Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–28

Bayesian Nonparametric Model-based Clustering with Intractable Distributions: An ABC Approach

Mario Beraha* and Riccardo Corradin†

Abstract. Bayesian nonparametric mixture models offer a rich framework for model-based clustering. We consider the situation where the kernel of the mixture is available only up to an intractable normalizing constant. In this case, the most commonly used Markov chain Monte Carlo (MCMC) methods are unsuitable. We propose an approximate Bayesian computational (ABC) strategy, whereby we approximate the posterior to avoid the intractability of the kernel. We derive an ABC-MCMC algorithm which combines (i) the use of the predictive distribution induced by the nonparametric prior as proposal and (ii) the use of the Wasserstein distance and its connection to optimal matching problems. To overcome the sensibility concerning the parameters of our algorithm, we further propose an adaptive strategy. We illustrate the use of the proposed algorithm with several simulation studies and an application on real data, where we cluster a population of networks, comparing its performance with standard MCMC algorithms and validating the adaptive strategy.

Keywords: approximate Bayesian computation, Markov chain Monte Carlo, adaptive sampling scheme, Bayesian nonparametric, Wasserstein distance, mixture models.

1 Introduction

For a generic dataset, cluster analysis consists of identifying a meaningful partition of observations into *homogeneous* clusters, that is, groups for which data in the same group are more similar than data in different groups. Clustering is a valuable tool in analyzing complex data as it allows for exploring the variability in a dataset and can constitute an effective pre-processing tool for downstream tasks. In a Bayesian model-based approach, clustering is performed by assuming a mixture likelihood for the data, where observations in each cluster are assumed i.i.d. from a (typically parametric) kernel density, $\mathcal{K}(\cdot \mid \theta)$ for some value of the parameters $\theta \in \Theta$. Then, the homogeneity of a cluster means that observations in that cluster are suitably modeled by $\mathcal{K}(\cdot \mid \theta)$. See, e.g. Frühwirth-Schnatter et al. (2019) and the references therein for a recent overview of Bayesian model-based clustering. The choice of $\mathcal{K}$ plays a crucial role in the interpretability of the clustering. In several real-world applications, the model assumed for the data-generating process leads to $\mathcal{K}$ being intractable, that is, impossible to evaluate analytically. Such intractability is due, for instance, to the presence of latent variables

*Department of Mathematics, Politecnico di Milano, Italy, mario.beraha@polimi.it
†School of Mathematical Sciences, University of Nottingham, Nottingham, UK, riccardo.corradin@nottingham.ac.uk

 https://doi.org/10.1214/24-BA1416

integrated out of the model or to the use of a physical model involving differential equations. Specific instances, addressed later in the paper, include stochastic volatility models for time series and the exponential random graph distribution for networks.

Approximate Bayesian computation (ABC) is a recent growing area of research dealing with statistical problems involving intractable distributions, i.e. distributions known up to a normalizing constant which is parameter-dependent (doubly-intractable distributions) or for which evaluating the probability density function is computationally prohibitive. We refer to the pioneering studies of these methodologies by mentioning the works of Rubin (1984); Tavaré et al. (1997); Pritchard et al. (1999); Beaumont et al. (2002). See also Sisson et al. (2018) and Karabatsos and Leisen (2018) for recent reviews. In the ABC setting, the direct application of Bayes' theorem to obtain the posterior distribution is infeasible due to the intractability of the likelihood. ABC strategies deal with this issue by introducing an approximation of the original posterior distribution, whereby the evaluation of the likelihood function is replaced by the evaluation of the distance between the observed data and a synthetic dataset generated from the model (or a surrogate) given parameters' values. The true and synthetic data are considered close if their distance is smaller than threshold ε, where both the distance and the threshold are specified by the user and problem-dependent. Intuitively, if the true and synthetic data are close, the parameters used to generate the synthetic dataset should be informative about the posterior distribution of parameters, given the true dataset. An approximation of the true posterior can then be constructed by considering the values of the parameters leading to synthetic datasets similar to the observed one. Therefore, ABC strategies require that simulating synthetic data is possible and feasible in a reasonable time.

The application of ABC methods spreads over many fields. Remarkable examples are recent usages in astronomy and cosmology (e.g. Cameron and Pettitt, 2012; Weyant et al., 2013), genetics (e.g. Beaumont and Rannala, 2004; Technow et al., 2015) and finance (e.g. Picchini, 2014; Calvet and Czellar, 2014), among others. Many ABC methods and extensions have been proposed in the literature over the last few decades, mainly by considering different strategies to approximate the posterior distribution, such as rejection sampler (e.g. Pritchard et al., 1999; Beaumont et al., 2002) and kernel methods (e.g. Beaumont et al., 2002; Wilkinson, 2013) among others. These strategies can be further combined with various standard computational methods, obtaining, for example, ABC rejection sampler (e.g. Tavaré et al., 1997; Pritchard et al., 1999), ABC importance sampler and sequential Monte Carlo (e.g. Fearnhead and Prangle, 2012; Sisson et al., 2007, 2009; Beaumont et al., 2009), ABC Markov chain Monte Carlo (e.g. Marjoram et al., 2003; Bortot et al., 2007), and ABC Variational Inference (e.g. Barthelmé and Chopin, 2014).

This work studies the Bayesian model-based clustering approach when the mixture kernel $\mathcal{K}(\cdot \mid \theta)$ is intractable. Standard techniques for estimating the posterior distribution are based on Markov chain Monte Carlo (MCMC) algorithms, which become either impractical or impossible in this case. We propose an ABC-MCMC algorithm (Marjoram et al., 2003) to sample from an approximation of the true posterior distribution of interest. The main quantities to define such an approximate strategy are the proposal

distribution of the MCMC scheme, the choice of a distance to compare observed and synthetic data, and a threshold. As far as the proposal is concerned, we exploit the predictive law induced by the *exchangeable partition probability function* (cf. Section 2.1) of the mixing measure, which ensures that, if the distance between observed and synthetic data is smaller than the threshold, we accept the proposed values with probability one. To compare two datasets, we employ the Wasserstein metric between the empirical probability distributions. See, e.g., Villani (2008) for an overview of foundations and theoretical results and Peyré et al. (2019) for the computational aspects. The primary motivation for this choice comes from the geometry of the underlying problem, i.e. partitions' estimation, and its connection with optimal transport. Recent attention was given in the literature to combining Wasserstein distance with ABC procedures. See, for example, Bernton et al. (2019b). Compared to previous approaches, where the Wasserstein distance was used as a mean to avoid summary statistics, here we make use of the optimal transport map to perform efficient inference on the latent partition of the observed data, starting from the latent partition of synthetic data. Recently, Nguyen et al. (2022) proposed a similar idea to define a coupling between Markov chains on the space of partitions, in the context of parallel MCMC for mixture models. Finally, we propose an adaptive strategy for the threshold, which improves the sampler's performance while simplifying its specification.

We validate our ABC-MCMC strategy through several simulated examples: when data are univariate, we consider mixtures of Gaussian and g-and-k distributions, showing that with an intractable kernel our approach yields better performance in terms of accuracy of the cluster detection, runtime and effective sample size, when compared to standard MCMC algorithms. Specifically, in the case of g-and-k distribution, we approximate the density numerically when running the standard MCMC sampler. When data are multivariate, we consider mixtures of bivariate g-and-k distributions and mixtures of the Lévy driven stochastic volatility model (Barndorff-Nielsen and Shephard, 2002), showing that our approach recovers the ground-truth clustering. Finally, we consider mixtures of exponential random graph densities and apply the model to cluster similar US air companies based on their connections among airports. In this case, each observation is represented by a network.

The paper is structured as follows: Section 2 reviews mixture modelling, latent random partition, intractable kernel distributions, and some results fundamental to the following sections. Section 3 introduces the ABC-MCMC sampling strategy for latent random partitions in mixture models in a general setting and discusses the use of an adaptive strategy for the rejection threshold. In Section 4, we present numerical illustrations where we compare our ABC-MCMC algorithm with standard MCMC samplers based on Gibbs sampling, demonstrating the usefulness of the adaptive threshold selection strategy. We conclude the paper with some final comments and remarks. Proofs of main results are deferred to the Supplementary Material (Beraha and Corradin, 2024). All the routines used for the analyses and simulations presented in the manuscript are available at `https://github.com/mberaha/abc_partition`.

2 Bayesian mixture models

Consider observations $\boldsymbol{y}_{1:n} = (y_1, \dots, y_n)$ such that each y_i belongs to a Polish space $(\mathbb{Y}, \mathcal{Y})$, for $i = 1, \dots n$. We will always assume the Borel σ-field and skip measure-theoretic details in the following. A possible way to account for sources of heterogeneity in the observed data $\boldsymbol{y}_{1:n}$ is to consider a mixture model specified through a mixing distribution $\tilde{p}$. We then assume that observations are i.i.d. conditionally on the mixing measure $\tilde{p}$, with

$$y_1, \dots, y_n \,|\, \tilde{p} \overset{\text{iid}}{\sim} \tilde{f}(\cdot) = \int_{\Theta} \mathcal{K}(\cdot\,; \theta)\tilde{p}(d\theta), \tag{1}$$

where $\mathcal{K}(\cdot, \cdot)$ is measurable in its two arguments, $\mathcal{K}(\cdot, \theta)$ is a probability density function for each value of $\theta \in \Theta$, and $\tilde{p}$ is an almost surely discrete random probability measure, i.e., $\tilde{p} \overset{a.s.}{=} \sum_h w_h \delta_{\theta_h^*}$ with both the weights w_h's and the atoms θ_h^*'s random quantities. Note that the number of components in $\tilde{p}$ can be either finite or infinite, depending on specific modelling choices. We further assume the distribution of the weights w_h's independent of the distribution of the locations θ_h^*'s, where the latter is usually assumed diffuse on Θ.

Although our methodology is valid regardless of the specific choice of $\mathcal{K}$, it is suited, in particular, to deal with cases when $\mathcal{K}$ is not analytically available. For instance, $\mathcal{K}$ could depend on the numerical solution of a differential equation, involve latent variables that are marginalized out, or simply be known up to an intractable normalizing constant depending on the parameters. In cases when $\mathcal{K}$ is known explicitly, a variety of efficient algorithms to fit mixture models have been proposed in the literature.

We can rewrite model (1) in a hierarchical fashion by assuming that $y_i \,|\, \theta_i \overset{\text{ind}}{\sim} \mathcal{K}(\cdot; \theta_i)$ and $\theta_i \,|\, \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$, $i = 1, \dots, n$. In particular, the sequence $\theta_1, \dots, \theta_n$ is exchangeable and, due to the almost sure discreteness of $\tilde{p}$, there is a positive probability of having ties among the θ_i's which identify the clusters. Exchangeability is paramount in designing our ABC-MCMC algorithm, cf. Section 3.2 below. Moreover, by the de Finetti's theorem (de Finetti, 1937), exchangeability further motivates the assumption of a prior Q for $\tilde{p}$. We highlight two possible specifications for Q below.

