optimal transport costs that contains the MWE as the most important special case, but only for



(a) WABC using the Wasserstein distance between marginal distributions.

(b) WABC using the Wasserstein distance between empirical distributions of delay reconstructions.

Figure 6: WABC posteriors of $(\phi, \log(\sigma))$ in the AR(1) example (colors from red to white to blue as ε decreases), corresponding to Examples 4.1 and 4.3. On the left, using the marginal empirical distribution of the series, WABC posteriors concentrate around a ridge of values such that $\sigma^2/(1-\phi^2)$ is constant. On the right, using delay reconstruction with lag $k = 1$, the WABC posteriors concentrate around the data-generating parameters, $\phi = 0.7, \log(\sigma) = 0.9$, indicated by full lines.

well-specified models with i.i.d data. We give a new proof that extends their results to cover misspecified models and certain types of non-i.i.d. data. Beyond consistency, a study of asymptotic distributions is presented in Bassetti and Regazzini (2006) for location-scale models on $\mathbb{R}$, in the case where $p = 1$, the metric is Euclidean, and the model is well-specified. We extend these results to cover generic models under similar conditions. Results for the minimum expected Wasserstein estimator are presented in the supplementary materials.

Next, we study the behavior of the Wasserstein ABC posterior under different regimes. First, we give conditions on a discrepancy measure for the associated ABC posterior to converge to the true posterior as the threshold ε goes to zero, while keeping the observed data fixed. We then discuss the behavior of the WABC posterior as $n \to \infty$ for fixed $\varepsilon > 0$. Finally, we establish rates of concentration of the WABC posterior around $\theta_\star = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\theta, \mu_\star)$, provided this parameter is well-defined, as the data size n grows and the threshold ε shrinks, similarly to Frazier et al. (2016) in the case of summary-based ABC. Proofs are deferred to the supplementary materials.

5.1 On MWE and MEWE

5.1.1 Existence, measurability, and consistency

Importantly, we assume that for any θ, the synthetic data-generating process $\mu_\theta^{(n)}$ defines an identifiable distribution μ_θ, informally referred to as the model. In the i.i.d. setting, $\mu_\star$ and μ_θ are simply the marginal distributions of the observed and synthetic data. In non-i.i.d. settings, they might be the stationary or limiting marginal distributions of reconstructions, denoted by $\tilde{\mu}_\star$ and $\tilde{\mu}_\theta$ in Section 4. As such, in this section, we take $\mathcal{Y}$ to denote a generic space in which our (potentially reconstructed) observations lie. Further assumptions are listed below.

Assumption 5.1. *The data-generating process is such that* $\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \to 0$, $\mathbb{P}$*-almost surely as* $n \to \infty$.

Assumption 5.2. *The map $\theta \mapsto \mu_\theta$ is continuous in the sense that $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$ implies $\mu_{\theta_n} \Rightarrow \mu_\theta$.*

Assumption 5.3. *For some $\varepsilon > 0$, the set $B_\star(\varepsilon) = \{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_\star + \varepsilon\}$ is bounded.*

Theorem 5.1 (Existence and consistency of the MWE)**.** *Under Assumptions 5.1-5.3, there exists a set $E \subset \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$, and there exists $n(\omega)$ such that, for all $n \geq n(\omega)$, the sets $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ are non-empty and form a bounded sequence with*

$$\limsup_{n\to\infty} \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \subset \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta).$$

For a generic function f, let $\varepsilon\text{-}\operatorname{argmin}_x f = \{x : f(x) \leq \varepsilon + \inf_x f\}$. Theorem 5.1 also holds with $\varepsilon_n\text{-}\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ in place of $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$, for any sequence $\varepsilon_n \to 0$. If $\theta_\star = \operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$ is unique, the result can be rephrased as $\hat{\theta}_n \to \theta_\star$, $\mathbb{P}$-almost surely.

Theorem 5.2 (Measurability of the MWE)**.** *Suppose that $\mathcal{H}$ is a σ-compact Borel measurable subset of $\mathbb{R}^{d_\theta}$. Under Assumption 5.2, for any $n \geq 1$ and $\varepsilon > 0$, there exists a Borel measurable function $\hat{\theta}_n : \Omega \to \mathcal{H}$ that satisfies*

$$\hat{\theta}_n(\omega) \in \begin{cases} \operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) & \text{if this set is non-empty,} \\ \varepsilon\text{-}\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) & \text{otherwise.} \end{cases}$$

Similar results for the MEWE are stated in supplementary materials, along with a few additional assumptions. Another result therein gives conditions for the MEWE to converge to the MWE as $m \to \infty$.

5.1.2 Asymptotic distribution

Under conditions guaranteeing the consistency of the minimum Wasserstein estimator, we study its asymptotic distribution in the case where $p = 1$, $\mathcal{Y} = \mathbb{R}$, and $\rho(x, y) = |x - y|$. Under this setup, it can be shown that $\mathfrak{W}_1(\mu, \nu) = \int_0^1 |F_\mu^{-1}(s) - F_\nu^{-1}(s)|ds = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)|dt$ (e.g. Ambrosio et al., 2005, Theorem 6.0.2), where F_μ and F_ν denote the cumulative distribution functions (CDFs) of μ and ν respectively. We also assume that the model is well-specified, and that $\mathcal{H}$ is endowed with a norm: $\rho_{\mathcal{H}}(\theta, \theta') = \|\theta - \theta'\|_{\mathcal{H}}$.

Our approach to derive asymptotic distributions follows Pollard (1980). Let F_θ, $F_\star$ and F_n denote the CDFs of μ_θ, $\mu_\star$ and $\hat{\mu}_n$ respectively. Informally speaking, we will show that $\sqrt{n}W_1(\hat{\mu}_n, \mu_\theta)$ can be approximated by $\int_{\mathbb{R}} |\sqrt{n}(F_n(t) - F_\star(t)) - \langle \sqrt{n}(\theta - \theta_\star), D_\star(t)\rangle|dt$ near $\theta_\star$, for some $D_\star \in (L_1(\mathbb{R}))^{d_\theta}$, with $\langle \theta, u\rangle = \sum_{i=1}^{d_\theta} \theta_i u_i$. Results in del Barrio et al. (1999) and Dede (2009) give conditions under which $\sqrt{n}(F_n - F_\star)$ converges to a zero mean Gaussian process $G_\star$ with given covariance structure, for both independent and certain classes of dependent data. Heuristically, the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_\star)$ is then close to that of $\operatorname{argmin}_{u\in\mathcal{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_\star(t)\rangle|dt$. We only state the result for i.i.d. data, but also prove it for certain classes of non-i.i.d. data in the supplementary materials. First, we give another assumption, which is stated with general p and under potential model misspecification. We will rely on this general form later, but for the moment only need it for $p = 1$ and well-specified models.

Assumption 5.4. *For all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\inf_{\theta\in\mathcal{H}:\rho_{\mathcal{H}}(\theta,\theta_\star)\geq\varepsilon} \mathfrak{W}_p(\mu_\star, \mu_\theta) > \mathfrak{W}_p(\mu_\star, \mu_{\theta_\star}) + \delta.$$

This assumption is akin to those made in the study of the asymptotic properties of the maximum likelihood estimator, where $\theta_\star$ is defined in terms of the KL divergence. A result in the supplementary materials gives conditions under which this assumption holds.

Theorem 5.3. *Suppose $Y_i \sim \mu_\star = \mu_{\theta_\star}$ i.i.d. for $\theta_\star$ in the interior of $\mathcal{H}$, and that $\int_0^\infty \sqrt{\mathbb{P}(|Y_0| > t)}dt < \infty$. Suppose that there exists a non-singular $D_\star \in (L_1(\mathbb{R}))^{d_\theta}$ such that*

$$\int_{\mathbb{R}} |F_\theta(t) - F_\star(t) - \langle \theta - \theta_\star, D_\star(t)\rangle| dt = o(\|\theta - \theta_\star\|_{\mathcal{H}}),$$

as $\|\theta - \theta_\star\|_{\mathcal{H}} \to 0$. Under Assumptions 5.1-5.4 and if $\operatorname{argmin}_{u\in\mathcal{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_\star(t)\rangle| dt$ is almost surely unique, the MWE of order 1 satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \Rightarrow \operatorname*{argmin}_{u\in\mathcal{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_\star(t)\rangle| dt,$$

as $n \to \infty$, where $G_\star$ is a zero mean Gaussian process with $\mathbb{E}G_\star(s)G_\star(t) = \min\{F_\star(s), F_\star(t)\} - F_\star(s)F_\star(t)$.

The theorem aligns with the concentration of the estimators observed in Figures 1-3, but does not fully explain their behavior in those models, because of misspecification. Computing confidence intervals using the asymptotic distribution is hard, due in part to its dependence on unknown quantities. The bootstrap appears as a practical alternative. We leave further discussion of the bootstrap for future research.

