Received: 14 July 2021 | Revised: 1 December 2021 | Accepted: 8 December 2021

DOI: 10.1111/sjos.12575

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Learning the two parameters of the Poisson–Dirichlet distribution with a forensic application

Giulia Cereda | **Fabio Corradi** | **Cecilia Viscardi**

Dipartimento di Statistica, Informatica, Applicazioni (DISIA), University of Florence, Florence, Italy

Correspondence
Giulia Cereda, Dipartimento di Statistica, Informatica, Applicazioni (DISIA), University of Florence, Florence, Italy.
Email: giulia.cereda@unifi.it

Funding information
Italian Ministry of University and Research (MIUR), Grant/Award Number: 58523_DIPECC; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: P2LAP2 178195

Abstract

In forensic science, the rare type match problem arises when the matching characteristic from the suspect and the crime scene is not in the reference database; hence, it is difficult to evaluate the likelihood ratio that compares the defense and prosecution hypotheses. A recent solution consists of modeling the ordered population probabilities according to the two-parameter Poisson–Dirichlet distribution, which is a well-known Bayesian nonparametric prior, and plugging the maximum likelihood estimates of the parameters into the likelihood ratio. We demonstrate that this approximation produces a systematic bias that fully Bayesian inference avoids. Motivated by this forensic application, we consider the need to learn the posterior distribution of the parameters that governs the two-parameter Poisson–Dirichlet using two sampling methods: Markov Chain Monte Carlo and approximate Bayesian computation. These methods are evaluated in terms of accuracy and efficiency. Finally, we compare the likelihood ratio

All authors contributed equally to this work.

that is obtained by our proposal with the existing solution using a database of Y-chromosome haplotypes.

KEYWORDS

ABC, MCMC, likelihood ratio, rare type match problem, two-parameter Poisson–Dirichlet distribution

1 | INTRODUCTION

The two-parameter Poisson–Dirichlet (PD) distribution, denoted by PD(α, θ), is a well-known Bayesian nonparametric prior for ordered infinite-dimensional vectors $\mathbf{p} = (p_1, p_2, \ldots)$ belonging to the infinite simplex. Our interest in learning the parameters of this distribution arises from a forensic application (Cereda & Gill, 2020). One of the authors of this paper proposed to use the PD(α, θ) to model the ranked frequencies of Y-chromosome short tandem repeat (Y-STR) haplotypes in a population of possible donors of a DNA stain found at the crime scene. The final aim of that work was to calculate the likelihood ratio (LR) for the so-called "rare type match problem," which occurs when the crime stain Y-STR profile that matches the suspect's profile has not been observed in the reference population sample. Following the forensic science terminology (as well as we will do hereafter) they refer to LR as what, in Bayesian circles, is called Bayes factor (Dorp et al., 2020; Taroni et al., 2016) and show that it depends on the posterior distribution of the PD parameters. Since this posterior distribution is analytically intractable, they proposed a plug-in solution that is based on the maximum likelihood estimate (MLE) of the parameters. Here we demonstrate that this approximation provides a systematic overestimation of the LR, which is an especially unpleasant result in forensic individualization. Instead, we address the problem of dealing with the lack of knowledge on α and θ by evaluating their posterior distribution via Markov Chain Monte Carlo (MCMC) and approximate Bayesian computation (ABC) methods. Furthermore, we compare the performance of the two sampling schemes on simulated datasets.

This paper is structured as follows. In Section 2, we provide the essentials about the PD prior, the related Pitman sampling formula, and the Chinese restaurant representation. In Section 3, we illustrate the Y-STR rare type match problem that motivates our interest in exploring inference strategies and the need of using a fully Bayesian inference instead of the plug-in of the MLEs of the PD parameters. Section 4 deals with the implementation of the MCMC and ABC simulation methods. Section 5 provides some proof-of-concept experiments based on simulated data in a controlled experiment. Section 6 considers a real and large Y-STR collection of profiles and performs 50 identification activities by using the plug-in and the fully Bayesian approaches. Finally, we give our conclusions and make some suggestions for possible developments.