Example 1 (Pitman-Yor process mixture model)**.** *The Pitman-Yor process (Pitman and Yor, 1997; Ishwaran and James, 2001) is a popular nonparametric prior for Bayesian mixture model and species sampling problems. We write* $\tilde{p} \sim \text{PY}(\vartheta, \sigma, G_0)$*, where* $\sigma \in [0, 1)$*,* $\vartheta > -\sigma$ *and* G_0 *is a diffuse probability measure on* $\mathbb{Y}$*. Then,* $\tilde{p} = \sum_{h=1}^{\infty} w_h \delta_{\theta_h^*}$ *with* $\theta_1^*, \theta_2^*, \dots \overset{iid}{\sim} G_0$ *and* $\{w_h\}_h$ *is a sequence of weights distributed according to a two-parameter Griffiths–Engen–McCloskey distribution, i.e.,* $w_1 = \nu_1$*,* $w_h = \nu_h \prod_{j \geq h}(1 - \nu_j)$ *for* $h > 1$*, with* $\nu_h \overset{iid}{\sim} \text{Beta}(1 - \sigma, \vartheta + h\sigma)$*.*

Example 2 (Mixture of finite mixtures)**.** *Introduced in Gnedin and Pitman (2006) and recently popularized by Miller and Harrison (2018), the mixture of finite mixtures (MFM) assumes* $\tilde{p} = \sum_{h=1}^{m} w_h \delta_{\theta_h^*}$ *where* $w_h \,|\, m \sim \text{Dir}_m(\alpha)$*, i.e. the symmetric Dirichlet distribution on the* $(m-1)$*-dimensional simplex, the* θ_i^**'s are independent and identically distributed from a diffuse probability measure* G_0*, and* $m \sim \pi(m)$*. Specifically, we will consider the case* $(m - 1) \sim \text{Poi}(\lambda)$*.*

2.1 Exchangeable random partitions and mixture models

For our purposes, it is easier to think of a mixture model in terms of a latent partition and a set of cluster-specific parameters. Specifically, let $[n] = \{1, \dots, n\}$, denote with $\boldsymbol{\rho}_n = \{A_1, \dots, A_k\}$ a random partition of $[n]$ (i.e. $\bigcup_j A_j = [n]$ and $A_i \cap A_j = \emptyset$ if $i \neq j$) and let $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_k^*)$. Writing $p(\cdot)$ for a generic density and $p(\cdot \,|\, \cdot)$ for a conditional density, we assume that

$$p(y_1, \dots, y_n \,|\, \boldsymbol{\theta}^*, \boldsymbol{\rho}_n) = \prod_{j=1}^{k} \prod_{i \in A_j} \mathcal{K}(y_i; \theta_j^*). \tag{2}$$

The Bayesian approach requires specifying a prior distribution for $(\boldsymbol{\rho}_n, \boldsymbol{\theta}^*)$. We assume that $\boldsymbol{\rho}_n$ is independent of $\boldsymbol{\theta}^*$ and that conditionally on the number of elements of the partition k (henceforth denoted as clusters), $\theta_1^*, \dots, \theta_k^*$ are independent and identically distributed random variables from a diffuse probability distribution G_0 that does not depend on $(\boldsymbol{\rho}_n, k)$, so that $\theta_i^* \neq \theta_j^*$ for $i \neq j$ almost surely. For the class of distributions considered here (see below), the representations in (1) and (2), together with prior assumptions, are indeed equivalent. In particular, (2) can be derived from (1) by marginalizing out $\tilde{p}$. See, for instance, James et al. (2009) and the references therein.

As far as the prior for $\boldsymbol{\rho}_n$ is concerned, we only require for our algorithm (cf. Section 3.2) that the law of the random partition allows for explicit formulas for $\mathrm{P}(\boldsymbol{\rho}_{n+1} \,|\, \mathbf{r}_n)$ where $\mathbf{r}_n = \{A_i\}_{i=1}^k$ is a partition of $[n]$ and $\boldsymbol{\rho}_{n+1}$ denotes the random partition of $[n+1]$. In particular, $\mathrm{P}(\boldsymbol{\rho}_{n+1} \,|\, \mathbf{r}_n)$ is the law of the extension of $\mathbf{r}_n$ to $[n+1]$, i.e. the partition of $[n+1]$ conditionally to the first n elements being partitioned according to $\mathbf{r}_n$.

A well studied class of random partitions is the one of *exchangeable* partitions (Kingman, 1978; Pitman, 1995), for which $\mathrm{P}(\boldsymbol{\rho}_n = \mathbf{r}_n)$ depends on $\mathbf{r}_n = \{A_i\}_{i=1}^k$ only through the cardinalities $n_j = |A_j|$ of each set A_j and k. Further, there exists a symmetric function $\mathrm{p}_{\mathrm{k}}^{(\mathrm{n})}(n_1, \dots, n_k)$ named *exchangeable partition probability function* (EPPF) such that

$$\mathrm{P}(\boldsymbol{\rho}_n = \mathbf{r}_n) = \int_{\Theta^k} \mathbb{E}_{\tilde{p}} \left[\prod_{j=1}^{k} \tilde{p}^{n_j}(\mathrm{d}\theta_j^*) \right] = \mathrm{p}_{\mathrm{k}}^{(\mathrm{n})}(|A_1|, \dots, |A_k|) = \mathrm{p}_{\mathrm{k}}^{(\mathrm{n})}(n_1, \dots, n_k),$$

where the first equality in the previous equation highlights the connection between (1) and (2). From the EPPF it is straightforward to derive the predictive law $\mathrm{P}(\boldsymbol{\rho}_{n+1} \,|\, \mathbf{r}_n)$ as

$$\begin{aligned} \mathrm{P}(\boldsymbol{\rho}_{n+1} = \{A_1, \dots, A_j \cup \{n+1\}, \dots, A_k\} \,|\, \mathbf{r}_n) &= \frac{\mathrm{p}_{\mathrm{k}}^{(\mathrm{n}+1)}(n_1, \dots, n_j + 1, \dots, n_k)}{\mathrm{p}_{\mathrm{k}}^{(\mathrm{n})}(n_1, \dots, n_k)}, \\ \mathrm{P}(\boldsymbol{\rho}_{n+1} = \{A_1, \dots, A_k, \{n+1\}\} \,|\, \mathbf{r}_n) &= \frac{\mathrm{p}_{\mathrm{k}+1}^{(\mathrm{n}+1)}(n_1, \dots, n_k, 1)}{\mathrm{p}_{\mathrm{k}}^{(\mathrm{n})}(n_1, \dots, n_k)}. \end{aligned} \tag{3}$$

Example 3. *(Pitman-Yor process mixture model (continued)) the EPPF of a PY process can be explicitly characterized as*

$$\mathrm{p}_{\mathrm{k}}^{(\mathrm{n})}(n_1,\dots,n_k) = \frac{\prod_{j=1}^{k-1}(\vartheta + j\sigma)}{(\vartheta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1},$$

where $(x)_n = x(x+1)\cdots(x+n-1)$ *denotes the Pochhammer symbol. Moreover it is straightforward to derive simpler expressions for the probabilities in* (3)

$$\begin{aligned} \mathrm{P}(\boldsymbol{\rho}_{n+1} = \{A_1,\dots,A_j\cup\{n+1\},\dots,A_k\} \,|\, \mathbf{r}_n) &\propto n_j - \sigma \\ \mathrm{P}(\boldsymbol{\rho}_{n+1} = \{A_1,\dots,A_k,\{n+1\}\} \,|\, \mathbf{r}_n) &\propto \vartheta + k\sigma, \end{aligned}$$

Example 4. *(MFM (continued)) the predictive probabilities for a MFM process satisfy*

$$\begin{aligned} \mathrm{P}(\boldsymbol{\rho}_{n+1} = \{A_1,\dots,A_j\cup\{n+1\},\dots,A_k\} \,|\, \mathbf{r}_n) &\propto n_j + \alpha \\ \mathrm{P}(\boldsymbol{\rho}_{n+1} = \{A_1,\dots,A_k,\{n+1\}\} \,|\, \mathbf{r}_n) &\propto \frac{V_n(k+1)}{V_n(k)}\alpha, \end{aligned}$$

where the weights $V_n(k)$*s are defined as*

$$V_n(k) = \sum_{\ell\geq 1} \frac{\Gamma(\ell+1)\Gamma(\gamma k)}{\Gamma(\ell-k+1)\Gamma(\gamma k+n)} \psi(k) \mathbb{1}_{[\ell<k]},$$

and the weights can be computed recursively using $V_{n+1}(k+1) = V_n(k)/\alpha - (n/\alpha + k)V_{n+1}(k)$*. See Miller and Harrison (2018) for further details.*

Other well-known EPPFs are the Ewens sampling formula (Blackwell and MacQueen, 1973), which corresponds to Example 3 when $\sigma = 0$, and the ones induced by Gibbs-type priors (Gnedin and Pitman, 2006; De Blasi et al., 2013), which include both our examples as special cases. Enriching the predictive structure of Gibbs-type EPPFs while maintaining the analytical tractability is an active area of research, see, e.g., Camerlenghi et al. (2023).

2.2 Dealing with intractable kernel density functions

Traditionally employed MCMC algorithms for fitting Bayesian mixture models require that $\mathcal{K}$ is known in closed form. To understand the issues arising with intractable kernels and motivate our approach, let us focus here on the case when $\mathcal{K}$ is known up to a normalizing constant, i.e. $\mathcal{K}(\cdot;\theta) = g(y_i;\theta)/Z_\theta$, where Z_θ is an intractable normalizing constant depending on the parameters. Algorithm 1 reports the celebrated algorithm 2 in Neal (2000), one of the cornerstone MCMC algorithms for Bayesian mixture models. Both updates present nontrivial challenges. Step (A) of Algorithm 1 requires sampling from a so-called doubly intractable distribution. Assuming that a perfect simulation algorithm from $\mathcal{K}$ is available, sampling from $p(\theta_h \,|\, \cdots)$ can be performed through an exchange algorithm as the one in Møller et al. (2006). However, as pointed out in Murray et al. (2006), the exchange algorithm can lead to low acceptance rates and a better

Algorithm 1: Neal's Algorithm 2 (Neal, 2000).

input a set of data $\boldsymbol{y}_{1:n}$

set admissible initial values for $\boldsymbol{\theta}^{*(0)}_{1:k}$ and $\boldsymbol{\rho}_n$;

for $r = 1, \dots, R$ **do**

(A) **for** $h = 1, \dots, k$ **do**

sample each $\theta_h^{*(r)}$ independently from

$$p(\theta_h^{*(r)} \mid \cdots) \propto \prod_{i \in A_h} \mathcal{K}(y_i; \theta_h^*) G_0(\theta_h^{*(r)}) = \left(Z_{\theta_h^{*(r)}}\right)^{-n_h} \prod_{i \in B_h} g(y_i; \theta_h^{*(r)}) G_0(\theta_h^*)$$

(B) **for** $i = 1, \dots, n$ **do**

update the cluster allocation of each observation sampling from

$$\mathrm{P}(i \in A_h \mid \cdots) \propto \begin{cases} \mathrm{p}^{(\mathrm{n})}_{\mathrm{k}}(n_1^{-i}, \dots, n_h^{-i} + 1, \dots n_k^{-i}) g(y_i; \theta_h^{*(r)}) / Z_{\theta_h^{*(r)}} & : h = 1, \dots, k \\ \mathrm{p}^{(\mathrm{n})}_{\mathrm{k}}(n_1^{-i}, \dots n_k^{-i}, 1) \int_\Theta g(y_i; \theta) / Z_\theta p(d\theta) & : h = k+1 \end{cases}$$

where the superscript $-i$ means that the i-th observation has been removed from the calculations.

end

solution would be to employ a sequence of tempered transitions, which still requires nontrivial implementations and fine-tuning. Step (B) involves a distribution over the integers $\{1, \dots, k+1\}$. The probability associated with $k+1$ involves an integral, but this can be overcome by using, for instance, Neal's Algorithm 8. Hence, for the sake of argument, let us ignore the last term. Usually, one computes the unnormalized probabilities, normalizes them and samples from the resulting discrete probability distribution. However, each term also contains Z_{θ_h}, which is unknown in this case, so this simple strategy is not possible. One could instead employ a Metropolis-Hastings step with a proposal over $\{1, \dots, k+1\}$, which would require again the use of some form of the exchange algorithm to get rid of the ratios of normalizing constants. In summary, the presence of an intractable normalizing constant in $\mathcal{K}$ severely impacts the feasibility of commonly used MCMC algorithms for mixture models and presents a major bottleneck for efficiency.