The condition $\int_0^\infty \sqrt{\mathbb{P}(|Y_0| > t)}dt < \infty$ implies the existence of second moments, and is itself implied by the existence of moments of order $2 + \varepsilon$ for some $\varepsilon > 0$ (see e.g. Section 2.9 in Wellner and van der Vaart, 1996). The uniqueness assumption on the argmin can be relaxed by considering convergence to the entire set of minimizing values, as in Section 7 of Pollard (1980). Still, uniqueness can sometimes be established, using e.g. Cheney and Wulbert (1969). This approach is taken in Bassetti and Regazzini (2006), who directly show that Theorem 5.3 holds when $\mathcal{M}$ is a location-scale family supported on a bounded open interval. The existence and form of $D_\star$ can in many cases be derived if the model is differentiable in quadratic mean (Le Cam, 1970), which is elaborated upon in the supplementary materials.

Theorem 5.3 also holds for approximations of the MWE, say $\tilde{\theta}_n$, provided that $\tilde{\theta}_n = \hat{\theta}_n + o_{\mathbb{P}}(1/\sqrt{n})$, as can be seen from the proof. In light of the convergence of the MEWE to the MWE as $m \to \infty$ illustrated in the supplementary materials, there exists a sequence $m(n)$ (depending on ω) such that the MEWE $\hat{\theta}_{n,m(n)}$ satisfies the conclusion of the theorem, provided that $m(n)$ increases sufficiently fast.

A similar result can be derived when $p = 2$ using results in Del Barrio et al. (2005). Extensions of the theorem to multivariate settings is left for future research; the main difficulty stems from the lack of convenient representations of the Wasserstein distance in such settings.

5.2 On WABC posteriors

We study the behavior of the Wasserstein ABC posterior under different regimes. First, we give conditions on a discrepancy measure for the associated ABC posterior to converge to the true posterior as the threshold ε goes to zero, while keeping the observed data fixed. We then discuss the behavior of the WABC posterior as $n \to \infty$ for fixed $\varepsilon > 0$. Finally, we establish rates of concentration of the WABC posterior around $\theta_\star = \operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$, as the data size n grows and the threshold ε shrinks, similarly to Frazier et al. (2016) in the case of summary-based ABC. Proofs are deferred to the appendix.

5.2.1 Behavior as $\varepsilon \to 0$ for fixed observations

The following result establishes some conditions under which a non-negative measure of discrepancy between data sets $\mathfrak{D}$ yields an ABC posterior that converges to the true posterior as $\varepsilon \to 0$, while the observations are kept fixed.

Proposition 5.1. *Assume that the posterior distribution is well-defined, and suppose that $\mu_\theta^{(n)}$ has a continuous density $f_\theta^{(n)}$, satisfying*

$$\sup_{y_{1:n} \in \mathcal{Y}^n, \theta \in \mathcal{H}} f_\theta^{(n)}(y_{1:n}) < \infty.$$

Suppose also that $\mathfrak{D}$ is continuous in the sense that, for any $y_{1:n}$, $\mathfrak{D}(y_{1:n}, z_{1:n}) \to \mathfrak{D}(y_{1:n}, x_{1:n})$ whenever $z_{1:n} \to x_{1:n}$ component-wise in the ground metric ρ. Suppose that either

1. *$f_\theta^{(n)}$ is n-exchangeable, such that $f_\theta^{(n)}(y_{1:n}) = f_\theta^{(n)}(y_{\sigma(1:n)})$ for any $\sigma \in \mathcal{S}_n$, and $\mathfrak{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{\sigma(1:n)}$ for some $\sigma \in \mathcal{S}_n$, or*

2. *$\mathfrak{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{1:n}$.*

Then, keeping $y_{1:n}$ fixed, the ABC posterior converges strongly to the posterior as $\varepsilon \to 0$.

The Wasserstein distance applied to unmodified data satisfies $\mathfrak{W}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{\sigma(1:n)}$ for some $\sigma \in \mathcal{S}_n$, making condition (a) of Proposition 5.1 applicable. Furthermore, taking $\mathfrak{D}$ to be the Wasserstein distance applied to delay reconstructed or curve matched data, condition (b) of the proposition holds.

5.2.2 Behavior as $n \to \infty$ for fixed ε

5.2.3 Concentration as n increases and ε decreases

A sequence of distributions $\pi_{y_{1:n}}$ on $\mathcal{H}$, depending on the data $y_{1:n}$, is consistent at $\theta_\star$ if, for any $\delta > 0$, $\mathbb{E}[\pi_{y_{1:n}}(\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta, \theta_\star) > \delta\})] \to 0$, where the expectation is taken with respect to $\mu_\star^{(n)}$. Finding rates of concentration for $\pi_{y_{1:n}}$ involves finding the fastest decaying sequence $\delta_n > 0$ such that the limit above holds. More precisely, we say that the rate of concentration of $\pi_{y_{1:n}}$ is bounded above by δ_n if $\mathbb{E}[\pi_{y_{1:n}}(\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta, \theta_\star) > \delta_n\})] \to 0$.

We establish rates of concentration of the sequence of WABC posteriors around $\theta_\star = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\theta, \mu_\star)$, as the data size n grows and the threshold shrinks towards $\varepsilon_\star = \mathfrak{W}_p(\mu_{\theta_\star}, \mu_\star)$ at a rate dependent on n. Although we focus on the Wasserstein distance in this section, the reasoning holds for other metrics on $\mathcal{P}(\mathcal{Y})$; see Section 6 and the supplementary materials.

As with the MWE, we assume that for any θ, the synthetic data-generating process $\mu_\theta^{(n)}$ defines an identifiable distribution μ_θ, informally referred to as the model, and take $\mathcal{Y}$ to denote a generic space in which our (potentially reconstructed) observations lie. Further assumptions are listed below, starting with a slightly weaker version of Assumption 5.1.

Assumption 5.5. *The data-generating process is such that $\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \to 0$, in $\mathbb{P}$-probability, as $n \to \infty$.*

The moment and concentration inequalities of Fournier and Guillin (2015) can be used to verify the above assumption, as well as the next assumption for i.i.d. data and certain classes of dependent processes.

Assumption 5.6. *For any $\varepsilon > 0$, $\mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \varepsilon) \leq c(\theta) f_n(\varepsilon)$, where $f_n(\varepsilon)$ is a sequence of functions that are strictly decreasing in ε for fixed n and $f_n(\varepsilon) \to 0$ for fixed ε as $n \to \infty$. The function $c : \mathcal{H} \to \mathbb{R}^+$ is π-integrable, and satisfies $c(\theta) \leq c_0$ for some $c_0 > 0$, for all θ such that, for some $\delta_0 > 0$, $\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \delta_0 + \epsilon_\star$.*

For well-specified models, Assumption 5.6 implies Assumption 5.5. The next assumption states that the prior distribution puts enough mass on the sets of parameters θ that yield distributions μ_θ close to $\mu_\star$ in the Wasserstein distance.

Assumption 5.7. *There exist $L > 0$ and $c_\pi > 0$ such that, for all ε small enough,*

$$\pi\left(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon + \varepsilon_\star\}\right) \geq c_\pi \varepsilon^L.$$

Under Assumption 5.4, note that the last part of Assumption 5.6 is implied by $c(\theta) \leq c_0$ for all θ with $\mathfrak{W}_p(\mu_{\theta_\star}, \mu_\theta) \leq \delta_0$, for some $\delta_0 > 0$. Indeed, $\mathfrak{W}_p(\mu_{\theta_\star}, \mu_\theta) \leq \delta_0$ implies that $\mathfrak{W}_p(\mu_\theta, \mu_\star) - \mathfrak{W}_p(\mu_\star, \mu_{\theta_\star}) \leq \delta_0$. Since $\varepsilon_\star = \mathfrak{W}_p(\mu_\star, \mu_{\theta_\star})$, the argument follows. By the same reasoning, Assumption 5.7 is implied by $\pi\left(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon\}\right) \geq c_\pi \varepsilon^L$, for some $c_\pi > 0$ and $L > 0$.

Theorem 5.4. *Under Assumptions 5.5-5.7, consider a sequence $(\varepsilon_n)_{n \geq 0}$ such that, as $n \to \infty$, $\varepsilon_n \to 0$, $f_n(\varepsilon_n) \to 0$, and $\mathbb{P}(\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon_n) \to 1$. Then, the WABC posterior with threshold $\varepsilon_n + \varepsilon_\star$ satisfies, for some $0 < C < \infty$ and any $0 < R < \infty$,*

$$\pi^{\varepsilon_n + \varepsilon_\star}\left(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) > \varepsilon_\star + 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R)\} \,|y_{1:n}\right) \leq \frac{C}{R},$$

with $\mathbb{P}$-probability going to 1 as $n \to \infty$.