2 | THE TWO-PARAMETER POISSON–DIRICHLET DISTRIBUTION

The two-parameter PD distribution, PD(α, θ) with $\alpha \in (0, 1)$ and $\theta > -\alpha$, is a distribution over the infinite simplex of the form

$$\nabla_\infty = \left\{ (p_1, p_2, \ldots) | p_1 \geq p_2 \geq \cdots > 0, \sum_{i=1}^{+\infty} p_i = 1 \right\}.$$

Actually, it is a probability distribution over a set of discrete probability distributions. This distribution was first introduced by Pitman (1992) and is the generalization of the one-parameter PD distribution due to Kingman (1975), which is obtained by introducing a discount parameter.

Operationally, the PD distribution can be constructed in two steps:

1. For $\alpha \in (0, 1)$, and $\theta > -\alpha$, the vector $\mathbf{W} = (W_1, W_2, \ldots)$ is distributed according to GEM(α, θ) if $\forall i, W_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$, and $V_i \sim$ Beta$(1 - \alpha, \theta + i\alpha)$. GEM$(\alpha, \theta)$ is also known as stick breaking prior (Pitman, 2002).
2. The random vector $\mathbf{P} = (P_1, P_2, \ldots)$, obtained sorting the elements in $\mathbf{W}$ in decreasing order, has the PD(α, θ) distribution. Parameters α and θ are called the discount and the concentration parameters, respectively.

Insights into the role of the two parameters in the characterization of the shape of the distribution can be found in Navarrete et al. (2008), and De Blasi et al. (2015). One of the exciting features of PD is that it shows an asymptotic Zipf's-law distribution for the ordered frequencies (Goldwater et al., 2006), meaning that their probabilities are inversely proportional to their rank. This is the fundamental reason why we use it for the forensic application in Section 3. Notice that the literature also includes the value 0 in the definition of the parameter space for α, corresponding to the (one parameter) Poisson–Dirichlet distribution (Kingman, 1975). However, we do not include the value 0 because this implies a critical distributional difference from the two-parameter PD: more specifically, in such a case, the ranked frequencies do not asymptotically show the desired Zipf's law behavior (Broderick et al., 2012).

The one-parameter Poisson–Dirichlet distribution, corresponding to $\alpha = 0$, is related to the well-known Dirichlet process (DP) (Ferguson, 1973) and similarly, the two-parameter PD distribution finds an extension in the so-called Pitman Yor process (PYP). The DP and the PYP are used for hierarchical mixture modeling and clustering problems, and are both functional on distributions: they take as an input a measurable space with domain Ψ, and a distribution over it (called the *base distribution* $G(\cdot)$), and they yield a discrete distribution as an output with a countable set of possible values of the form

$$\sum_{i=1}^{k} p_i \delta_{\psi_i}(\cdot),$$

where $\psi_1, \psi_2, \ldots$ are i.i.d from G, and the vector $\mathbf{p}$ is distributed according to a PD$(0, \theta)$ for the Dirichlet process, or according to PD(α, θ) with $\alpha > 0$ for the PYP. Note that here we are interested in the PD distribution, modeling the probability of the ordered vector $\mathbf{p}$, and we are not interested in the entire process, which is also characterized by the base measure.

Another possible set of values for the two parameters traditionally associated to the PD is $\alpha < 0, \theta = -m\alpha$ for some positive integer m (Pitman, 1996). However, we are not interested in this choice because it is only suitable in the case of a finite k; hence, it is neither appropriate as a Bayesian nonparametric prior nor for the application at hand because we do not know in advance the number of types.

In what follows, we report two alternative characterizations of the PD(α, θ) distribution.