3 Approximate inference for random partitions

By applying Bayes' theorem, the posterior of the partition $\boldsymbol{\rho}_n$ given data $\boldsymbol{y}$ can be written as

$$\pi(\boldsymbol{\rho}_n \mid \boldsymbol{y}_{1:n}) = \frac{p(\boldsymbol{\rho}_n) p(\boldsymbol{y}_{1:n} \mid \boldsymbol{\rho}_n)}{\sum_{\boldsymbol{\rho}_n \in \mathcal{P}_{1:n}} p(\mathrm{d}\boldsymbol{\rho}_n) p(\boldsymbol{y}_{1:n} \mid \boldsymbol{\rho}_n)}, \tag{4}$$

where $\mathcal{P}_{1:n}$ is the space of all possible partitions of n elements, $p(\boldsymbol{\rho}_n)$ is the prior distribution, and

$$p(\boldsymbol{y}_{1:n} \mid \boldsymbol{\rho}_n) = \prod_{j=1}^{k} \int \prod_{i \in A_j} \mathcal{K}(y_i; \theta) G_0(\mathrm{d}\theta) = \prod_{j=1}^{k} \int \prod_{i \in A_j} \frac{g(y_i; \theta)}{Z_\theta} G_0(\mathrm{d}\theta).$$

To overcome the analytical intractability of the mixture kernel, we propose to consider an approximation of the posterior in (4), namely π_ε defined as:

$$\pi_\varepsilon(\boldsymbol{\rho}_n \mid \boldsymbol{y}_{1:n}) = \frac{p(\boldsymbol{\rho}_n) \int_{\mathbb{Y}^n} \mathbb{1}_{[d(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n}) < \varepsilon]} p(d\boldsymbol{s}_{1:n} \mid \boldsymbol{\rho}_n)}{\sum_{\boldsymbol{\rho}_n \in \mathcal{P}_{1:n}} p(\mathrm{d}\boldsymbol{\rho}_n) \int_{\mathbb{Y}^n} \mathbb{1}_{[d(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n}) < \varepsilon]} p(d\boldsymbol{s}_{1:n} \mid \boldsymbol{\rho}_n)}, \tag{5}$$

where $d : \mathbb{Y}^n \times \mathbb{Y}^n \to [0, +\infty)$ is a metric (specific choices will be discussed later) and $p(d\boldsymbol{s}_{1:n} \mid \boldsymbol{\rho}_n)$ is the distribution of a *synthetic* dataset generated from (2) conditional on the partition $\boldsymbol{\rho}_n$. In particular (5) is an ABC posterior as in Equation (1.5) of Sisson et al. (2018), where the *ABC kernel* is $\mathbb{1}_{[d(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n}) < \varepsilon]}$.

Algorithm 2: ABC rejection sampling for random partitions.

input a set of data $\boldsymbol{y}_{1:n}$
for $r = 1, \ldots, R$ **do**
  repeat
    sample a partition $\widetilde{\boldsymbol{\rho}}_n^{(r)} = \{A_j\}_{j=1}^k$ from the prior.
    sample $\theta_j^{*(r)} \overset{\text{iid}}{\sim} G_0$ $j = 1, \ldots, k$ and $\{s_i\}_{i \in A_j} \mid \theta_j^{*(r)} \overset{\text{iid}}{\sim} \mathcal{K}(\cdot; \theta_j^{*(r)})$.
  until $d(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n}) < \varepsilon$;
end

Many different techniques can be considered to obtain a sample from (5). A basic acceptance-rejection ABC algorithm can be straightforwardly derived. Although it is not the one we will employ (see Section 3), it is instructive to report it in Algorithm 2 for the discussion below. First, we note that the distance $d(\cdot, \cdot)$ has not been specified yet. Traditionally, ABC algorithms employed statistics $\nu : \mathbb{Y}^n \to \mathbb{R}^d$ and considered $d(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n}) = \|\nu(\boldsymbol{y}_{1:n}) - \nu(\boldsymbol{s}_{1:n})\|$, where $\|\cdot\|$ is the Euclidean metric on $\mathbb{R}^d$. For instance, ν could compute the mean and variance of $\boldsymbol{y}_{1:n}$. Using summary statistics simplifies the computations as it allows for great dimensionality reduction but also causes a loss of information. Further, the choice of which summary statistics to use is not obvious (Fearnhead and Prangle, 2012). More recently, the use of statistical distances to compare the empirical distributions of $\boldsymbol{y}_{1:n}$ and $\boldsymbol{s}_{1:n}$ has been proposed to overcome the issues related to summarization. See, for instance, Drovandi and Frazier (2022) and the references therein. Moreover, another issue is evident when inspecting the output of the acceptance-rejection algorithm: the partition $\widetilde{\boldsymbol{\rho}}_n$ in Algorithm 2 describes the clusters associated to $\boldsymbol{s}_{1:n}$ and provides little information about the clustering of the observations $\boldsymbol{y}_{1:n}$. In the following, we show how to overcome both issues by a suitable choice of distance, namely the Wasserstein distance.

Given two measures μ_1, μ_2 over $\mathbb{Y}$ with finite q-th moment and a cost function $c : \mathbb{Y} \times \mathbb{Y} \to [0, +\infty)$, assumed convex in the following, the Wasserstein distance of order q is defined as

$$\mathcal{W}_q(\mu_1, \mu_2) := \left\{ \inf_{\gamma \in \Gamma(\mu_1, \mu_2)} \int_{\mathbb{Y} \times \mathbb{Y}} c(x_1, x_2)^q \mathrm{d}\gamma(x_1, x_2) \right\}^{\frac{1}{q}}, \tag{6}$$

where $\Gamma(\mu, \nu)$ denotes the Radon space of all measures defined on $\mathbb{Y} \times \mathbb{Y}$ with marginals μ_1 and μ_2. Letting $\mu_1 = n^{-1} \sum \delta_{y_i}$ and $\mu_2 = n^{-1} \sum \delta_{s_i}$, we can use $\mathcal{W}_q$, with a suitable choice of the cost function, to compare $\boldsymbol{y}_{1:n}$ and $\boldsymbol{s}_{1:n}$. We will write $\mathcal{W}_q(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n})$ to make this explicit. This is the case of the Wasserstein-ABC algorithm in Bernton et al. (2019a), where the authors propose to use the Wasserstein distance principally to avoid the choice of statistics for the ABC-SMC scheme. See also, e.g., Bassetti et al. (2006) and Bernton et al. (2019b) for further uses of the Wasserstein distance in the statistical framework.

In this work, the Wasserstein distance is valuable for avoiding summarization, but it is also the key ingredient that allows inference on the partition of $\boldsymbol{y}_{1:n}$ starting from $\widetilde{\boldsymbol{\rho}}_n$ the partition of $\boldsymbol{s}_{1:n}$. First, we note that since μ_1 and μ_2 are always discrete measures, (6) reduces to

$$\mathcal{W}_q(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n}) := \left\{ \min_{P \in M_{n \times n}} \sum_{i=1}^{n} \sum_{j=1}^{n} c(y_i, s_j)^q P_{i,j} \right\}^{\frac{1}{q}} = \left\{ \min_{P \in M_{n \times n}} < C^{(q)}, P > \right\}^{\frac{1}{q}}, \tag{7}$$

where, referring to an optimal transport notation, $C^{(q)}$ denotes the cost matrix of order q, with i, j-th element $C^{(q)}_{i,j} = c(y_i, s_j)^q$ and P is an $n \times n$ matrix encoding a discrete distribution with n^2 support points $\boldsymbol{y}_{1:n} \times \boldsymbol{s}_{1:n}$ with marginals equal to μ_1 and μ_2 respectively. In particular, $P \in M_{n \times n}$ where $M_{n \times n}$ is the space of $n \times n$ matrices with positive entries summing to one and whose rows and columns all sum to n^{-1}. Observe also how the infimum in (6) has been replaced with a minimum in (7). For our purposes, it is useful to think of P as a *transport* matrix associating to each element of the synthetic dataset one or more elements of the observed dataset. If this mapping is one-to-one, then it becomes straightforward to refer the partition $\widetilde{\boldsymbol{\rho}}_n$ of the synthetic data to a partition of the observed data. This is indeed the case as clarified by the following proposition.

Proposition 1. *Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{i=1}^{m} b_i \delta_{y_i}$. If $m = n$ and $\boldsymbol{a} = \boldsymbol{b}$, $a_i = b_i = 1/n$ for all $i \in \{1, \dots, n\}$, then there exists an optimal solution to problem (7) $P^* = P_{\lambda^*}$, which is a permutation matrix associated to an optimal permutation λ^* in the class of permutations of n elements.*

We refer to Proposition 2.1 of Peyré et al. (2019) for a detailed proof of Proposition 1. Hence, by computing the Wasserstein distance between $\boldsymbol{y}_{1:n}$ and $\boldsymbol{s}_{1:n}$, we are also matching the partition $\widetilde{\boldsymbol{\rho}}_n$ of $\boldsymbol{s}_{1:n}$ to a corresponding partition $\boldsymbol{\rho}_n$ of $\boldsymbol{y}_{1:n}$, by considering $\boldsymbol{\rho}_n = \lambda^*(\widetilde{\boldsymbol{\rho}}_n)$. This result is remarkable as it allows us to find a solution to the assignment problem in a polynomial time using, for instance, the well-known simplex

algorithm (Peyré et al., 2019), while the space of all permutations of n objects has size $n!$. Hence, when the distance $\mathcal{W}_q(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n})$ is less than the threshold ε, we accept $\lambda^*(\widetilde{\boldsymbol{\rho}}_n)$ as a realization from $\pi_\varepsilon(\boldsymbol{\rho}_n \,|\, \boldsymbol{y}_{1:n})$. We further remark that such rearrangement is legitimate in force of the exchangeability of the observed data.

3.1 Computation of the Wasserstein distance

When the data are univariate, computing the Wasserstein distance between $\boldsymbol{y}_{1:n}$ and $\boldsymbol{s}_{1:n}$ and the related optimal permutation can be efficiently done. The minimization problem is available in close form, as described in the following remark (remark 2.30 in Peyré et al., 2019).