The assumptions that $f_n(\varepsilon_n) \to 0$ and that $\mathbb{P}(\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon_n) \to 1$ imply that ε_n has to be the slowest of the two convergence rates: that of $\hat{\mu}_n$ to $\mu_\star$ and that of $\hat{\mu}_{\theta,n}$ to μ_θ. We can further relate concentration on the sets $\{\theta : \mathfrak{W}_p(\mu_\theta, \mu_\star) < \delta' + \varepsilon_\star\}$, for some $\delta' > 0$, to concentration on the sets $\{\theta : \rho_{\mathcal{H}}(\theta, \theta_\star) < \delta\}$, for some $\delta > 0$, assuming the parameter $\theta_\star = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$ is well-defined. In turn, this leads to concentration rates of the WABC posteriors. To that end, consider the following assumption.

Assumption 5.8. *There exist $K > 0$, $\alpha > 0$ and an open neighborhood $U \subset \mathcal{H}$ of $\theta_\star$, such that, for all $\theta \in U$,*

$$\rho_{\mathcal{H}}(\theta, \theta_\star) \leq K(\mathfrak{W}_p(\mu_\theta, \mu_\star) - \varepsilon_\star)^\alpha.$$

Corollary 5.1. *Under Assumptions 5.4-5.8, consider a sequence $(\varepsilon_n)_{n \geq 0}$ such that, as $n \to \infty$, $\varepsilon_n \to 0$, $f_n(\varepsilon_n) \to 0$, $f_n^{-1}(\varepsilon_n^L) \to 0$ and $\mathbb{P}(\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon_n) \to 1$. Then the WABC posterior with threshold $\varepsilon_n + \varepsilon_\star$ satisfies, for some $0 < C < \infty$ and any $0 < R < \infty$,*

$$\pi^{\varepsilon_n + \varepsilon_\star}\left(\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta, \theta_\star) > K(4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R))^\alpha\} \,|y_{1:n}\right) \leq \frac{C}{R},$$

with $\mathbb{P}$-probability going to 1.

This result gives concentration rates through the expression $K(4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R))^\alpha$. We verify the assumptions and derive explicit rates for certain classes of models and data-generating processes in the supplementary materials. The main messages are: the concentration rate is sensitive to the dimension of the

observation space $\mathcal{Y}$, to the choice of the order p of the Wasserstein distance, and to the misspecification of the model.

However, it is unclear what happens when ε_n decays to a value smaller than $\varepsilon_\star$ at a rate faster than that prescribed by Corollary 5.1. For example, as shown in Proposition 5.1, the WABC posterior converges to the true posterior when $\varepsilon \to 0$ for n fixed. The posterior itself is known to concentrate around the point in $\mathcal{H}$ minimizing the KL divergence between $\mu_\star$ and μ_θ (e.g. Müller, 2013).

6 Distance calculations

Here we discuss algorithms and costs associated with computing the Wasserstein distance, and possible ways of reducing these costs using approximations and alternative distances. The supplementary materials feature theoretical results validating the use of these alternative distances, as referenced below.

6.1 Exact Wasserstein distance

Computing the Wasserstein distance between the distributions $\hat{\mu}_n = n^{-1}\sum_{i=1}^n \delta_{y_i}$ and $\hat{\nu}_n = n^{-1}\sum_{i=1}^n \delta_{z_i}$ reduces to a linear sum assignment problem, as in Eq. (3). In the univariate case, finding the optimal permutation can be done by sorting the vectors $y_{1:n}$ and $z_{1:n}$ in increasing order, obtaining the orders $\sigma_y(i)$ and $\sigma_z(i)$ for $i \in \{1, \ldots, n\}$. Then, one associates each y_i with $z_{\sigma(i)}$ where $\sigma(i) = \sigma_z \circ \sigma_y^{-1}(i)$. The cost of the Wasserstein distance computation is thus of order $n \log n$.

In multivariate settings, Eq. (3) can be solved by the Hungarian algorithm for a cost of order n^3. Other algorithms have a cost of order $n^{2.5}\log(n\,C_n)$, with $C_n = \max_{1\leq i,j\leq n}\rho(y_i, z_j)$, and can therefore be more efficient when C_n is small (Burkard et al., 2009, Section 4.1.3). In our numerical experiments, we use the short-list method presented in Gottschlich and Schuhmacher (2014) and implemented in Schuhmacher et al. (2017). This simplex algorithm-derived method also solves the more general Eq. (2), which is needed when approximating the MEWE for $m \neq n$. In general, simplex algorithms come without guarantees of polynomial running times, but Gottschlich and Schuhmacher (2014) show empirically that their method tends to have sub-cubic cost.

The cubic cost of computing Wasserstein distances in the multivariate setting can be prohibitive for large data sets. However, some applications of ABC and indirect inference involve relatively small numbers of observations from complex models which are expensive to simulate, while the cost of computing distances is model-free. Note that the dimension d_y of the observation space only enters the ground distance ρ, and thus the cost of computing the Wasserstein distance under a Euclidean ground metric is linear in d_y. For cases where the cubic cost in n is prohibitive, one can resort to various approximations of the Wasserstein distance, as described in the following sections.

6.2 Hilbert distance

The assignment problem in Eq. (3) can be solved in $n \log n$ in the univariate case by sorting samples. We propose a new distance generalizing this idea when $d_y > 1$, by sorting samples according to their projection via the Hilbert space-filling curve. As shown in Gerber and Chopin (2015); Schretter et al. (2016), transformations through the Hilbert space-filling curve and its inverse preserve a notion of distance between probability measures. The Hilbert curve $H : [0,1] \to [0,1]^{d_y}$ is a Hölder continuous mapping from $[0,1]$ into $[0,1]^{d_y}$. One can define a measurable pseudo-inverse $h : [0,1]^{d_y} \to [0,1]$ verifying $h(H(x)) = x$ for

all $x \in [0,1]$. We assume in this subsection that $\mathcal{Y} \subset \mathbb{R}^{d_y}$ is such that there exists a mapping $\psi : \mathcal{Y} \to (0,1)^{d_y}$ verifying, for $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$, $\psi(y) = \big(\psi_1(y_1), ..., \psi_{d_y}(y_{d_y})\big)$ where the ψ_i's are continuous and strictly monotone. For instance, if $\mathcal{Y} = \mathbb{R}^{d_y}$, one can take ψ to be the component-wise logistic transformation; see Gerber and Chopin (2015) for more details. By construction, the mapping $h_{\mathcal{Y}} := h \circ \psi : \mathcal{Y} \to (0,1)$ is one-to-one. For two vectors $y_{1:n}$ and $z_{1:n}$, denote by σ_y and σ_z the permutations obtained by mapping the vectors through $h_{\mathcal{Y}}$ and sorting the resulting univariate vectors in increasing order. We define the Hilbert distance $\mathfrak{H}_p$ between the empirical distributions of $y_{1:n}$ and $z_{1:n}$ by

$$\mathfrak{H}_p(y_{1:n}, z_{1:n})^p = \frac{1}{n} \sum_{i=1}^{n} \rho\big(y_i, z_{\sigma(i)}\big)^p, \tag{8}$$

where $\sigma(i) = \sigma_z \circ \sigma_y^{-1}(i)$ for all $i \in \{1, \ldots, n\}$.

Proposition 6.1. *For any integer $n \geq 1$ and real number $p \geq 1$, $\mathfrak{H}_p$ defines a distance on the space of empirical distributions of size n.*

The Hilbert distance can be computed at a cost in the order of $n \log n$ and an implementation is provided by the function `hilbert_sort` in The CGAL Project (2016). From a practical point of view, this implementation has the attractive property of not having to map the samples to $(0,1)^{d_y}$ and hence having to choose a specific mapping ψ. Instead, this function directly constructs the Hilbert curve around the input point set.

The Hilbert distance approximates the assignment problem of Eq. (3). Therefore, it is always greater than the Wasserstein distance, which minimizes such sums. This property plays an important role in analyzing the associated ABC posterior. In the supplementary materials, we provide a version of Theorem 5.4 that holds for the ABC posterior based on the Hilbert distance, assuming that the model is well-specified. A theoretical analysis of this approach under milder conditions is left for future research.

6.3 Swapping distance

Viewing the Wasserstein distance calculation as the assignment problem in Eq. (3), Puccetti (2017) proposed a swapping algorithm to approximate the optimal assignment. Consider an arbitrary permutation σ of $\{1, \ldots, n\}$, and the associated transport cost $\sum_{i=1}^{n} \rho(y_i, z_{\sigma(i)})^p$. The swapping algorithm consists in checking, for all $1 \leq i < j \leq n$, whether $\rho(y_i, z_{\sigma(i)})^p + \rho(y_j, z_{\sigma(j)})^p$ is less or greater than $\rho(y_i, z_{\sigma(j)})^p + \rho(y_j, z_{\sigma(i)})^p$. If it is greater, then one swaps $\sigma(i)$ and $\sigma(j)$, resulting in a decrease of the transport cost. One can repeat these sweeps over $1 \leq i < j \leq n$, until the assignment is left unchanged, and denote it by $\tilde{\sigma}$. Each sweep has a cost of order n^2 operations. There is no guarantee that the resulting assignment $\tilde{\sigma}$ corresponds to the optimal one. Note that we can initialize the algorithm with the assignment obtained by Hilbert sorting for a negligible cost of $n \log n$. We refer to the resulting distance $(n^{-1} \sum_{i=1}^{n} \rho(y_i, z_{\tilde{\sigma}(i)})^p)^{1/p}$ as the swapping distance.