2.1 | Pitman sampling formula

Consider a sequence of integer-valued random variables, $I_1, \dots, I_n$, representing some characteristics of n units. Let the equivalence relation $i \sim j$ hold if and only if $I_i = I_j$ (i.e., the ith and the jth units have the same characteristics). The equivalence classes based on subsets of indices corresponding to the same value of I form a random partition of $[n] = \{1, 2, \dots, n\}$, which will be denoted as $\Pi_{[n]}(I_1, I_2, \dots, I_n)$. For instance, the following realization of a random partition

$$\pi_{[10]} = \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}\} \tag{1}$$

corresponds to $I_1 = I_3$, $I_2 = I_4 = I_{10}, I_5 = I_6$, while I_7, I_8, and I_9 are singletons. We have retained equalities and inequalities but we have lost information about the value of each I_i.

It holds that if $\forall n \in \mathbb{N}$

$$\begin{aligned} I_1, \dots, I_n | \mathbf{P} = \mathbf{p} &\overset{\text{i.i.d.}}{\sim} \mathbf{p}, \\ \mathbf{P} | \alpha, \theta &\sim \text{PD}(\alpha, \theta), \end{aligned} \tag{2}$$

then the random partition $\Pi_{[n]} = \Pi_{[n]}(I_1, \dots, I_n)$ has the following distribution:

$$\Pr(\Pi_{[n]} = \pi_{[n]} | \alpha, \theta) = \frac{[\theta + \alpha]_{k-1;\alpha}}{[\theta + 1]_{n-1;1}} \prod_{i=1}^{k} [1 - \alpha]_{n_i - 1;1}, \tag{3}$$

where n_i is the size of the ith class of $\pi_{[n]}$, k is the number of classes, and $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$,

$$[x]_{a;b} := \begin{cases} \prod_{i=0}^{a-1} (x + ib) & \text{if } a \in \mathbb{N} \setminus \{0\} \\ 1 & \text{if } a = 0. \end{cases}$$

The *Pitman sampling formula* (3), which was first derived by Pitman (1995), will be used as a likelihood for obtaining the MLE and the MCMC posterior distribution for (α, θ).

2.2 | Chinese restaurant representation

An alternative characterization of this model is called anecdotically the "Chinese restaurant process" (CRP), due to Aldous (1985) for the one-parameter case, and studied by Pitman (2006) for the two-parameter version.

Consider a restaurant with infinitely many tables, each table is infinitely large. Let $S_1, S_2, \dots$, be integer valued random variables representing the seating plan of the restaurant: $S_i = j$ means that the ith customer seats at the jth table. Let $S_1 = 1$. After that, the process depends on parameters $\alpha \in (0, 1)$, and $\theta > -\alpha$ and is described by the following conditional probability:

$$\Pr(S_{n+1} = j | S_1, \dots, S_n) = \begin{cases} \frac{\theta + k\alpha}{n + \theta} & \text{if } j = k + 1 \\ \frac{n_j - \alpha}{n + \theta} & \text{if } 1 \le j \le k, \end{cases} \tag{4}$$

where k is the number of tables occupied by the first n customers, and n_j is the number of customers already seating at table j. Clearly, $S_1, \dots, S_n$ are not i.i.d., nor exchangeable, but $\Pi_{[n]}(S_1, \dots, S_n)$ is distributed according to the Pitman sampling formula (3) (for a proof, see Pitman (2006)). This process provides the generative model required to perform ABC, see Section 4.2.

From Equation (4), it is apparent that allowing $\alpha < 0, \theta = -m\alpha$, implies that after the observation of m types, the probability of observing an unseen type becomes 0; hence, no new types can be observed, making the probability distribution inappropriate in the nonparametric framework.