Remark 1. *For measures μ, ν on $\mathbb{R}$, denote with F_μ (F_ν) the cumulative distribution function of μ (ν) from $\mathbb{R}$ to $[0,1]$, defined as $F_\mu(x) = \int_{-\infty}^{x} d\mu$ for all x and its pseudoinverse $F_\mu^{-1}(x) = \min_z\{z \in \mathbb{R} \cup \{-\infty\} : F_\mu(z) \geq x\}$. Then for any $q \geq 1$ one has $\mathcal{W}_q(\mu,\nu)^q = \int_0^1 |F_\mu^{-1}(x) - F_\nu^{-1}(x)|^q dx$.*

Letting μ and ν be equal to the empirical measures of $\boldsymbol{y}_{1:n}$ and $\boldsymbol{s}_{1:n}$, it is apparent that the optimal solution is given by sorting both the vectors $\boldsymbol{y}_{1:n}$ and $\boldsymbol{s}_{1:n}$. The computational cost of solving the problem in an optimal way is of order $n \log n$. In the multivariate setting, (7) can be solved exactly using the Hungarian algorithm, which has a cost of order n^3. This cost can become prohibitive for large sample sizes, but we can resort to approximating the Wasserstein distance, which significantly saves computational time.

Let $\tau \geq 0$ be a real-valued regularization term. By introducing an entropic regularization factor in (6), we obtain the so-called Sinkhorn distance

$$\left\{W_q^\tau(\mu,\nu)\right\}^q = \min_{\gamma \in \Gamma(\mu,\nu)} \int c(x_1,x_2)^q d\gamma(x_1,x_2) - \tau KL(\gamma||\mu \otimes \nu),$$

where KL is the Kullback Leibler divergence and $\mu \otimes \nu$ denotes the product measure.

In the case of discrete measures, Cuturi (2013) showed that the solution to the Sinkhorn distance could be computed by an iterative algorithm, which requires a cost of n^2 per iteration and converges in $O(\tau^{-2})$ iterations, up to a logarithmic factor. Moreover, the Sinkhorn distance converges to the regular Wasserstein distance as $\tau \to 0$. More recently, Altschuler et al. (2017) proposed a greedy variant of the original Sinkhorn algorithm, which runs in a nearly linear time. Nonetheless, both these algorithms require the computation of the full pairwise cost matrix C, which is still $O(n^2)$.

3.2 ABC-MCMC approach for random partitions

The main problem we face when sampling from π_ε is that the size of the partitions' space $\mathcal{P}_{1:n}$ escalates quickly as n increases (its growth is super-exponential), which makes the acceptance-rejection ABC algorithm useless in practical applications. Below, we outline an ABC-MCMC sampling scheme that overcomes these difficulties. In ABC-MCMC,

the value for the parameters at the r-th iteration is sampled from a transition kernel which depends on the value of the parameters at the $(r-1)$-th iteration. Specifically, we propose to use the predictive distribution of an additional sample of size n as transition kernel. At each iteration, we sample a candidate partition and the associated synthetic dataset until the distance between the true and synthetic data is less than a threshold. Once that such condition is satisfied, we perform a Metropolis-Hastings step to accept the proposed value or remain on the current state of the latent partition. We further show that with our choice of proposal distribution, the acceptance rate of the Metropolis-Hastings step is always equal to one.

First, observe that model (2) together with a prior for $\pi(\boldsymbol{\rho}_n, \boldsymbol{\theta}^*)$ that factorizes into the EPPF of $\boldsymbol{\rho}_n$ times $\prod_{j=1}^{k} G_0(\mathrm{d}\theta_j^*)$ is equivalent to assuming $y_i \,|\, \theta_i \sim \mathcal{K}(\cdot \,|\, \theta_i)$ and

$$P(\boldsymbol{\theta} \in \mathrm{d}\boldsymbol{\theta}) = p_k^{(n)}(n_1, \dots, n_k) \prod_{j=1}^{k} G_0(\mathrm{d}\theta_j^*),$$

where the θ_j^*'s are the unique values in $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, each appearing with frequencies n_j. Then, the predictive distribution for the $(n+1)$-th latent parameter θ_{n+1}, conditionally on $\theta_1, \dots, \theta_n$, is given by

$$\begin{aligned} \mathrm{P}(\theta_{n+1} \in \mathrm{d}t \,|\, \theta_1, \dots, \theta_n) = \frac{\mathrm{p}_{\mathrm{k}+1}^{(\mathrm{n}+1)}(n_1, \dots, n_k, 1)}{\mathrm{p}_{\mathrm{k}}^{(\mathrm{n})}(n_1, \dots, n_k)} G_0(\mathrm{d}t) \\ + \sum_{j=1}^{k} \frac{\mathrm{p}_{\mathrm{k}}^{(\mathrm{n}+1)}(n_1, \dots, n_j + 1, \dots n_k)}{\mathrm{p}_{\mathrm{k}}^{(\mathrm{n})}(n_1, \dots, n_k)} \delta_{\theta_j^*}(\mathrm{d}t). \end{aligned} \tag{8}$$

Observe that the predictive distribution is a convex combination of the prior guess, expressed in terms of G_0, and the empirical information of the previous values of the latent parameters, driven by the EPPFs' ratios.

We can exploit the chain rule to produce a sample n step further from the current state, obtaining

$$\begin{aligned} \mathrm{P}(\boldsymbol{\theta}_{n+1:2n} \,|\, \boldsymbol{\theta}_{1:n}) = \mathrm{P}(\theta_{n+1} \,|\, \boldsymbol{\theta}_{1:n}) \mathrm{P}(\theta_{n+2} \,|\, \boldsymbol{\theta}_{1:n}, \theta_{n+1}) \\ \dots \mathrm{P}(\theta_{2n} \,|\, \boldsymbol{\theta}_{1:n}, \theta_{n+1}, \dots, \theta_{2n-1}). \end{aligned} \tag{9}$$

Since at each step of the chain rule we are using the predictive distribution in (8), the resulting $\boldsymbol{\theta}'_{1:n} = \boldsymbol{\theta}_{n+1:2n}$, is a combination of the prior guess and the empirical information of $\boldsymbol{\theta}_{1:n}$. Thanks to the fact that $\boldsymbol{\theta}_{1:2n}$ is an exchangeable sequence, we can think on $\boldsymbol{\theta}'_{1:n}$ as a standalone sample form $\tilde{p}$, with latent partition $\boldsymbol{\rho}'_n$ here termed *raw candidate*. We can then sample a set of synthetic data $\boldsymbol{s}_{1:n}$ conditionally on $\boldsymbol{\theta}'_{1:n}$, with the generic $S_i \sim \mathcal{K}(s_i; \theta'_i)$ for all $i \in \{1, \dots, n\}$.

Once we have produced a set of synthetic data, we evaluate its distance from the observed data $\boldsymbol{y}_{1:n}$ via the Wasserstein metric by solving $\mathcal{W}_q(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n})$. From Proposition 1, as a byproduct of the computation of the Wasserstein metric, we also compute

an *optimal* permutation λ^* of $[n]$ in the sense of Proposition 1. This is used to match the synthetic data to the observed one and, in particular, the partition of the synthetic data $\boldsymbol{\rho}'_n$ with the corresponding partition of the observed data $\boldsymbol{\rho}''_n := \lambda^*(\boldsymbol{\rho}'_n)$, providing also a permuted version of the latent parameters $\boldsymbol{\theta}''_{1:n} := \lambda^*(\boldsymbol{\theta}'_{1:n})$. The fact that we can propose $\boldsymbol{\rho}'_n$ and then permute it into $\boldsymbol{\rho}''_n$ without impacting the limiting distribution of the chain is due to the exchangeability assumption underlying the mixture model, that is, the invariance with respect to permutation of the joint distribution of the observations y_i and the latent parameters θ_i. Whenever $\mathcal{W}_q(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n})$ is smaller than a threshold ε, we can perform a Metropolis-Hastings step to update the state of the latent partition or stay on the current value. We notice that the acceptance rate of this latter Metropolis-Hastings step is always 1, since we are proposing a partition according to (8) and (9). Let $q(\boldsymbol{\theta}_{1:n} \to \boldsymbol{\theta}''_{1:n}) = \mathrm{P}(\boldsymbol{\theta}''_{1:n} \,|\, \boldsymbol{\theta}_{1:n})$ be the proposal distribution for the latent parameters, with $\boldsymbol{\theta}''_{1:n}$ the optimal permuted version of $\boldsymbol{\theta}'_{1:n}$, and $q(\boldsymbol{\rho}_n \to \boldsymbol{\rho}''_n)$ the proposal distribution induced on the latent partition. Indeed, the acceptance rate of the Metropolis-Hasting step is equal to

$$\alpha(\boldsymbol{\rho}''_n, \boldsymbol{\rho}_n) = 1 \wedge \frac{\mathrm{P}(\boldsymbol{\theta}_{1:n}) q(\boldsymbol{\theta}_{1:n} \to \boldsymbol{\theta}''_{1:n})}{\mathrm{P}(\boldsymbol{\theta}''_{1:n}) q(\boldsymbol{\theta}''_{1:n} \to \boldsymbol{\theta}_{1:n})} = 1 \wedge \frac{\mathrm{P}(\boldsymbol{\theta}_{1:n}) \mathrm{P}(\boldsymbol{\theta}''_{1:n}, \boldsymbol{\theta}_{1:n}) \mathrm{P}(\boldsymbol{\theta}''_{1:n})}{\mathrm{P}(\boldsymbol{\theta}''_{1:n}) \mathrm{P}(\boldsymbol{\theta}_{1:n}, \boldsymbol{\theta}''_{1:n}) \mathrm{P}(\boldsymbol{\theta}_{1:n})} = 1,$$

where $\mathrm{P}(\boldsymbol{\theta}''_{1:n}, \boldsymbol{\theta}_{1:n}) = \mathrm{P}(\boldsymbol{\theta}_{1:n}, \boldsymbol{\theta}''_{1:n})$ in force of the exchangeability of the latent parameters. Such behaviour is caused by the usage of the predictive distribution as proposal distribution.

Algorithm 3: ABC-MCMC for latent random partitions.

```
[1] input a set of data y_{1:n}, a threshold ε, and possibly hyperparameters for
    K(·; θ);
[2] set admissible initial values for θ^{(0)}_{1:n};
[3] for r = 1, ..., R do
[4]     repeat
[5]         propose a move from θ^{(r-1)}_{1:n} to θ'_{1:n} according to a transition kernel
              q(θ^{(r-1)}_{1:n} → θ'_{1:n}), with related partition ρ'_n;
[6]         sample s_{1:n} | θ'_{1:n} vector of synthetic data, where S_i ~ K(·, θ'_i);
[7]     until W_q(y_{1:n}, s_{1:n}) < ε; denote with λ* the optimal permutation, cf.
          Proposition 1;
[8]     accept ρ''_n := λ*(ρ'_n) as realization from π_ε(ρ_n | y_{1:n});
[9] end
```

An implementation of the previous strategy is reported in Algorithm 3. We can easily prove that Algorithm 3 is effectively producing R realizations from a Markov chain which has invariant distribution corresponding to the ε-approximation of the posterior distribution. We synthesize in the following Lemma the convergence of Algorithm 3.