The swapping distance between $y_{1:n}$ and $z_{1:n}$ takes values that are, by construction, between the Wasserstein distance $\mathfrak{W}_p(y_{1:n}, z_{1:n})$ and the Hilbert distance $\mathfrak{H}_p(y_{1:n}, z_{1:n})$. Thanks to this important property, the associated ABC posterior can be analyzed using results obtained for the Wasserstein and the Hilbert distances. A version of Theorem 5.4 for the swapping distance is provided in the supplementary materials.

6.4 Other distances

Recent articles introduce other algorithms to approximate the Wasserstein distance in a cost of n^2 (Cuturi, 2013; Genevay et al., 2016; Ye et al., 2017), or $n \log n$ using random projections (Rabin et al., 2011). In particular, Cuturi (2013) convexifies the optimization problem of Eq. (2) using an entropic constraint on the joint distribution γ. Consider the regularized version of Eq. (3) $\gamma^\zeta = \text{argmin}_{\gamma \in \Gamma_{n,n}} \sum_{i,j=1}^n \rho(y_i, z_j)^p \gamma_{ij} + \zeta \sum_{i,j=1}^n \gamma_{ij} \log \gamma_{ij}$, which includes a penalty on the entropy of γ, and define the dual-Sinkhorn divergence $S_p^\zeta(y_{1:n}, z_{1:n})^p = \sum_{i,j=1}^n \rho(y_i, z_j)^p \gamma_{ij}^\zeta$. The regularized problem can be solved iteratively by Sinkhorn's algorithm, which involves matrix-vector multiplications resulting in a total cost of order n^2. If ζ goes to zero, the dual-Sinkhorn divergence goes to the Wasserstein distance. More properties of the dual-Sinkhorn divergence are discussed in Cuturi (2013).

Following Ramdas et al. (2017), one can show that the so-called Energy Distance (ED) and Maximum Mean Discrepancy (MMD) arise as the respective limits, as $\zeta \to \infty$, of $\tilde{S}_1^\zeta(y_{1:n}, z_{1:n}) = 2S_1^\zeta(y_{1:n}, z_{1:n}) - S_1^\zeta(y_{1:n}, y_{1:n}) - S_1^\zeta(z_{1:n}, z_{1:n})$, under the ground distances $\rho(x, y)$ and $\tilde{\rho}(x, y) = 1 - \exp(-\rho(x, y)^2/(2s^2))$, where s is a tuning parameter. The ED takes the form

$$\tilde{S}_1^\infty(y_{1:n}, z_{1:n}) = \frac{2}{n^2} \sum_{i,j=1}^n \rho(y_i, z_j) - \frac{1}{n^2} \sum_{i,j=1}^n \rho(y_i, y_j) - \frac{1}{n^2} \sum_{i,j=1}^n \rho(z_i, z_j), \tag{9}$$

which in the limit $n \to \infty$ converges to $2\mathbb{E}[\rho(X, Y)] - \mathbb{E}[\rho(Y, Y)] - \mathbb{E}[\rho(X, X)]$. The same expressions hold for the MMD, with $\tilde{\rho}$ in place of ρ. For both ground distances, the latter quantity is zero if and only if X and Y have the same distribution. The cost of computing the ED and MMD is of order n^2. The article Park et al. (2016) proposes to use a variant of the MMD as a discrepancy between empirical distributions in ABC, and heuristics to choose the parameter s. A numerical comparison between ABC based on various distances is provided in Section 7.1.

Contrarily to the Hilbert and the swapping distances, the distances mentioned here either are not proper distances or do not upper bound the Wasserstein distance. This complicates the analysis of the associated ABC posterior distributions and minimum distance estimators.

6.5 Choosing and combining distances

Since we have theoretical support for ABC and minimum distance estimation with Hilbert, swapping and exact Wasserstein distances, we suggest to choose among these three according to the size of the data set. They can be calculated in a cost of order $n \log n$, n^2 and n^3 respectively. We compare the three distances in the numerical experiments of Section 7.1. In univariate settings, these three distances coincide, and we then refer to all as the Wasserstein distance (e.g. in Sections 7.2 and 7.3).

It might be useful to combine distances. For instance, one might want to start exploring the parameter space with the Hilbert distance, and switch to the exact Wasserstein distance in a region of interest; or use the Hilbert distance to save computations in a delayed acceptance scheme within ABC. One might also combine a transport distance with a distance between summaries. We can combine distances in the ABC framework by introducing a threshold for each distance, and define the ABC posterior as in Eq. (6), with a product of indicators corresponding to each distance.

In the minimum distance estimation setting, we can initialize the optimization of one distance with the minimum distance estimator based on another. The optimization could either be constrained to parameters yielding distances in the initial metric that are smaller than some threshold, or it could be unconstrained.

The first option might be particularly useful when combining transport distances with summaries. Other options include defining a new distance as a weighted average of two or more other distances. We explore the combination of distances in the numerical experiments of Section 7.4.

Any of the aforementioned distances can be computed faster by first sub-sampling $m < n$ points from $y_{1:n}$ and $z_{1:n}$, and then computing the distance between the resulting distributions. This increases the variance of the resulting distances, introducing a trade-off with computation time. In the case of the Wasserstein distance, this could be studied formally using the results of Sommerfeld and Munk (2017). Other multiscale approaches could also be used to accelerate computation (Mérigot, 2011).

7 Numerical experiments

As outlined in Section 2.3, the application of Monte Carlo Expectation-Maximization to approximate the MEWE is limited to models that generate synthetic data according to an equation of the form $z_{1:m} = g_m(u, \theta)$. We therefore choose to focus on the WABC approach in the numerical experiments, which does not make such assumptions. In the spirit of Monte Carlo optimization, the algorithms we use to approximate the WABC posterior could be modified to give approximations of the MEWE.

7.1 Quantile "g-and-k" distribution

We start by illustrating the proposed approach on an example where the likelihood can be approximated with high precision, which allows comparisons between the standard posterior and WABC approximations. We also compare the use of the Wasserstein distance with some of the other distances from Section 6.

A classical example used in the ABC literature is the g-and-k distribution (see e.g. Fearnhead and Prangle, 2012; Mengersen et al., 2013), whose quantile function is given by

$$r \in (0,1) \mapsto a + b\left(1 + 0.8\frac{1-\exp(-gz(r))}{1+\exp(-gz(r))}\right)\left(1+z(r)^2\right)^k z(r), \tag{10}$$

where $z(r)$ refers to the r-th quantile of the standard Normal distribution. Sampling from the g-and-k distribution can be done by plugging standard Normal variables into Eq. (10) in place of $z(r)$. We consider the bivariate extension of the g-and-k distribution (Drovandi and Pettitt, 2011), where one generates bivariate Normals with mean zero, variance one, and correlation ρ, and plugs them in place of $z(r)$ in Eq. (10), with parameters (a_i, b_i, g_i, k_i) for each component $i \in \{1, 2\}$. We generate $n = 500$ observations from the model using $a_1 = 3, b_1 = 1, g_1 = 1, k_1 = 0.5, a_2 = 4, b_2 = 0.5, g_2 = 2, k_2 = 0.4, \rho = 0.6$, as in Section 5.2 of Drovandi and Pettitt (2011). The parameters (a_i, b_i, g_i, k_i) are assigned a uniform prior on $[0, 10]$, for $i \in \{1, 2\}$, and ρ a uniform prior on $[-1, 1]$.

The probability density function is intractable but can be numerically calculated with high precision since it only involves one-dimensional inversions and differentiations of the quantile function of Eq. (10), as described in Rayner and MacGillivray (2002). Therefore, Bayesian inference can be carried out with e.g. Markov chain Monte Carlo. We run 8 Metropolis–Hastings chains for $100,000$ iterations to approximate the posterior distribution, and discard the first $10,000$ as burn-in. For the WABC approximation, we use $N = 1,024$ particles and run the sequential algorithm of Section 3.3 until 2×10^6 simulations from the model have been performed.

Figure 7 shows the marginal posterior and WABC posterior distributions of the nine parameters (Figure 7a-7i). The WABC approximation is accurate for the marginals of a_1, b_1, g_1, k_1, but less so for a_2, b_2, g_2, k_2.

In particular, the WABC marginal posterior of g_2 seems very similar to the prior. In Figure 7j, the last 10 WABC approximations are shown, overlaid with a horizontal line indicating the prior density. According to the plot, the WABC posterior of g_2 starts concentrating around the data-generating value (indicated by a vertical line), but more simulations from the model would be necessary to accurately approximate the posterior.