2.3 | Review of the literature

In the literature, there are a few contributions to methods for learning the PD parameters. Carlton (1999) and Zhou et al. (2017) consider MLE for α and θ separately and jointly, providing (Carlton only) caveats on the consistency of the estimators under some conditions. Sibuya and Yamato (2001), following Carlton, show the suboptimality of MLE and present some alternative estimators. Hoshino (2001) deals with computational aspects of deriving MLE. In the Bayesian framework, Lijoi et al. (2008) pose a prior on the two parameters to evaluate a Bayes factor required for model choice, while Favaro et al. (2009), and Lijoi et al. (2007) solve the problem through an empirical Bayes estimation.

In Jara et al. (2010), Carmona et al. (2018), and Murphy et al. (2019), MCMC inference on the parameters of a PYP mixture models is explored providing posterior uncertainty on the cluster locations and their relative sizes. Earlier contributions can be found in West (1992), Escobar (1994), Escobar and West (1995), and Escobar and West (1998) for the DP mixture model, where they derive (also) the full conditional for the parameter α.

Lijoi et al. (2008) attempted to derive the posterior for two parameters PD (α, θ), via a Gibbs scheme. However, the computation of their full conditionals requires us to resort to a discretization for the two parameters. To avoid constraining the choice of the prior to convenience reasons, which misrepresents the parameter's support, we choose to perform a MH inference.

As far as we know, there are no attempts in the literature to derive inference for the PD parameters via ABC.

3 | THE RARE TYPE MATCH PROBLEM

In a forensic case, the characteristic of a stain found on the crime scene turns out to match the characteristic of a suspect. Let us name as D the two profiles along with a database of reference containing the characteristics of n statistical units. Throughout this paper, the database is assumed to be a random sample from the population of possible donors, even though this assumption might not always be satisfied.

To "weight the data" D under the prosecution's (identification) and the defense's (no-identification) hypotheses, h_p and h_d, the forensic statistician has to calculate the likelihood ratio, defined as:

$$\mathrm{LR} = \frac{\Pr(D|h_p)}{\Pr(D|h_d)}. \tag{5}$$

In this basic forensic setting, the condition most favorable to identification happens if the suspect's profile and the crime stain's profile are identical (i.e., whenever they match). The support to the identification depends on the rarity of the trait in the database: the rarer the evidence is, the stronger the support to h_p will be.

The rare type match case is the situation in which the profile shared by the suspect and the stain is not among those contained in the database of reference. In this framework, the rarity principle is not operational because there are no frequencies to empirically evaluate the probability of observing the characteristic exhibited by the crime and the suspect sample. Actually, one can always enlarge the dataset with the crime's sample (or equivalently, the suspect's sample) as recommended by, for example, Dawid and Mortera (1996), but the estimate of the probability of the matching characteristic would still be based on one observation only.

This state of affairs is not unusual because the evidence considered for forensic purposes (e.g., tire marks, glass fragments, and others) is often made of a large and unknown number of features. The same happens when using Y-STR profiles, which are made of a number (varying from 7 to 23) of STR polymorphisms belonging to the nonrecombining part of the Y-chromosome. The lack of recombination implies that there is no biological reason to assume independence among Y-STR loci. This dependency, which was confirmed by Caliebe et al. (2015), makes the available databases too small with respect to the very large number of possible profiles. This does not represent a problem when dealing with autosomal STR profiles, for which the assumption of independence among the loci holds and allows considering each locus separately.

To address this problem, we propose a "change of glasses strategy" that considers only the event that a never observed characteristic has been observed twice as relevant, and hence contributes to the augmented database partition with a further class of size two. What becomes of interest is how rare is to observe two traces with the same characteristic in the augmented database, with no mention of its value which becomes irrelevant.

This change of glasses strategy ignores information about the type of each observation, and allows us to model the ordered vector $\mathbf{p}$ of the relative frequencies as a PD distribution. The reasons that motivate this choice are detailed in Cereda and Gill (2020), and are fundamentally due to the Zipf's law behavior shown by the Y-STR ordered frequencies in the YHRD database. In addition, assuming that the number of possible Y-STR haplotypes is infinite simplifies our task because we are not required to fix their number in advance and we are always ready to accommodate for the observation of a not yet observed type.