Lemma 1. *Assume $\{\boldsymbol{\rho}_{n,1}, \boldsymbol{\rho}_{n,2}, \ldots\}$ be a sample from an ABC-MCMC scheme according to algorithm 3, with proposal $q(\boldsymbol{\rho}_n \to \boldsymbol{\rho}'_n)$ described in (8) and (9). Then the produced chain has invariant distribution $\pi_\varepsilon(\boldsymbol{\rho}_n \,|\, \boldsymbol{y}_{1:n})$.*

The proof of Lemma 1 is trivial, and follows from Marjoram et al. (2003). To help the understanding of the proof of Theorem 1, we report in Section S1 of the Supplementary Material a proof of Lemma 1. The presented strategy could be a first simple approach to perform approximate inference of latent random partitions. Nevertheless we can relax the assumption of a fixed threshold ε along the chain.

3.3 An adaptive strategy for ε

The threshold ε strongly impacts the computational time and the quality of the results of Algorithm 3. See, e.g., the simulation studies in Section S4 of the Supplementary Material where small ε leads to a poor mixing (it is hard to accept a proposed value) and large ε provides a rough approximation of the true posterior. Choosing a suitable threshold seems an essential task, but as remarked by Vihola and Franks (2020), threshold selection (see e.g. Beaumont et al., 2002; Wegmann et al., 2009) may not be suitable in an MCMC regime with weakly informative prior.. Instead of a fixed ε, a possible strategy is to consider a sequence $\{\varepsilon_l\}_{l\geq 1}$, which allows for larger thresholds in the early phase of the chain, leading to a larger acceptance rate in the early phase of the algorithm.

Algorithm 4: adaptive ABC-MCMC for latent random partitions.

input a set of data $\boldsymbol{y}_{1:n}$, a threshold ε_0, and possibly hyperparameters for $\mathcal{K}(\cdot;\theta)$;
set admissible initial values for $\boldsymbol{\theta}^{(0)}_{1:n}$, set $l=1$;
for $r=1,\dots,R$ **do**
  repeat
    propose a move from $\boldsymbol{\theta}^{(r-1)}_{1:n}$ to $\boldsymbol{\theta}'_{1:n}$ according to a transition kernel $q(\boldsymbol{\theta}^{(r-1)}_{1:n} \to \boldsymbol{\theta}'_{1:n})$, with related partition $\boldsymbol{\rho}'_n$;
    sample $\boldsymbol{s}_{1:n} \,|\, \boldsymbol{\theta}'_{1:n}$ vector of synthetic data, where $S_i \sim \mathcal{K}(\cdot, \theta'_i)$;
    update ε_l and set $l = l+1$;
  until $\mathcal{W}_p(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n}) \leq \varepsilon_\ell$; denote with λ^* the optimal permutation, cf. Proposition 1;
  accept $\boldsymbol{\rho}''_n := \lambda^*(\boldsymbol{\rho}'_n)$, as realization from $\pi_{\varepsilon_l}(\boldsymbol{\rho}_n \,|\, \boldsymbol{y}_{1:n})$;
end

Algorithm 4 describes an implementation of the adaptive strategy. We remark that while we are sampling R values from the approximate posterior distribution, the threshold update can also be done when we reject the proposed values.

By assuming that the sequence $\{\varepsilon_l\}_{l\geq 1}$ converges to a fixed threshold ε^*, we are able to characterize the limit behaviour of the target distribution, showing that the MCMC has invariant distribution corresponding to the ε^*-approximation of the posterior distribution $\pi_{\varepsilon^*}(\boldsymbol{\rho}_n \,|\, \boldsymbol{y}_{1:n})$, as stated in the following Theorem.

Theorem 1. *Let* $\{\varepsilon_l\}_{l\geq 1}$ *be an* $\mathbb{R}^+$*-valued sequence of elements, such that* $\lim_{l\to+\infty}|\varepsilon_l - \varepsilon^*| = 0$*. Let* $\left\{\boldsymbol{\rho}_n^{(1)}, \boldsymbol{\rho}_n^{(2)}, \dots\right\}$ *be a sample from an ABC-MCMC scheme according to Algorithm 4, with proposal* $q(\boldsymbol{\rho}_n \to \boldsymbol{\rho}'_n)$ *according to (8) and (9). Let* $p(w)$ *denotes the density function of* $\mathcal{W}_q(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n})$*, where* $\boldsymbol{s}_{1:n}$ *denotes the* l*-th synthetic sample, and assume* $0 < p(w) < M$ *for all* l*. Then, for* $l \to +\infty$*, we have that* $\pi_{\varepsilon^*}(\boldsymbol{\rho}_n \,|\, \boldsymbol{y}_{1:n})$ *is the invariant distribution of the chain.*

To prove Theorem 1 we exploit the continuity of $\mathcal{W}_p(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n})$, and the convergence of $\{\varepsilon_t\}_{t\in T}$ to its limit. A detailed proof is reported in Section S2 of the Supplementary Material.

The sequence of thresholds $\{\varepsilon_l\}_{l\geq 1}$ can be specified in many ways. One can, for example, define a decreasing sequence from a large initial value ε_0 to a smaller target value ε^*. Such a strategy provides more flexibility than a fixed threshold, as in Algorithm 3, but it requires particular care. For example, if the sequence is quickly approaching the optimal value and the algorithm visits a local mode, it can be stuck in the neighbourhood of such value. To overcome this issue it is reasonable to assume a sequence which is actually adapting over the sampling times, i.e. it becomes smaller or larger depending on if we are accepting too many or too few proposed values, but also the adaptation is vanishing as far as the number of values sampled from the posterior is increasing. We define the sequence of thresholds on a log-scale, according to Vihola and Franks (2020), with

$$\log(\varepsilon_l) = \log(\varepsilon_0) + \sum_{j=1}^{l} \frac{(\alpha^* - \mathbb{1}_{[\mathcal{W}_p(\boldsymbol{y}_{1:n}, \boldsymbol{s}_{1:n}) < \varepsilon_{j-1}]})}{j^{2/3}}, \tag{10}$$

where α^* denotes the target acceptance rate. A strategy as in (10) produces a sequence with diminishing adaptation of order $o(l^{-3/2})$. Further, in Section S3 of the Supplementary Material we show that the adaptation scheme in Equation 10 satisfies the hypotheses of Theorem 1. Vihola and Franks (2020) suggest stopping the adaptation after the burn-in phase, allowing control of the approximation level by imposing a fixed threshold from a certain point of the chain (strategy ABCad1 in Section 4). However, in some scenarios, we found that updating the threshold along the entire sampling led to slightly better numerical results (strategy ABCad2 in Section 4). Within this strategy, we waive controlling the degree of approximation of the target distribution to increase the flexibility of the sampler since the threshold might increase or decrease over the entire sampled chain. Nevertheless, we remark that it is not possible to provide a proper interpretation of the degree of approximation included in the sampling strategy even if the threshold is fixed.

Alternatively, one can refer to a post-processing algorithm as in Vihola and Franks (2020). Let the threshold ε_l vary across the whole chain and then select a new $\delta < \varepsilon_l$ threshold and weight the posterior samples taking into account the difference between δ and ε. See their Theorem 1. However, in our simulations, we found that this was superfluous. Although letting ε_l vary across the whole chain results in a loss of control on the degree of approximation introduced by the ABC likelihood, the posterior inference we obtain is consistent with the true data generating process and similar to the one obtained using competitor algorithms.

4 Numerical illustrations

We present some numerical illustrations of the ABC-MCMC sampling scheme for latent partitions. Section 4.1 shows a comparison with a marginal sampler for the non-conjugate case (Neal, 2000). Section 4.2 illustrates the effect of the Sinkhorn approximation in a multivariate setting. In Section 4.3 we present a synthetic example where the data generating process and the kernel match the Lévy-driven stochastic volatility model (Barndorff-Nielsen and Shephard, 2002; Chopin et al., 2013), and we are able to infer the latent partition of a sample of time series. Section 4.4 discusses an example where the data lies on a more abstract space. Further examples are deferred to the Supplementary Material. In particular, in Section S4 we consider a tractable case where the data are generated from a mixture of Gaussian distributions and the kernel function matches a Gaussian distribution. From this scenario we can appreciate that the inclusion of an adaptation step slightly increases the computational time, but produces more precise estimates of the latent partition.

For all the illustrations, once we have run the algorithms we estimate the optimal latent partition of the data by resorting to a decision theoretic approach based on the *variation of information* loss function (Wade and Ghahramani, 2018; Rastelli and Friel, 2018). When comparing the estimated latent partition with the true partition, we resort to the normalized variation of information, i.e.

$$\mathrm{VI}(\boldsymbol{r}_1, \boldsymbol{r}_2) = \frac{1}{\log n}(\mathrm{H}(\boldsymbol{r}_1) + \mathrm{H}(\boldsymbol{r}_2) - 2\mathrm{I}(\boldsymbol{r}_1, \boldsymbol{r}_2)),$$

where $\boldsymbol{r}_\ell = \{A_{1,\ell}, \dots, A_{k_\ell,\ell}\}$, $\ell = 1, 2$, $\mathrm{H}(\boldsymbol{r}) = \sum_{j>1}^{k} p_j(\boldsymbol{r}) \log p_j(\boldsymbol{r})$ represents the entropy associated to the partition $\boldsymbol{r}$, while

$$\mathrm{I}(\boldsymbol{r}_1, \boldsymbol{r}_2) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} p_{ij}(\boldsymbol{r}_1, \boldsymbol{r}_2) \log[p_{ij}(\boldsymbol{r}_1, \boldsymbol{r}_2)/(p_i(\boldsymbol{r}_1)p_j(\boldsymbol{r}_2))]$$

denotes the mutual information of $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, with $p_j(\boldsymbol{r}_\ell) = |A_{j,\ell}|/n$, $p_{ij}(\boldsymbol{r}_1, \boldsymbol{r}_2) = |A_{i,1} \cap A_{j,2}|/n$, and k, k_1, k_2 denote the cardinality of $\boldsymbol{r}$, $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ respectively. Lower values of the variation of information indicate that $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are close. To measure the mixing of the MCMC, we report the effective sample size (see, e.g., Equation (11.6) in Chapter 11 of Gelman et al., 2013) of the chains of a functional of the visited partitions $\{\boldsymbol{r}^{(j)}\}_{j\geq 1}$, namely the entropy $\mathrm{H}(\boldsymbol{r}^{(j)})$. When considering the adaptive ABC-MCMC, we set the target acceptance rate α^* equal to 0.1, (i.e., the optimal rate as defined in Vihola and Franks, 2020), and we update the thresholds according to (10). The chains are sampled for 15 000 iterations, discarding the first 5 000 as burn-in.