Figure 7k shows the $\mathfrak{W}_1$-distance between the WABC posterior samples, obtained with various distances, and a sample of $1,024$ points thinned out from the Markov chains targeting the posterior. This distance is plotted against the number of model simulations. It shows that all distances yield WABC posteriors getting closer to the actual posterior. On the other hand, for a finite number of model simulations, all WABC posteriors are significantly different from the actual posterior. For comparison, the $\mathfrak{W}_1$-distance between two samples of size $1,024$ thinned out from the Markov chains is on average 0.08. In this particular example, it appears that the MMD leads to ABC posteriors that are not as close to the posterior as the other distances. The Hilbert distance provides a particularly cheap and efficient approximation to the Wasserstein distance in this bivariate case.

For data sets of size $n = 500$ simulated using the data-generating parameter, the average wall-clock times to compute distances between simulated and observed data, on an Intel Core i7-5820K (3.30GHz), are as follows: $0.002s$ for the Hilbert distance, $0.01s$ for the MMD, $0.03s$ for the swapping distance, and $0.22s$ for the exact Wasserstein distance; these average times were computed on $1,000$ independent data sets. In this example, simulating from the model takes a negligible time, even compared to the Hilbert distance. Calculating the likelihood over $1,000$ parameters drawn from the prior, we find an average compute time of $0.05s$.

7.2 Toggle switch model

We borrow the system biology "toggle switch" model used in Bonassi et al. (2011); Bonassi and West (2015), inspired by studies of dynamic cellular networks. This provides an example where a sophisticated design of summaries can be replaced by the Wasserstein distance between empirical distributions. For $i \in 1:n$ and $t \in 1:T$, let $(u_{i,t}, v_{i,t})$ denote the expression levels of two genes in cell i at time t. Starting from $(u_{i,0}, v_{i,0}) = (10, 10)$, the evolution of $(u_{i,t}, v_{i,t})$ is given by

$$u_{i,t+1} = u_{i,t} + \alpha_1/(1 + v_{i,t}^{\beta_1}) - (1 + 0.03u_{i,t}) + 0.5\xi_{i,1,t},$$
$$v_{i,t+1} = v_{i,t} + \alpha_2/(1 + u_{i,t}^{\beta_2}) - (1 + 0.03v_{i,t}) + 0.5\xi_{i,2,t},$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are parameters, and ξ's are standard Normal variables, truncated so that $(u_{i,t}, v_{i,t})$ only takes non-negative values. For each cell i, we only observe a noisy measurement of the terminal expression level $u_{i,T}$. Specifically, the observations y_i are assumed to be independently distributed as Normal variables with mean $\mu + u_{i,T}$ and standard deviation $\mu\sigma/u_{i,T}^{\gamma}$, where μ, σ, γ are parameters. We generate $n = 2,000$ observations using $\alpha_1 = 22$, $\alpha_2 = 12$, $\beta_1 = 4$, $\beta_2 = 4.5$, $\mu = 325$, $\sigma = 0.25$, $\gamma = 0.15$. A histogram of the data is shown in Figure 8a.

We consider the task of estimating the data-generating values, using uniform prior distributions on $[0, 50]$ for α_1, α_2, on $[0, 5]$ for β_1, β_2, on $[250, 450]$ for μ, $[0, 0.5]$ for σ and on $[0, 0.4]$ for γ. These ranges are derived from Figure 5 in Bonassi and West (2015). Instead of using 11-dimensional tailor-made summaries as in Bonassi et al. (2011); Bonassi and West (2015), we use the Wasserstein distance with $p = 1$. The SMC sampler is run with $N = 2,048$, for a total number of 10^6 model simulations.

The seven marginal ABC posterior distributions, obtained for decreasing values of ε, are shown in Figure 8. We find that the marginal WABC posterior distributions concentrate at different rates depending on the parameters, similarly to the results of Bonassi and West (2015), Figure 5. Comparing the results, we see that the design of custom summaries can be by-passed thanks to the use of a distance between empirical distributions: the resulting posterior approximations are very similar, while our proposed approach is fully black-box and guaranteed to retrieve the exact posterior distribution in the limit of the number of model simulations.

7.3 Queueing model

We turn to the M/G/1 queueing model, which has appeared as a test case in the ABC literature, see e.g. Fearnhead and Prangle (2012). It provides an example where the data are dependent, but where the parameters can be identified from the marginal distribution of the data. In the model, customers arrive at a server with independent interarrival times w_i, exponentially distributed at rate θ_3. Each customer is served with independent service times u_i, taken to be uniformly distributed on $[\theta_1, \theta_2]$. We observe only the interdeparture times y_i, given by the process $y_i = u_i + \max\{0, \sum_{j=1}^{i} w_j - \sum_{j=1}^{i-1} y_j\}$. The prior on $(\theta_1, \theta_2 - \theta_1, \theta_3)$ is Uniform on $[0, 10]^2 \times [0, 1/3]$.

We use the data set given in Shestopaloff and Neal (2014), which was generated using the parameters $(\theta_1, \theta_2, \theta_3) = (4, 7, 0.15)$ and $n = 50$. The WABC posterior based on the empirical distribution of $y_{1:n}$ is approximated using the SMC algorithm of Section 3.3, with $N = 1,024$, for more than 3×10^7 model simulations. The actual posterior distribution is approximated with a particle marginal Metropolis–Hastings (PMMH) run (Andrieu et al., 2010), using $4,096$ particles and 10^5 iterations. The use of PMMH was suggested in Shestopaloff and Neal (2014), as an alternative to their proposed, model-specific Markov chain Monte Carlo algorithm.

Upon observing $y_{1:n}$, θ_1 has to be less than $\min_{i \in 1:n} y_i$, which is implicitly encoded in the likelihood, but not in an ABC procedure. One can add this constraint explicitly, rejecting parameters that violate it, which is equivalent to redefining the prior on θ_1 to be uniform on $[0, \min_{i \in 1:n} y_i]$. Figure 9 shows the marginal distributions of the parameters obtained with PMMH and with WABC, with or without the additional constraint. Overall, the WABC approximations are very close to the posterior, in comparison to the relatively vague prior distribution on (θ_1, θ_2). Furthermore, we see that incorporating the constraint helps estimating θ_1, without much effect on the other parameters. Similar results were found for the other two data sets considered in Shestopaloff and Neal (2014) (not shown).

7.4 Lévy-driven stochastic volatility model

We consider a Lévy-driven stochastic volatility model (e.g. Barndorff-Nielsen and Shephard, 2002), used in Chopin et al. (2013) as a challenging example of parameter inference in state space models. We demonstrate how ABC with transport distances can identify some of the parameters in a black-box fashion, and can be combined with summaries to identify the remaining parameters. The observation y_t at time t is the log-return of a financial asset, assumed Normal with mean $\mu + \beta v_t$ and variance v_t, where v_t is the actual volatility. Together with the spot volatility z_t, the pair (v_t, z_t) constitutes a latent Markov chain, assumed to follow a Lévy process. Starting with $z_0 \sim \Gamma(\xi^2/\omega^2, \xi/\omega^2)$ (where the second parameter is the rate), and

an arbitrary v_0, the evolution of the process goes as follows:

$$k \sim \mathcal{P}\text{oisson}\left(\lambda\xi^2/\omega^2\right), \quad c_{1:k} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(t, t+1), \quad e_{1:k} \overset{\text{i.i.d.}}{\sim} \mathcal{E}\text{xp}\left(\xi/\omega^2\right),$$
$$z_{t+1} = e^{-\lambda} z_t + \sum_{j=1}^{k} e^{-\lambda(t+1-c_j)} e_j, \quad v_{t+1} = \frac{1}{\lambda}[z_t - z_{t+1} + \sum_{j=1}^{k} e_j] \tag{11}$$

The random variables $(k, c_{1:k}, e_{1:k})$ are generated independently for each time period, and $1:k$ is the empty set when $k = 0$. The parameters are $(\mu, \beta, \xi, \omega^2, \lambda)$. We specify the prior as Normal with mean zero and variance 2 for μ and β, Exponential with rate 0.2 for ξ and ω^2, and Exponential with rate 1 for λ.

We generate synthetic data with $\mu = 0$, $\beta = 0$, $\xi = 0.5$, $\omega^2 = 0.0625$, $\lambda = 0.01$, which were used also in the simulation study of Barndorff-Nielsen and Shephard (2002); Chopin et al. (2013), of length $n = 10,000$. We use delay reconstruction with a lag $k = 1$, and the Hilbert distance $\mathfrak{H}_p$ of Section 6.2 with $p = 1$. Given the length of the time series, the cost of computing the Hilbert distance is much smaller than that of the other distances discussed in Section 6. The algorithm is run with $N = 1,024$ particles until 10^5 data sets have been simulated in total. Figure 10 shows the resulting quasi-posterior marginals for (μ, β), (ξ, ω^2), and λ. The parameters $(\mu, \beta, \xi, \omega^2)$ are accurately identified, from a vague prior to a region close to the data-generating values. On the other hand, the approximation of λ is barely different from the prior distribution. Indeed, the parameter λ represents a discount rate which impacts the long-range dependencies of the process, and is thus not captured by the bivariate marginal distribution of (y_t, y_{t-1}).