Consider the partition $\pi_{[10]}$ shown in Section 2.1, representing a small database of 10 individuals. By augmenting the database with the crime stain and suspect's characteristics, in the rare type match case we would obtain:

$$\pi_{[12]} = \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}, \{11, 12\}\}.$$

The 11th and the 12th characteristics (of the suspect's and the crime stain's) constitute a new class by themselves because they are equal to each other but different from all of those previously observed. The change of glasses allows one to focus on the classes of the partition; for instance, disregarding the genetic information and taking into account only equalities and inequalities among Y-STR profiles. In this framework, the computation of (5) requires us to evaluate the probability of the partition conditionally to the hypothesis $H \in \{h_p, h_d\}$. In fact, the hypotheses matter: when $H = h_p$, the suspect and the crime characteristics belong to the same person and are actually two observations of the suspect's characteristic; when $H = h_d$, they are two equal characteristics belonging to two different persons.

Let $\pi_{[n+1]}$ denote the partition obtained from the database augmented with the suspect's characteristic, and $\pi_{[n+2]}$ denote the partition obtained by adding the crime stain in the same class as the suspect sample. Since defense and prosecution agree about the origin of the first $n+1$ observations and they also assume the same model, $\Pi_{[n+1]} \perp\!\!\!\perp H|\alpha, \theta$ holds. In such a case, the likelihood ratio becomes

$$\begin{aligned}
\text{LR} &= \frac{p(\pi_{[n+2]}|h_p)}{p(\pi_{[n+2]}|h_d)} \\
&= \frac{p(\pi_{[n+2]}, \pi_{[n+1]}|h_p)}{p(\pi_{[n+2]}, \pi_{[n+1]}|h_d)} \quad \text{because } \pi_{[n+1]} \subset \pi_{[n+2]} \\
&= \frac{p(\pi_{[n+2]}|\pi_{[n+1]}, h_p)p(\pi_{[n+1]}|h_p)}{p(\pi_{[n+2]}, \pi_{[n+1]}|h_d)} \\
&= \frac{1 \cdot p(\pi_{[n+1]}|h_p)}{p(\pi_{[n+2]}, \pi_{[n+1]}|h_d)} \\
&= \frac{\int p(\pi_{[n+1]}, \alpha, \theta)d\alpha, d\theta}{\int p(\pi_{[n+2]}|\pi_{[n+1]}, \alpha, \theta, h_d)p(\pi_{[n+1]}|\alpha, \theta)p(\alpha, \theta)d\theta d\alpha} \qquad (6) \\
&= \frac{1}{\int \frac{1-\alpha}{n+1+\theta}\, p(\alpha, \theta|\pi_{[n+1]})d\alpha d\theta}, \qquad (7)
\end{aligned}$$

where (a) $p(\pi_{[n+2]}|\pi_{[n+1]}, h_p) = 1$ because according to h_p the suspect and the crime stain characteristics must be equal, and hence belong to the same class; (b) (6) derives from $\Pi_{[n+1]} \perp\!\!\!\perp H|\alpha, \theta$; and (c) the last line comes from (4) and Bayes' theorem.

This result (7) is formally reminiscent of the LR employed in usual forensic identification whenever the crime stain and the suspect characteristics coincide, as also shown by Dawid (2017, Sect 4.1.2). Even there, the LR is obtained by integrating the probability of observing the crime stain's evidence with respect to the posterior distribution of unknown population parameters given the database enlarged with the suspect's characteristic. In our change of glasses perspective, one has to find the marginal probability of the event of observing the $(n+2)$th characteristic, identical to the $(n+1)$th, both never observed before, by integrating with respect to the PD parameters. Using a PD distribution, the probability of this event, conditionally to the model parameters, is provided by (4) (bottom line with $n_i = 1$) and is equal to $\frac{1-\alpha}{n+1+\theta}$, further mixed by the posterior of α and θ, conditioning on $\pi_{[n+1]}$.