4.1 Comparing ABC and standard MCMC approaches

We consider a scenario where the density is known up to an intractable constant. We simulate sets of data from an unbalanced mixture of two g-and-k distributions, i.e.

$$f_0(y) = 0.75\psi(y; -3, 0.75, -0.9, 0.1, 0.8) + 0.25\psi(y; 3, 0.5, 0.4, 0.5, 0.8),$$

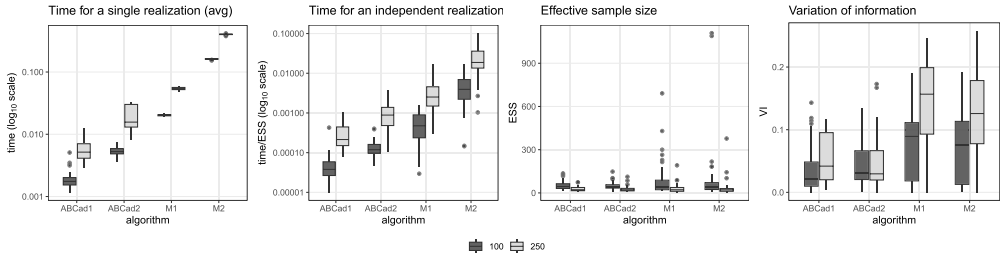

Figure 1: Simulation summaries for the mixture of g-and-k distributed data. Different sample sizes $n \in \{100, 250\}$ (dark gray and light gray respectively). The results are averaged over 100 replications. Different sampling strategies: adaptive ABC-MCMC algorithm (ABCad1) with adaptation stopped after the burn-in phase; adaptive ABC-MCMC algorithm (ABCad2); marginal sampler (M1 - M2).

where $\psi(x; a, b, g, k, c)$ denotes the density function of a g-and-k distribution with location parameter a, scale parameter b, shape parameter g (mainly affecting the skewness), shape parameter k (mainly affecting the kurtosis), and the parameter c fixed and equal to 0.8. The g-and-k distribution is defined through its quantile function $F^{-1}_{GK}(u) : [0, 1] \to \mathbb{R}$ with

$$F^{-1}_{GK}(u) = a + b\,(1 + c\tanh(gu/2))\,\Phi^{-1}(u)\left(1 + \Phi^{-1}(u)^2\right)^k \tag{11}$$

and $\Phi^{-1}(u)$ denotes the quantile function of a standard Gaussian distribution. We consider different sample sizes, with $n \in \{100, 250\}$. Our prior model specification consists of a g-and-k mixture with MFM mixing measure, with $\alpha = 1$ and $\lambda = 1$. Moreover, the base measure G_0 equals a product of independent distribution for the relevant parameters of the model, with $a \sim \mathrm{N}(0, 25)$, $b \sim \textsc{Inv-Gamma}(1, 2)$, $g \sim \mathrm{N}(0, 25)$, and $k \sim \textsc{Inv-Gamma}(1, 2)$.

We consider two cases of adaptive ABC-MCMC: a case where the adaptation is stopped after the burn-in phase (ABCad1) and a case where the adaptation is carried over the entire sampled chains (ABCad2). The performances of the ABC-MCMC sampler are further compared with a marginal sampling scheme with a Monte-Carlo integration to estimate the probability of sampling a new value, in the spirit of Algorithm 8 of Neal (2000), where we consider $m \in \{10, 100\}$ temporary values for the Monte-Carlo integration (algorithms M1 and M2 respectively). We remark that while sampling a realization from a g-and-k distribution can be done efficiently, the evaluation of the density requires numerical optimizations, which also impacts the Monte Carlo integration step.

Figure 1 shows the computational time required to obtain a single realization, the computational time required to perform a single independent realization, the effective sample size of the entropy of the partitions and the distance of the latent partition estimate and the true latent partition, for different sample sizes n and different sampling strategies. The ABC-MCMC algorithms are significantly faster than the marginal

strategies. Further, the increased computational cost of the marginal strategies does not translate into more precise estimates of the partition, as shown in Figure 1. The marginal sampler, known for its performances in terms of mixing of the sampled chains, is showing performances comparable to the ABC-MCMC adaptive strategy. Further insights on this example, but with larger numbers of components in the data generating process, are deferred to Section S5 of the Supplementary Material. In particular, as the number of components increases, the adaptive ABC approach maintain its relative efficiency with respect to the competitor, but the accuracy of the estimates is getting similar for all the different methods.

4.2 The effect of Sinkhorn approximation

Here we want to illustrate the effect of the Sinkhorn approximation on the clustering. As a multivariate extension of the example in Section 4.1, we consider data from the multivariate g-and-k distribution, in dimension $p = 2$. The multivariate g-and-k distribution shares the same intractability as the univariate one, with the further addition that, to the best of our knowledge, it is not possible to approximate the probability density function numerically. To generate from the bivariate g-and-k distribution it suffices to simulate (u_1, u_2) from a bivariate Gaussian distribution with zero mean, unit marginal variances and correlation ρ and then let

$$y_i = a_i + b_i \left(1 + c_i \tanh(g_i u_i/2)\right) u_i \left(1 + u_i^2\right)^{k_i}, \quad i = 1, 2.$$

As in the previous example, we assume c_i fixed and equal to 0.8 and the correlation between the u_i's fixed as $\rho = 0.5$. A priori, we assume a multivariate g-and-k mixture model with Pitman-Yor process mixing measure, with $\vartheta = 1$ and $\sigma = 0.1$. The base measure G_0 over parameters $\{a_i, b_i, g_i, k_i\}$ $i = 1, 2$ factorizes into the product of the marginal distributions, that are assumed identical to the ones in Section 4.1 for each $i = 1, 2$. We simulated data from an equally-weighted mixture of bivariate g-and-k distributions, with parameters (along each direction $i = 1, 2$) equal to the ones in Section 4.1. We exploit an adaptive ABC-MCMC scheme with the adaptation carried over the entire sampled chains.

Figure 2 shows an example of simulated data (left) with the posterior estimate of the similarity matrix (middle), while the right column shows the effective sample size of the entropy $\mathrm{H}(\boldsymbol{r})$ of the visited partitions (right-top) and the VI distance (right-bottom) evaluated for 100 of replications, for different sample sizes. Both methods show comparable effective sample sizes, while the Sinkhorn algorithm produces slightly more precise estimates of the latent partitions. Regarding the computational cost, we did not observe any significant difference comparing the runtimes when using the Wasserstein or Sinkhorn distance within this particular scenario.

4.3 Time series stratification

We consider another scenario where observations are multivariate and the kernel is intractable. Specifically, let $y_i = (y_{i,1}, \ldots, y_{i,T})$, $i = 1, \ldots, n$ and think of each observation as a time series. The kernel $\mathcal{K}(\cdot; \theta)$ equals the Lévy-driven stochastic volatility

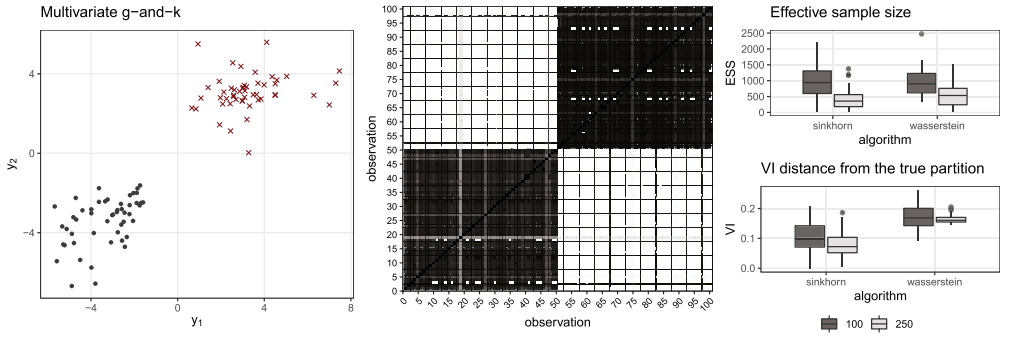

Figure 2: Left column: an example of data generated from the mixture of multivariate g-and-k distributions. Middle column: an example of estimated posterior similarity matrix. Right column: effective sample size of the entropy (top) and VI distance from the true partition (bottom) of different replications for the multivariate g-and-k scenario.

model (Barndorff-Nielsen and Shephard, 2002; Chopin et al., 2013) with parameters $\theta = (\mu, \beta, \xi, \omega, \eta)$, i.e., we assume

$$y_{i,t+1} \,|\, \mu_i, \beta_i, v_{i,t+1} \sim \mathrm{N}(\mu_i + \beta_i v_{i,t+1}, v_{i,t+1}), \quad t = 1, \dots, T$$

$$v_{i,t+1} = \eta_i^{-1}\Big(z_{i,t} - z_{i,t-1} + \sum_{j=1}^{k} e_{i,j}\Big), \qquad z_{i,t+1} = e^{-\eta_i} z_{i,t} + \sum_{j=1}^{k} e^{-\eta_i(t+1)-c_{i,j}} e_{i,j}$$

$$c_{i,1}, \dots, c_{i,k_i} \,|\, k_i \overset{\text{iid}}{\sim} \textsc{Unif}(t, t+1), \;\; e_{i,1}, \dots, e_{i,k_i} \,|\, k_i \overset{\text{iid}}{\sim} \textsc{Exp}(\xi_i/\omega_i^2), \;\; k_i \sim \textsc{Poi}(\eta_i \xi_i^2/\omega_i^2),$$

where we suppressed the dependence of k, the $c_{i,j}$'s and $e_{i,j}$'s on the time t. Indeed, these are generated independently at each time.

The Lévy-driven stochastic volatility model is popular in financial applications, where it is used to model the log-return of stocks, i.e., $y_t = \log(x_{t+1} - x_t)/x_t$ where x_t is the price of the stock at time t. A similar scenario, but with a single time series, was analyzed with ABC tools in Bernton et al. (2019a) as a challenging example in the context of state-space models. Observe how their goal is different from ours: they set to perform inference on the parameters of the stochastic volatility model that generated a single time series, while our aim is to cluster similar elements belonging to a sample of multiple time series.

As central part of our methodology, we need to select a distance function between two time series $\{y_t\}_{t=1}^T$ and $\{s_t\}_{t=1}^T$. In our example, we follow Bernton et al. (2019a) in their choice of distance as detailed below, but note here that our methodology is valid for any choice of cost. See Dyer et al. (2021) for other possibilities. To define our distance, we consider the 1-lagged time series, i.e. the pointclouds $\{Y_i\}_{i=1}^{T-1} \subset \mathbb{R}^2$, $Y_i = (y_i, y_{i+1})$, and $\{S_i\}_{i=1}^{T-1}$ defined analogously from $\{s_t\}_{t=1}^T$. Then, we consider the Hilbert space-filling curve and obtain one-dimensional projections of the Y_i's and S_i's on the curve and compute an optimal matching, say σ^*, between the one-dimensional projections. Finally we set $d_H^2(\{y_t\}_{t=1}^T, \{s_t\}_{t=1}^T) = \sum_{i=1}^{T-1} \|Y_i - S_{\sigma_i^*}\|^2$. From a practical

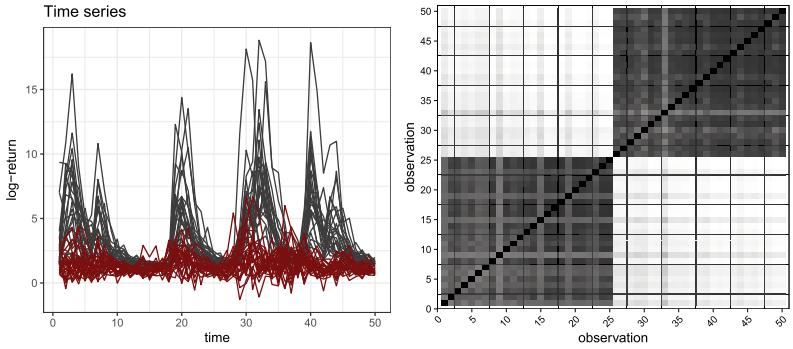

Figure 3: Left column: data generated from the mixture of Lévy-driven stochastic volatility models. Right column: an example of estimated posterior similarity matrix.

viewpoint, we use the `hilbertsort` function in the CGAL C++ library that directly finds the optimal sorting σ^* without needing to compute the Hilbert curve.