Hoping to capture long-range dependencies in the series, we define a summary $\eta(y_{1:n})$ as the sum of the first 50 sample autocorrelations among the squared observations. For each of the parameters obtained with the first run of WABC described above, we compute the summary of the associated synthetic data set. We plot the summaries against λ in Figure 11a. The dashed line indicates the value of the summary calculated on the observed data. It appears from the plot that the summaries closest to the observed summary are those obtained with the smallest values of λ. Therefore, we might be able to learn more about λ by combining the previous Hilbert distance with a distance between summaries.

Denote by $\mathfrak{H}_1(\tilde{y}_{1:n}, \tilde{z}_{1:n})$ the Hilbert distance between delay reconstructions, and by ε_h the threshold obtained after the first run of the algorithm. A new distance between data sets is defined as $|\eta(y_{1:n}) - \eta(z_{1:n})|$ if $\mathfrak{H}_1(\tilde{y}_{1:n}, \tilde{z}_{1:n}) < \varepsilon_h$, and $+\infty$ otherwise. We then run the SMC sampler of Section 3.3, starting from the result of the first run, and using the new distance. In this second run, a new threshold is introduced and adaptively decreased, keeping the first threshold ε_h fixed. One could also decrease both thresholds together or alternate between decreasing either. Note that the Hilbert distance and the summaries could have been combined in other ways, for instance in a weighted average.

We run the algorithm with the new distance for an extra 2×10^5 model simulations. Figures 11b and 11c show the evolution of the WABC posterior distributions of ω^2 and λ during the second run, using the summary. The WABC posteriors concentrate closer to the data-generating values, particularly for λ; for (μ, β, ξ), the effect is minimal and not shown. The WABC posterior could then be used to initialize a particle MCMC algorithm (Andrieu et al., 2010) targeting the posterior. The computational budget of 3×10^5 model simulations, as performed in total by the WABC procedure in this section, would be equivalent to 300 iterations of particle MCMC with $1,000$ particles at each iteration, in terms of number of model transitions. Therefore, the cost of initializing a particle MCMC algorithm with the proposed ABC approach is likely to be negligible. The gains could be considerable given the difficulty of initializing particle MCMC algorithms, mostly due to the large variance of the likelihood estimator for parameters located away from the posterior

mode. This is for instance illustrated in Figure 2 (c) of Murray et al. (2013).

8 Discussion

Using the Wasserstein distance in approximate Bayesian computation and minimum distance estimation leads to principled ways of inferring parameters in generative models, by-passing the choice of summaries. The approaches can also be readily used for deterministic models. We have demonstrated how the proposed ABC approach identifies high posterior density regions, in settings of multivariate (Section 7.1) and dependent data (Section 7.3). In the toggle switch model of Section 7.2, we have obtained posterior approximations similar to those obtained with sophisticated and case-specific summaries. Furthermore, we have shown how summaries and transport distances can be fruitfully combined in Section 7.4.

We have proposed multiple ways of defining meaningful empirical distributions of time series data, in order to identify model parameters. The proposed approaches have tuning parameters, such as λ in the curve matching approach of Section 4.1 or the lag k in delay reconstruction in Section 4.2. The choice of these parameters would deserve further research, leveraging the literature on Skorokhod distances for λ (Majumdar and Prabhu, 2015), and dynamical systems for k (Moeckel and Murray, 1997; Stark et al., 2003). Similar ideas could be explored in the setting of spatial data.

We have discussed some asymptotic properties of the minimum Wasserstein and minimum expected Wasserstein estimators, establishing existence, measurability and consistency under model misspecification and certain classes of dependent data, generalizing the results of Bassetti et al. (2006). The asymptotic distribution of the estimators can be derived in certain special cases, but deserves more research. One could also study the point estimators associated with the various distances of Section 6.

We have also established some properties of the WABC posterior distribution as the threshold goes to zero for a fixed number of observations n, and as n goes to infinity with a decreasing sequence ε_n. Our results on concentration rates highlight the impact of the order p of the Wasserstein distance, of model misspecification, and of the dimension of the observation space. These results add to the existing literature on asymptotic properties of ABC posteriors (Frazier et al., 2016; Li and Fearnhead, 2015). However, little is known about the ABC posterior for fixed ε. Viewing it as a coarsened posterior (Miller and Dunson, 2015), one can justify the use of the ABC posterior in terms of robustness to model misspecification. On the other hand, we do not claim that the WABC posterior for a fixed ε yields conservative statements about the actual posterior. For instance, Figure 5c shows that ABC posteriors can have little overlap with the actual posterior, for a fixed threshold ε, despite having shown signs of concentration away from the prior distribution.

As Wasserstein distance calculations scale super-quadratically with the number of observations n, we have introduced a new distance based on the Hilbert space-filling curve, computable in order $n \log n$, which can be used to initialize a swapping distance with a cost of order n^2. We have derived posterior concentration results for the ABC posterior distributions using the Hilbert and swapping distances, similarly to Theorem 5.4 obtained for the Wasserstein distance. Avenues of research include weakening the assumptions of these results, and investigating comparable results for the ABC posterior associated with the dual-Sinkhorn divergence and the Maximum Mean Discrepancy.

Acknowledgements We are grateful to Marco Cuturi, Jeremy Heng, Guillaume Pouliot and Neil Shephard for helpful feedback.

References

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser Verlag AG, Basel, second edition. 15

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo (with discussion). *Journal of the Royal Statistical Society: Series B*, 72(4):357–385. 24, 25

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. PMLR. 2

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B*, 64(2):253–280. 24, 25

Bassetti, F., Bodini, A., and Regazzini, E. (2006). On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302. 2, 4, 13, 26

Bassetti, F. and Regazzini, E. (2006). Asymptotic properties and robustness of minimum dissimilarity estimators of location-scale parameters. *Theory of Probability and its Applications*, 50(2):171–186. 7, 14, 16

Basu, A., Shioya, H., and Park, C. (2011). *Statistical inference: the minimum distance approach*. CRC Press. 4

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025. 1

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA. 11

Bonassi, F. V. and West, M. (2015). Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Analysis*, 10(1):171–187. 23, 24

Bonassi, F. V., You, L., and West, M. (2011). Bayesian learning from marginal data in bionetwork models. *Statistical applications in genetics and molecular biology*, 10(1). 23

Buchin, K., Buchin, M., and Wenk, C. (2008). Computing the Fréchet distance between simple polygons. *Computational Geometry*, 41(1-2):2–20. 11

Burkard, R., Dell'Amico, M., and Martello, S. (2009). *Assignment Problems*. Society for Industrial and Applied Mathematics (SIAM). 19

Cheney, E. W. and Wulbert, D. E. (1969). The existence and unicity of best approximations. *Mathematica Scandinavica*, 24:113–140. 16

Chopin, N., Jacob, P., and Papaspiliopoulos, O. (2013). SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B*, 75(3):397–426. 24, 25

Cuturi, M. (2013). Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2292–2300. 2, 21

Dede, S. (2009). An empirical central limit theorem in l^1 for stationary sequences. *Stochastic Processes and their Applications*, 119:3494 – 3515. 15

del Barrio, E., Giné, E., and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, pages 1009–1071. 15

Del Barrio, E., Giné, E., Utzet, F., et al. (2005). Asymptotics for l2 functionals of the empirical quantile process, with applications to tests of fit based on weighted wasserstein distances. *Bernoulli*, 11(1):131–189. 16

Del Barrio, E. and Loubes, J.-M. (2017). Central limit theorems for empirical transportation cost in general dimension. *arXiv preprint arXiv:1705.01299*. 5

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020. 9

Drovandi, C. C. and Pettitt, A. N. (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556. 22

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*, 74(3):419–474. 8, 22, 24

Forneron, J.-J. and Ng, S. (2015). The ABC of simulation estimation with auxiliary statistics. *arXiv preprint arXiv:1501.01265*. 1

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738. 17

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2016). Asymptotic properties of approximate Bayesian computation. *arXiv preprint arXiv:1607.06903*. 14, 16, 26

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3432–3440. 21

Gerber, M. and Chopin, N. (2015). Sequential quasi-Monte Carlo. *Journal of the Royal Statistical Society: Series B*, 77(3):509–579. 19, 20

Gottschlich, C. and Schuhmacher, D. (2014). The shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *PloS one*, 9(10):e110214. 19

Gouriéroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8:85–118. 1, 3, 5

Graham, M. and Storkey, A. (2017). Asymptotically exact inference in differentiable generative models. In *Artificial Intelligence and Statistics*, pages 499–508. 3

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054. 4

Kantz, H. and Schreiber, T. (1997). *Nonlinear time series analysis*. Cambridge University Press. 12

Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimators. *Annals of Mathematical Statistics*, 41:802–828. 16

Lee, A. (2012). On the choice of MCMC kernels for approximate Bayesian computation with SMC samplers. In *Proceedings of the 2012 Winter Simulation Conference*, pages 304–315. 9