The fully Bayesian evaluation of the LR (7) differs from the proposal of Cereda and Gill (2020), who proceeded to plug-in the MLE estimates of the PD parameters derived by the likelihood (3), obtaining:

$$\text{LR}_{\text{MLE}} = \frac{n+1+\theta_{\text{MLE}}}{1-\alpha_{\text{MLE}}}. \qquad (8)$$

Expression (8) is an heterogeneous result because the LR comes from a Bayesian formulation but its evaluation is obtained by considering the parameters as fixed quantities estimated by MLE. In this work, the evaluation of the MLE for α and θ has been obtained numerically by the routine *nlm* available in R, using Newton-type algorithms (see Dennis and Schnabel (1996)). This is not computationally intensive.

Actually, the question is whether the plug-in approach (8) produces a good approximation of (7) or introduces a well-characterized bias in the evaluation of the LR compared to the fully

Bayesian solution. The issue seems a specific instance of a contrived dispute opposing two approaches to inference but, in this case, we believe that a reason motivating the preference for the Bayesian solution does exist, as detailed in Section 3.1.

3.1 | The plug-in and the fully Bayesian solution compared

Let us define $\phi(\alpha, \theta) = \frac{n+1+\theta}{1-\alpha}$. The fully Bayesian solution (7) can be estimated by resorting to a MC estimate of the integral. Accordingly,

$$\mathrm{LR}_{\mathrm{MC}} = \frac{1}{\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}} \frac{1-\alpha_i}{n+1+\theta_i}} = \frac{n_{sim}}{\sum_{i=1}^{n_{sim}} \phi(\alpha_i, \theta_i)^{-1}}, \tag{9}$$

which corresponds to the harmonic mean of $\phi(\alpha, \theta)$ evaluated using pairs (α_i, θ_i), $i \in \{1, ..n_{sim}\}$, drawn from the posterior.

If the posterior distribution of the two parameters was concentrated around $(\alpha_{\mathrm{MLE}}, \theta_{\mathrm{MLE}})$, then the estimate in (8) would represent the first-order Taylor approximation of the expected value of $\phi(\theta, \alpha)$. This suggests that, given n_{sim} samples from the posterior, the arithmetic mean $\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}} \phi(\alpha_i, \theta_i)$ would be very similar to $\mathrm{LR}_{\mathrm{MLE}}$.

Recalling that the harmonic mean is always smaller than the arithmetic mean, $\mathrm{LR}_{\mathrm{MLE}}$ represents a systematic overestimation of the fully Bayesian solution (7).

This implies that (8) provides a greater and unjustified support to identification, and is consequently not desirable in the forensic setting. In other words, only if the posterior distribution is strictly concentrated around $(\alpha_{\mathrm{MLE}}, \theta_{\mathrm{MLE}})$, that is, when $p(\alpha, \theta|\pi_{[n+1]}) \approx \delta_{(\alpha_{\mathrm{MLE}}, \theta_{\mathrm{MLE}})}(\alpha, \theta)$ do the two estimates approximately coincide. In such a case, the computational effort of sampling from the posterior could be avoided by resorting to the plug-in estimate. In all other cases, a proper evaluation of the Bayesian solution in (7) requires computation of the MC estimate in (9) based on simulations from the posterior distribution.

Similarly, Meester and Slooten (2021) show that in the standard context of identification (i.e., when the specific characteristic of the matching trace is taken into account), it holds that

$$\begin{aligned}\mathrm{LR} &= \frac{\int p(X_s = x|\theta, h_p)p(\theta)d\theta}{\int p(X_c = x|X_s = x, \theta, h_d)p(X_s = x|\theta, h_d)p(\theta)d\theta} = \frac{E(\theta)}{E(\theta^2)} \\ &= \frac{E(\theta)}{(E(\theta))^2 + var(\theta)} = \frac{1}{E(\theta) + \frac{var(\theta)}{E(\theta)}},\end{aligned}$$

where X is the characteristic of the matching profile, $p(X = x|\theta) = \theta$, $\theta \sim p(\theta)$, and $X_c = X_s = x$ are the crime and the suspect sample, respectively.