We generated $n = 50$ time series with $T = 50$ observed times, from a two-component mixture with equal weights. In the first component, data are generated from the Lévy-driven stochastic volatility model with parameters $(1.5, 2.75, 1.0, 2.5, 1.0)$ while in the second with parameters $(1.0, 2.0, 0.6, 1, 0.4)$, see Figure 3 (left panel). The base measure G_0 is the product of independent distributions, namely $\mu \sim \text{N}(1, 4^2)$, $\beta \sim \text{N}(1, 4^2)$, $\xi \sim \text{Gamma}(1, 2)$, $\omega \sim \text{Gamma}(1, 1)$ and $\eta \sim \text{Gamma}(1, 1)$. We further set a Pitman-Yor process as mixing measure, with $\vartheta = 1$ and $\sigma = 0.1$.

We ran the adaptive ABC-MCMC (ABCad2) algorithm. The right column of Figure 3 shows the posterior similarity matrix (right) and the point estimate of the random partition obtained using the greedy algorithm in Rastelli and Friel (2018) highlighted with different colours (left), the adjusted rand index between the estimated and true partition is equal to one. To give a rough estimate of the computational cost, the runtime required by this simulation is approximately three hours on a Macbook Pro M1 with 16GB of RAM.

4.4 Clustering a population of networks

As a final illustration, we analyze data from $n = 52$ airline companies serving the US airports.[1] For each airline, we represent the covered routes as the edges of a graph (also called network) $\boldsymbol{G}_i = \{\mathcal{V}_i, \mathcal{E}_i\}$, where $\mathcal{V}_i$ represents an M-dimensional set of nodes (or vertexes) for the i-th observation, $i = 1, \dots, n$, and $\mathcal{E}_i$ denotes the set of tuples $(j, k) \in \mathcal{V}_i \times \mathcal{V}_i$. The airports are shared by all the companies, so that $\mathcal{V}_i = \mathcal{V}$ for all $i = 1, \dots, n$, and, in particular, the $M = 100$ nodes correspond to the 100 most served airports in the US. We further assume the graph to be undirected. With the aim of clustering together graphs with similar topology, we consider unlabelled networks, so that, for instance, the two networks in Figure 4 are completely identical.

[1] Data are available on the OpenFlights database (`https://openflights.org/`).

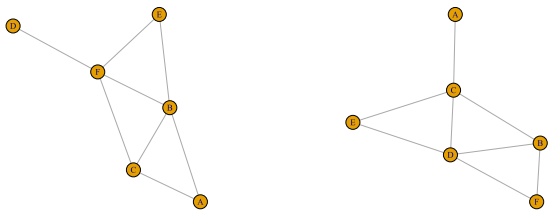

Figure 4: An example of two graphs differing only on the labeling of the nodes, but with the same topology. The graph in the right picture is recovered starting from the graph in the left picture by renaming the nodes as $D \to A$, $C \to B$, $F \to C$, $B \to D$, $A \to F$, $E \to E$.

To measure the distance between two specific graphs $\boldsymbol{G}_i$ and $\boldsymbol{G}_j$ we use as cost operator $C(\boldsymbol{G}_i, \boldsymbol{G}_j)$, the spectral distance between graphs, as defined in Gu et al. (2015), that is

$$C(\boldsymbol{G}_i, \boldsymbol{G}_j)^2 = \frac{\big||\mathcal{E}_i| - |\mathcal{E}_j|\big| + 1}{\min\{|\mathcal{E}_i|, |\mathcal{E}_j|\} + 1} \|\boldsymbol{\lambda}_{G_i} - \boldsymbol{\lambda}_{G_j}\|_2^2,$$

where $|\mathcal{E}_i|$ is the number of vertices of the i-th graph, and $\boldsymbol{\lambda}_G$ is the vector of eigenvalues of the Laplacian of the graph G, defined as $I - D^{-1/2} G D^{-1/2}$ where D is a diagonal matrix with entries $D_{ii} = \sum_j G_{i,j}$. Among different possible choices for a metric to compare graphs, the spectral distance is particularly suited for our purpose, as it focuses on the topology of the networks rather than on the labeling of the nodes.

Remark 2. *The main difference between this application and the previous sections is that the observed data $\boldsymbol{G}_{1:n}$ are not a subset of $\mathbb{R}^d$ anymore. Nonetheless, the formulation of the Wasserstein distance in Equation (7), and then the consequent results, remains valid for general choices of the cost operator C.*

We consider as data generating process an Exponential Random Graph Model (ERGM, see, e.g., Robins et al., 2007). Recall that we denote by M the number of nodes, assumed fixed, and let $\boldsymbol{Y}$ an $M \times M$ binary matrix such that $Y_{jk} = 0$ if j and k are not connected and $Y_{jk} = 1$ otherwise. In this context, the matrix $\boldsymbol{Y}$ is usually termed adjacency matrix, and it is in one-to-one correspondence with $\boldsymbol{G} = (\mathcal{V}, \mathcal{E})$, when we assume $\boldsymbol{G}$ an unlabeled network. The assumption underlying ERGMs is that the topology of an observed graph $\boldsymbol{y}$ can be explained by a set of statistics $\boldsymbol{s}(\boldsymbol{y})$. In particular we assume

$$P(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}^\intercal \boldsymbol{s}(\boldsymbol{y})\right)}{Z_{\boldsymbol{\theta}}}, \tag{12}$$

where $Z_{\boldsymbol{\theta}}$ is a normalizing constant, not available in closed form. Simulation strategies from (12) are discussed in Morris et al. (2008). The model specification is completed by specifying the statistics $\boldsymbol{s}(\boldsymbol{y})$. Generally the choice of these statistics is problem specific, and there is no one-fits-all choice. For the airlines networks, we have the following structural behaviour: (i) several networks have a strong hub-and-spoke behaviour, meaning that there is one node connected to most of the other ones and, apart from that particular node, the rest of the network is barely connected, i.e. companies with a main

airport connected to most other ones; (ii) several networks instead have a more connected topology, meaning that most of the nodes are connected to many other nodes, i.e. companies which are diffuse over the airports; (iii) in both cases, there are nodes that are not connected to any other node, i.e. connections not served by the company. These insights led us to consider:

$$\boldsymbol{s}(\boldsymbol{y}) = \left(\sum_{i,j=1}^{M} y_{ij}, \sum_{j=1}^{M} \mathbb{1}_{[y_{j\bullet}=0]}, \sum_{j=1}^{M} \mathbb{1}_{[y_{j\bullet}=1]}, \sum_{j=1}^{M} \mathbb{1}_{[y_{j\bullet}\in[2,10]]}, \sum_{j=1}^{M} \mathbb{1}_{[y_{j\bullet}\in[11,50]]} \right),$$

where $y_{j\bullet} = \sum_{k=1}^{M} y_{jk}$. Despite the simplicity of the model, maximum likelihood estimates of the parameters $\boldsymbol{\theta}$ for our sample of networks are hard to compute, and most of the time we incur in numerical errors. We set a Pitman-Yor process as mixing distribution, with $\vartheta = 1$ and $\sigma = 0.1$, and we complete the specification of the model by letting G_0 be a five dimensional Gaussian distribution with covariance equal to $10\mathbf{I}_5$, where $\mathbf{I}_5$ denotes the identity matrix of dimension 5, and mean $(-4, 3, 3, 15, -20)$. The values of the mean parameters were chosen via an empirical Bayes procedure as the maximum likelihood estimates when considering all the data together. We obtain a sample from the posterior distribution of interest using an adaptive ABC-MCMC scheme where the adaptation is carried over the entire sampled chain.

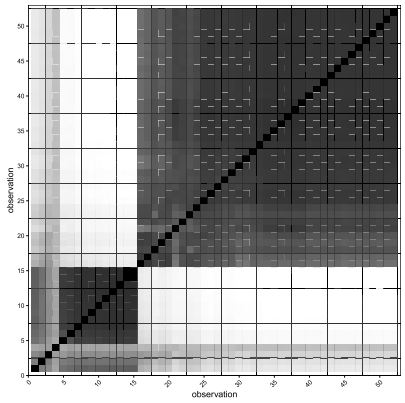

Figure 5: Posterior similarity matrix for the airline dataset.

The point estimate of the latent partition identifies 2 clusters, highlighted in Figure 5, with the cluster sizes n_j reported in Table 1. We can appreciate that the two clusters strongly differ from each other: the first cluster is characterized by observations with a large number of nodes with no connections and a small number of nodes with few connections, i.e. the airline companies belonging to this cluster are serving few airports in the network. On the counterpart, the second cluster is composed by airline companies which are serving several airports.

Statistics	1st cluster	2nd cluster
Cardinality	37	15
$\sum_{i,j=1}^{M} y_{ij}$	(2, 6, 12)	(142, 198, 913)
$\sum_{j=1}^{M} \mathbb{1}_{[y_{j\bullet}=0]}$	(93, 96, 98)	(11, 25, 54)
$\sum_{j=1}^{M} \mathbb{1}_{[y_{j\bullet}=1]}$	(2, 2, 4)	(7, 22, 34.5)
$\sum_{j=1}^{M} \mathbb{1}_{[y_{j\bullet}\in[2,10]]}$	(0, 1, 3)	(10, 31, 54)
$\sum_{j=1}^{M} \mathbb{1}_{[y_{j\bullet}\in[11,50]]}$	(0, 0, 0)	(1, 5, 11.5)

Table 1: Summary statistics in the two clusters. Values in the second and third columns correspond to the first, second, and third quartile of the statistics in the first and second clusters respectively.

5 Discussion

In this paper, we introduced an approximate sampling strategy to deal with model-based clustering whenever the kernel function is known up to an intractable normalizing constant, but it is easy to define a distance between pairs of observations. We proposed an ABC-MCMC algorithm, exploiting the predictive distribution induced by the underlying random probability measure, and using the Wasserstein distance, in connection to the optimal transportation problem. Further, we proposed an adaptive strategy to avoid the arduous choice of the threshold ε, providing theoretical and numerical results as support. In extensive simulation studies we have shown that our proposal is a suitable choice in many contexts where the problem is hardly tractable or intractable. Despite its simplicity, we have obtained good performance for both computational cost and quality of the estimates, especially for the adaptive extension. The generality of the model allows us to work on abstract spaces, as shown for example in the case study described in Section 4.4.

Our algorithm suffers from the curse of dimensionality, as all the other MCMC algorithms for mixture models. In particular, when the dimension of the parameter space increases it becomes more and more difficult to propose suitable values for the cluster parameters, while when the dimension of the data increases both the observed and synthetic data suffer from sparsity. In this situation, we would advise to first project the data on a lower dimensional subspace, via, e.g., principal component analysis, and then perform model-based clustering on the lower dimensional projections.

ABC-MCMC arguably does not receive as much attention as ABC-SMC. In the context explored in this paper, the design of an ABC-SMC strategy is cumbersome due to the combinatorial nature of the problem. As argued in Bernton et al. (2019a), one possibility is to combine our ABC-MCMC proposal within an ABC-SMC scheme. However, the resulting strategy may lack effectiveness as the SMC particles do not tend to diversify, so that, essentially, one ends up using ABC-MCMC paying a higher computational price. We leave it as an interesting question for future investigation to design effective ABC-SMC schemes for nonparametric mixture models.