Lee, A. and Łatuszyński, K. (2014). Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, 101(3):655–671. 9

Li, W. and Fearnhead, P. (2015). On the asymptotic efficiency of ABC estimators. *arXiv preprint arXiv:1506.03481*. 26

Majumdar, R. and Prabhu, V. S. (2015). Computing the Skorokhod distance between polygonal traces. In *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, pages 199–208. ACM. 11, 26

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180. 1

Mengersen, K. L., Pudlo, P., and Robert, C. P. (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, 110(4):1321–1326. 13, 22

Mérigot, Q. (2011). A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library. 22

Miller, J. W. and Dunson, D. B. (2015). Robust Bayesian inference via coarsening. *arXiv preprint arXiv:1506.06101*. 2, 9, 26

Moeckel, R. and Murray, B. (1997). Measuring the distance between time series. *Physica D*, 102:187–194. 12, 26

Montavon, G., Müller, K.-R., and Cuturi, M. (2016). Wasserstein training of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 3711–3719. 2, 5

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849. 19

Murray, L. M., Jones, E. M., and Parslow, J. (2013). On disturbance state-space models and the particle marginal Metropolis-Hastings sampler. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):494–521. 26

Muskulus, M. and Verduyn-Lunel, S. (2011). Wasserstein distances in the analysis of time series and dynamical systems. *Physica D*, 240:45–58. 12

Neath, R. C. et al. (2013). On convergence properties of the Monte Carlo EM algorithm. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, pages 43–62. Institute of Mathematical Statistics. 5

Owen, A. B. (2001). *Empirical likelihood*. CRC press. 4

Park, M., Jitkrittum, W., Sejdinovic, D., and Unit, G. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 398–407. 21

Parr, W. C. and Schucany, W. R. (1980). Minimum distance and robust estimation. *Journal of the American Statistical Association*, 75(371):616–624. 7

Pollard, D. (1980). The minimum distance method of testing. *Metrika*, 27:43–70. 15, 16

Prangle, D., Everitt, R. G., and Kypraios, T. (2017). A rare event approach to high-dimensional approximate Bayesian computation. *Statistics and Computing*. 3, 8

Puccetti, G. (2017). An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *Journal of Mathematical Analysis and Applications*, 451(1):132–145. 20

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 6

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer. 21

Ramdas, A., Trillos, N. G., and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47. 2, 21

Rayner, G. D. and MacGillivray, H. L. (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75. 22

Rubio, F. J., Johansen, A. M., et al. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654. 5

Sagan, H. (1994). *Space-filling curves*. Springer-Verlag New York. 2

Schretter, C., He, Z., Gerber, M., Chopin, N., and Niederreiter, H. (2016). Van der Corput and golden ratio sequences along the Hilbert space-filling curve. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 531–544. Springer. 19

Schuhmacher, D., Bähre, B., Gottschlich, C., and Heinemann, F. (2017). `transport: Optimal Transport in Various Forms`. R package version 0.8-2. 19

Shestopaloff, A. Y. and Neal, R. M. (2014). On Bayesian inference for the M/G/1 queue with efficient MCMC sampling. *arXiv preprint arXiv:1401.5548*. 24

Sommerfeld, M. and Munk, A. (2017). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B*. 2, 22

Sousa, V. C., Fritz, M., Beaumont, M. A., and Chikhi, L. (2009). Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics*, 181(4):1507–1519. 8

Srivastava, S., Cevher, V., Dinh, Q., and Dunson, D. (2015). WASP: Scalable Bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920. 2

Stark, J., Broomhead, D. S., Davies, M. E., and Huke, J. (2003). Delay embeddings for forced system: II. Stochastic forcing. *Journal of Nonlinear Science*, 13(6):519–577. 12, 26

The CGAL Project (2016). *CGAL: User and Reference Manual*. CGAL Editorial Board, 4.8 edition. 20

Villani, C. (2003). *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society. 3

Villani, C. (2008). *Optimal transport, old and new*. Springer-Verlag New York. 3

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704. 5

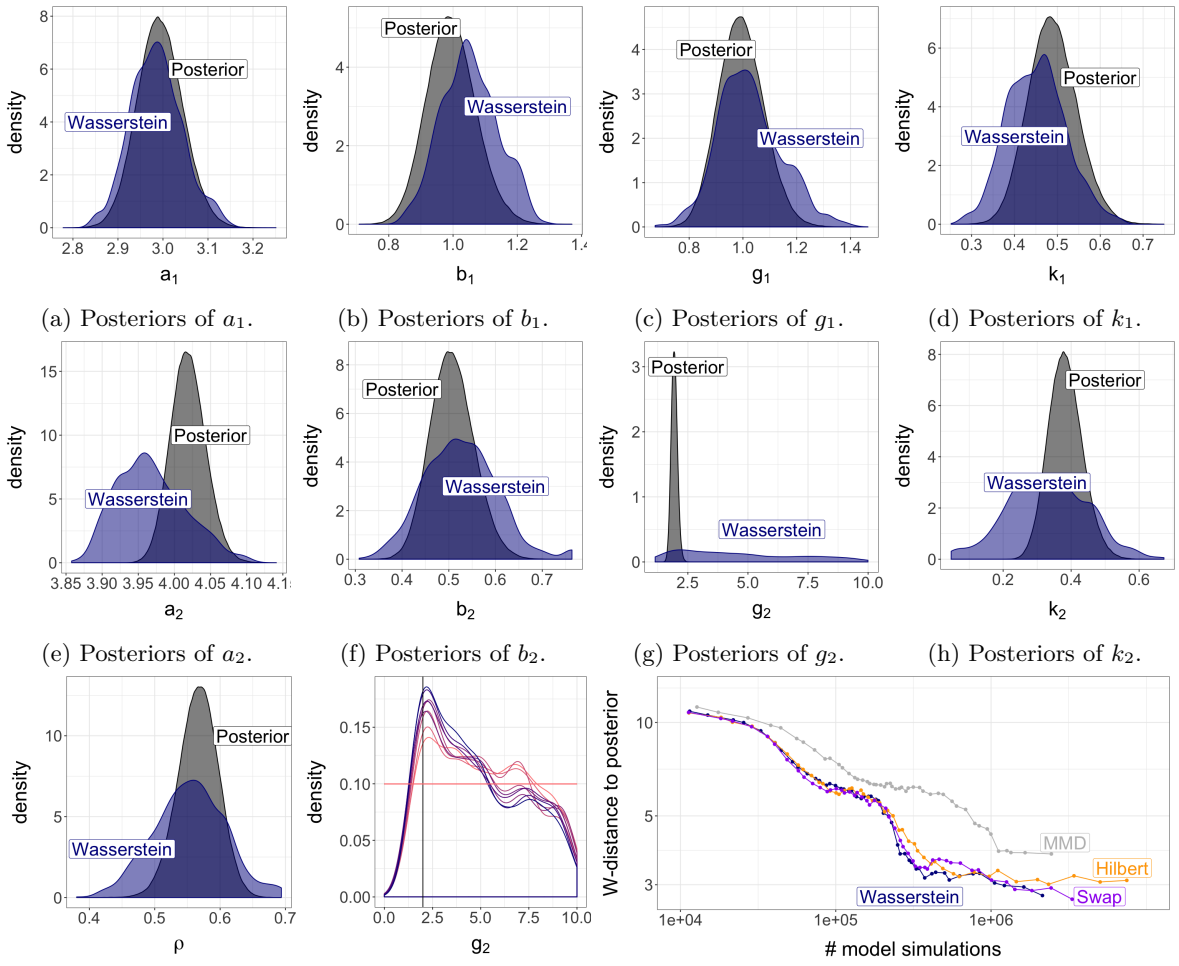
Wellner, J. A. and van der Vaart, A. W. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag New York. 16

Wolfowitz, J. (1957). The minimum distance method. *The Annals of Mathematical Statistics*, 28(1):75–88. 4

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104. 5

Wu, C. and Tabak, E. G. (2017). Statistical archetypal analysis. *arXiv preprint arXiv:1701.08916*. 2

Ye, J., Wang, J. Z., and Li, J. (2017). A simulated annealing based inexact oracle for Wasserstein loss minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3940–3948. PMLR. 21