Several extensions are possible. On the algorithmic side, it could be interesting to add an *acceleration* step to sample the unique values, similar to the algorithms in Neal

(2000). This would bring our approach closer to the Gibbs-like algorithm in Clarté et al. (2020). As far as the model is concerned, we could consider generalizations beyond the exchangeable case. Note in fact that our approach holds only if the observations are assumed exchangeable. We believe that partial exchangeability could be easily dealt with, while dependence on continuous covariates such as in the PPMx model (Müller and Quintana, 2010) would require more work. Finally, it would be interesting to investigate the estimation of the group-specific parameters, either by assuming non-exchangeable prior distributions (see, e.g., Kunkel and Peruggia, 2020) or by developing post-processing procedures akin to the ones in Egidi et al. (2018).

Acknowledgments

Mario Beraha gratefully acknowledges the DataCloud laboratory (https://datacloud.polimi.it); experiments in Sections 4.2, 4.3 and 4.4 have been performed thanks to the Cloud resources offered by the DataCloud laboratory. Riccardo Corradin gratefully acknowledges the DEMS Data Science Lab for supporting this work through computational resources.

Funding

Mario Beraha acknowledges the support by MUR, grant Dipartimento di Eccellenza 2023-2027, and received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257.

Supplementary Material

Supplementary Material for "Bayesian nonparametric model based clustering with intractable distributions: an ABC approach" (DOI: 10.1214/24-BA1416SUPP; .pdf). The Supplementary Material contains the proofs of the theoretical results as well as additional simulation studies.

References

Altschuler, J., Niles-Weed, J., and Rigollet, P. (2017). "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration." In *Advances in Neural Information Processing Systems*, 1964–1974. 10

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 (Statistical Methodology)*, 64(2): 253–280. MR1904704. doi: https://doi.org/10.1111/1467-9868.00336. 3, 15, 18

Barthelmé, S. and Chopin, N. (2014). "Expectation propagation for likelihood-free inference." *Journal of the American Statistical Association*, 109(505): 315–333. MR3180566. doi: https://doi.org/10.1080/01621459.2013.864178. 2

Bassetti, F., Bodini, A., and Regazzini, E. (2006). "On minimum Kantorovich dis-

tance estimators." *Statistics & Probability Letters*, 76(12): 1298–1302. MR2269358. doi: https://doi.org/10.1016/j.spl.2006.02.001. 9

Beaumont, M. and Rannala, B. (2004). "The Bayesian revolution in genetics." *Nature Reviews Genetics*, 5: 251–261. 2

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). "Adaptive approximate Bayesian computation." *Biometrika*, 96(4): 983–990. MR2767283. doi: https://doi.org/10.1093/biomet/asp052. 2

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). "Approximate Bayesian computation in population genetics." *Genetics*, 162(4): 2025–2035. 2, 13

Beraha, M. and Corradin, R. (2024). "Supplementary Material for "Bayesian nonparametric model based clustering with intractable distributions: an ABC approach"", *Bayesian Analysis*, doi: https://doi.org/10.1214/24-BA1416SUPP. 3

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019a). "Approximate Bayesian computation with the Wasserstein distance." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2): 235–269. MR3928142. doi: https://doi.org/10.1111/rssb.12312. 9, 18, 22

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019b). "On parameter estimation with the Wasserstein distance." *Information and Inference: A Journal of the IMA*, 8(4): 657–676. 3, 9

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1(2): 353–355. MR0362614. 6

Bortot, P., Coles, S. G., and Sisson, S. A. (2007). "Inference for stereological extremes." *Journal of the American Statistical Association*, 102(477): 84–92. MR2345549. doi: https://doi.org/10.1198/016214506000000988. 2

Calvet, L. E. and Czellar, V. (2014). "Accurate methods for approximate Bayesian computation filtering." *Journal of Financial Econometrics*, 13(4): 798–838. 2

Camerlenghi, F., Corradin, R., and Ongaro, A. (2023). "Contaminated Gibbs-type priors." *Bayesian Analysis*, 1–30. 6

Cameron, E. and Pettitt, A. N. (2012). "Approximate Bayesian Computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift." *Monthly Notices of the Royal Astronomical Society*, 425(1): 44–65. 2

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). "SMC2: an efficient algorithm for sequential analysis of state space models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3): 397–426. MR3065473. doi: https://doi.org/10.1111/j.1467-9868.2012.01046.x. 15, 18

Clarté, G., Robert, C. P., Ryder, R. J., and Stoehr, J. (2020). "Componentwise approximate Bayesian computation via Gibbs-like steps." *Biometrika*. MR4298766. doi: https://doi.org/10.1093/biomet/asaa090. 23

Cuturi, M. (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, 2292–2300. Red Hook, NY, USA: Curran Associates Inc. 10

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I., and Ruggiero, M. (2013). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37. 6

de Finetti, B. (1937). "La prévision: ses lois logiques, ses sources subjectives." *Annales de l'Institut Henri Poincaré*, 7(1): 1–68. MR1508036. 4

Drovandi, C. and Frazier, D. T. (2022). "A comparison of likelihood-free methods with and without summary statistics." *Statistics and Computing*, 32(3): 42. MR4426803. doi: https://doi.org/10.1007/s11222-022-10092-4. 8

Dyer, J., Cannon, P., and Schmon, S. M. (2021). "Approximate bayesian computation with path signatures." *arXiv preprint arXiv:2106.12555*. 18

Egidi, L., Pappadá, R., Pauli, F., and Torelli, N. (2018). "Relabelling in Bayesian mixture models by pivotal units." *Statistics and Computing*, 28(4): 957–969. MR3766053. doi: https://doi.org/10.1007/s11222-017-9774-2. 23

Fearnhead, P. and Prangle, D. (2012). "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3): 419–474. MR2925370. doi: https://doi.org/10.1111/j.1467-9868.2011.01010.x. 2, 8

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2019). *Handbook of Mixture Analysis*. Chapman and Hall/CRC. MR3889980. 1

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis (Third Edition)*. Chapman and Hall/CRC. MR3235677. 15

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Sciences*, 138(3): 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 4, 6

Gu, J., Hua, B., and Liu, S. (2015). "Spectral distances on graphs." *Discrete Applied Mathematics*, 190: 56–74. MR3351721. doi: https://doi.org/10.1016/j.dam.2015.04.011. 20

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 4

James, L. F., Lijoi, A., and Prünster, I. (2009). "Posterior analysis for normalized random measures with independent increments." *Scandinavian Journal of Statistics*, 36(1): 76–97. MR2508332. doi: https://doi.org/10.1111/j.1467-9469.2008.00609.x. 5

Karabatsos, G. and Leisen, F. (2018). "An approximate likelihood perspective on ABC methods." *Statistics Surveys*, 12: 66–104. MR3812816. doi: https://doi.org/10.1214/18-SS120. 2

Kingman, J. F. C. (1978). "Random partitions in population genetics." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 361(1704): 1–20. MR0526801. doi: https://doi.org/10.1098/rspa.1978.0089. 5

Kunkel, D. and Peruggia, M. (2020). "Anchored Bayesian Gaussian mixture models." *Electronic Journal of Statistics*, 14(2): 3869–3913. MR4165496. doi: https://doi.org/10.1214/20-EJS1756. 23

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). "Markov chain Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences*, 100(26): 15324–15328. 2, 13

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113(521): 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 4, 6

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants." *Biometrika*, 93(2): 451–458. MR2278096. doi: https://doi.org/10.1093/biomet/93.2.451. 6

Morris, M., Handcock, M. S., and Hunter, D. R. (2008). "Specification of exponential-family random graph models: terms and computational aspects." *Journal of Statistical Software*, 24(4): 1548. 20

Müller, P. and Quintana, F. (2010). "Random partition models with regression on covariates." *Journal of Statistical Planning and Inference*, 140(10): 2801–2808. MR2651966. doi: https://doi.org/10.1016/j.jspi.2010.03.002. 23

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). "MCMC for doubly-intractable distributions." In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, 359–366. Arlington, Virginia, USA: AUAI Press. 6

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 6, 7, 15, 16, 22

Nguyen, T. D., Trippe, B. L., and Broderick, T. (2022). "Many processors, little time: MCMC for partitions via optimal transport couplings." In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 3483–3514. PMLR. URL https://proceedings.mlr.press/v151/nguyen22a.html 3

Peyré, G., Cuturi, M., et al. (2019). "Computational optimal transport." *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607. 3, 9, 10

Picchini, U. (2014). "Inference for SDE models via approximate Bayesian computation." *Journal of Computational and Graphical Statistics*, 23(4): 1080–1100. MR3270712. doi: https://doi.org/10.1080/10618600.2013.866048. 2

Pitman, J. (1995). "Exchangeable and partially exchangeable random partitions." *Probability Theory and Related Fields*, 102(2): 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 5

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 25(2): 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 4

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." *Molecular Biology and Evolution*, 16(12): 1791–1798. 2

Rastelli, R. and Friel, N. (2018). "Optimal Bayesian estimators for latent variable cluster models." *Statistics and Computing*, 28(6): 1169–1186. MR3850389. doi: https://doi.org/10.1007/s11222-017-9786-y. 15, 19

Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007). "Recent developments in exponential random graph (p*) models for social networks." *Social Networks*, 29(2): 192–215. 20

Rubin, D. B. (1984). "Bayesianly justifiable and relevant frequency calculations for the applied statistician." *The Annals of Statistics*, 12(4): 1151–1172. MR0760681. doi: https://doi.org/10.1214/aos/1176346785. 2

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. CRC Press. MR3889281. 2, 8

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). "Sequential Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences*, 104(6): 1760–1765. MR2301870. doi: https://doi.org/10.1073/pnas.0607208104. 2

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2009). "Correction for Sisson et al., Sequential Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences*, 106(39): 16889–16889. MR2301870. doi: https://doi.org/10.1073/pnas.0607208104. 2

Tavaré, S., Balding, D., Griffiths, R., and P., D. (1997). "Inferring coalescence times from DNA sequence data." *Genetics*, 145: 505–18. 2

Technow, F., Messina, C. D., Totir, L. R., and Cooper, M. (2015). "Integrating crop growth models with whole genome prediction through approximate Bayesian computation." *PLOS ONE*, 10(6): 1–20. 2

Vihola, M. and Franks, J. (2020). "On the use of approximate Bayesian computation Markov chain Monte Carlo with inflated tolerance and post-correction." *Biometrika*, 107(2): 381–395. MR4108936. doi: https://doi.org/10.1093/biomet/asz078. 13, 14, 15

Villani, C. (2008). *Optimal Transport – Old and New*, volume 338, xxii+973. MR2459454. doi: https://doi.org/10.1007/978-3-540-71050-9. 3

Wade, S. and Ghahramani, Z. (2018). "Bayesian cluster analysis: Point estimation and credible balls (with discussion)." *Bayesian Analysis*, 13(2): 559–626. MR3807860. doi: https://doi.org/10.1214/17-BA1073. 15

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). "Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood." *Genetics*, 182(4): 1207–1218. 13

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

Wilkinson, R. (2013). "Approximate Bayesian Computation (ABC) gives exact results under the assumption of model error." *Statistical Applications in Genetics and Molecular Biology*, 12: 129–141. MR3071024. doi: https://doi.org/10.1515/sagmb-2013-0010. 2