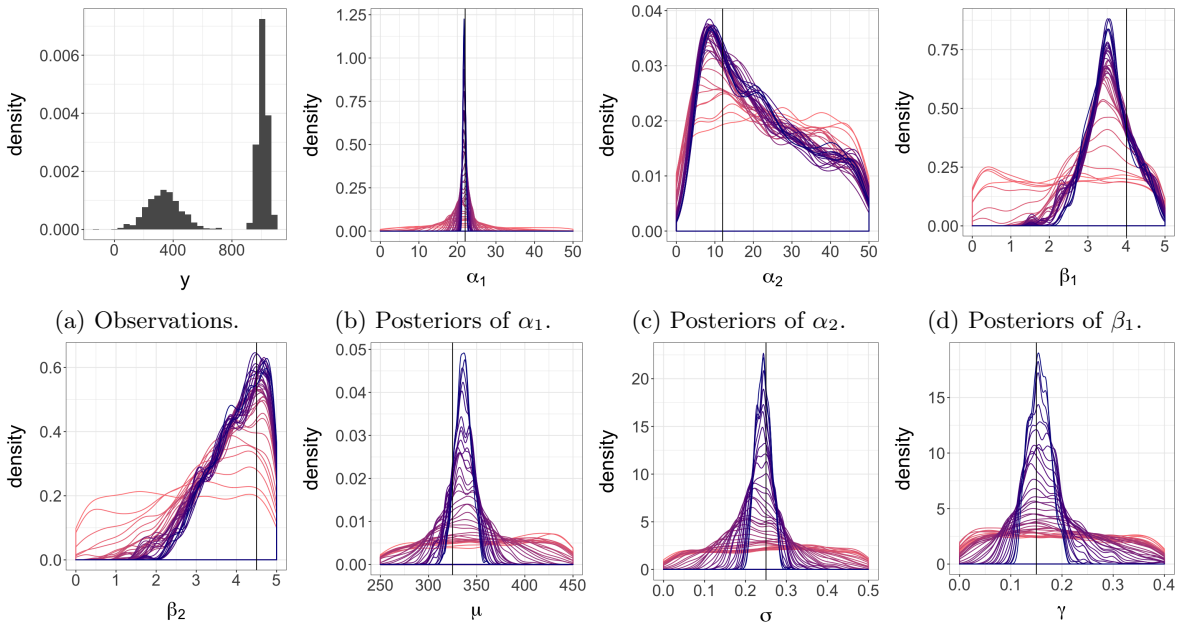


(a) Posteriors of a_1. (b) Posteriors of b_1. (c) Posteriors of g_1. (d) Posteriors of k_1.

(e) Posteriors of a_2. (f) Posteriors of b_2. (g) Posteriors of g_2. (h) Posteriors of k_2.

(i) Posteriors of ρ. (j) Posteriors of g_2 over last 10 steps. (k) $\mathfrak{W}_1$-distance to posterior sample, versus number of model simulations (log-log scale).

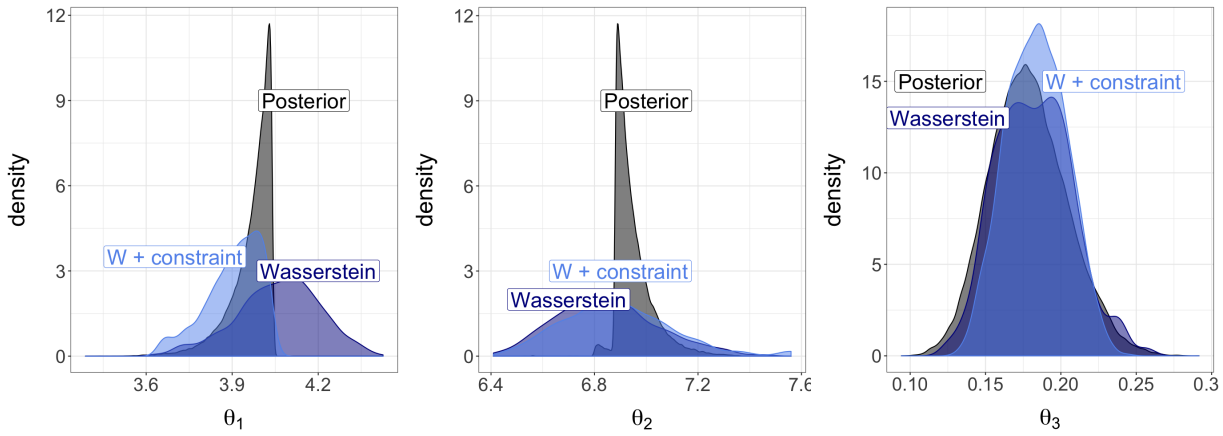
Figure 7: 7a-7i: posterior marginals in the g-and-k example of Section 7.1 (black, obtained via exact MCMC), and approximations by Wasserstein ABC (blue), for a budget of 2×10^6 model simulations. 7j: WABC posterior of g_2 over the last 10 steps of the algorithm, with prior density as horizontal line, and data-generating value as vertical line. 7k: $\mathfrak{W}_1$-distance between ABC posterior samples, obtained with various distances, and exact posterior samples against the number of model simulations (in log-log scale).



(a) Observations. (b) Posteriors of α_1. (c) Posteriors of α_2. (d) Posteriors of β_1.

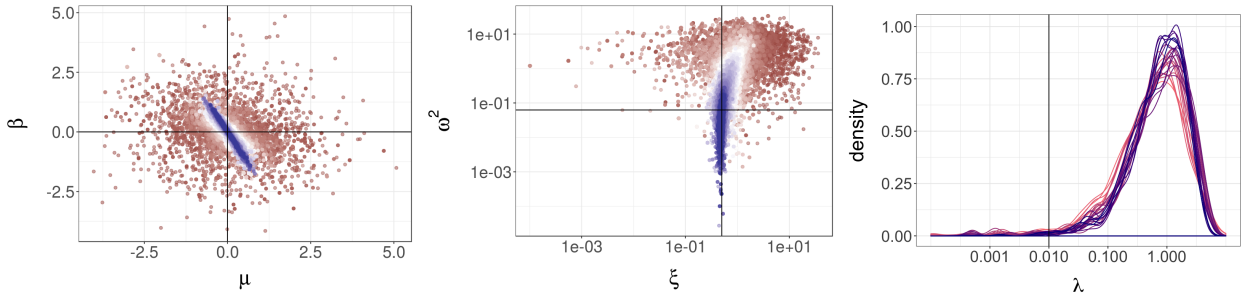
(e) Posteriors of β_2. (f) Posteriors of μ. (g) Posteriors of σ. (h) Posteriors of γ.

Figure 8: Histogram of observations (8a), and WABC marginal posteriors in the toggle switch model (all others). Data-generating values are indicated by vertical lines. Different values of ε, automatically obtained over 39 steps of the SMC sampler of Section 3.3, are indicated by different colored full lines (colors from light red to dark blue as ε decreases).



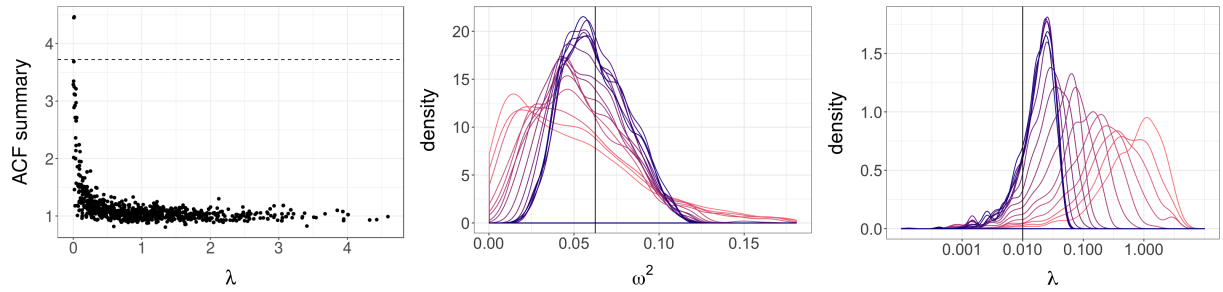
(a) Posteriors of θ_1. (b) Posteriors of θ_2. (c) Posteriors of θ_3.

Figure 9: ABC posteriors of parameters in queueing model. Fifty observations were generated with $\theta_1 = 4$, $\theta_2 = 7$ and $\theta_3 = 0.15$, indicated by vertical lines. The actual posterior is obtained by particle marginal Metropolis–Hastings. The ABC approach is run with the Wasserstein distance between the marginal empirical distributions of synthetic and observed data. The additional constraint that θ_1 has to be less than $\min_{i \in 1:n} y_i$ can be encoded in the prior ("W + constraint").



(a) Posteriors of (μ, β). (b) Posteriors of (ξ, ω^2) (in log-log scale). (c) Posteriors of λ (in log scale).

Figure 10: ABC approximations in the Lévy-driven stochastic volatility model, using the Hilbert distance between delay reconstructions with lag $k = 1$. The plots show the bivariate marginals of (μ, β) (left), (ξ, ω^2) (middle), and the marginal distributions of λ (right). These are obtained for up to 10^5 model simulations. Data-generating parameter values are indicated by full lines.



(a) Summary against λ. (b) Distributions of ω^2. (c) Distributions of λ (log-scale).

Figure 11: Left: summary, defined as the sum of the first 50 sample autocorrelations of the squared series, against λ, computed for the output of the WABC algorithm using the Hilbert distance between delay reconstructions. Middle and right: approximations of ω^2 and λ, from the second run of WABC using the Hilbert distance between delayed reconstructions combined with the summary. The horizontal axis in the right plot is on the log-scale, illustrating the significant concentration of the ABC posterior on the data-generating value of λ.