Therefore, also in the standard context, if we replace $E(\theta)$ with θ_{MLE} and we ignore the parameter's uncertainty and its variability, we obtain an overestimation of the LR.

4 | INFERENCE OF THE PD PARAMETERS

In this section, we specify how to make inference on α and θ by considering that a solution in closed form does not exist. In this case, a likelihood and a generative model are both available.

Therefore, we planned to explore the use of MCMC and ABC sampling scheme to compare their respective merits and demerits.

4.1 | Markov Chain Monte Carlo

Nowadays, MCMC methods (Robert & Casella, 2013), especially Gibbs sampler and Metropolis Hastings (MH), are the main tool to perform calculations in Bayesian inference. The MH algorithm generates samples from the posterior distribution of a set of parameters, which are generically indicated by γ and indexing a possible multivariate random variable X through a specific model, say $p(x|\gamma)$, by building an ergodic Markov chain through an acceptance-rejection mechanism: at each iteration, t, the proposal γ^* is accepted or rejected after a comparison with the last accepted parameter, γ_{t-1}. The algorithm relies on three ingredients: (a) the likelihood function, $\ell(\gamma; x)$, providing the link between the model parameters, γ, and the data, x; (b) the prior distribution, $p(\gamma)$, and (c) the proposal distribution, $q(\gamma^*|\gamma_{t-1})$. Upon assessment of the convergence, the result is a sample from the posterior distribution of γ.

MCMC for PD(α, θ). For a budget of T simulations, Algorithm 1 summarizes how to implement MH for the parameters $\gamma = (\alpha, \theta)$ of the PD distribution, conditionally to an observed random partition $\pi_{[n]}$.

Algorithm 1. MH

Initialize $\alpha_0 \sim p(\alpha), \theta_0 \mid \alpha_0 \sim p(\theta|\alpha_0)$,
for $t = 1, \dots, T$ **do**
Draw $\alpha^*, \theta^* \sim q(\alpha, \theta|\alpha_{t-1}, \theta_{t-1})$
Evaluate the ratio

$$R = \frac{p(\alpha^*, \theta^*|\pi_{[n]})}{p(\alpha_{t-1}, \theta_{t-1}|\pi_{[n]})} \frac{q(\alpha_{t-1}, \theta_{t-1}|\alpha^*, \theta^*)}{q(\alpha^*, \theta^*|\alpha_{t-1}, \theta_{t-1})}$$

Set $(\alpha_t, \theta_t) = (\alpha^*, \theta^*)$ with probability equal to $\min(R, 1)$
end for

The acceptance ratio R in Algorithm 1 can be expanded as

$$R = \frac{p(\alpha^*, \theta^*)}{p(\alpha_{t-1}, \theta_{t-1})} \frac{\ell(\alpha^*, \theta^*; \pi_{[n]})}{\ell(\alpha_{t-1}, \theta_{t-1}; \pi_{[n]})} \frac{q(\alpha_{t-1}, \theta_{t-1}|\alpha^*, \theta^*)}{q(\alpha^*, \theta^*|\alpha_{t-1}, \theta_{t-1})},$$

where the intractable posterior normalizing constants cancel. We propose the following prior and proposal distributions:

1. A joint prior $p(\alpha, \theta) = p(\theta|\alpha)p(\alpha)$ made of:
   - A vague prior for α: $\alpha \sim \text{Unif}(0, 1)$.
   - A truncated Gaussian distribution for $\theta|\alpha$:

$$\theta|\alpha,\tilde{\theta} \sim \overline{N}\left(\tilde{\theta}, 2|\tilde{\theta}|, (-\alpha, +\infty)\right) \tag{10}$$

where $\tilde{\theta}$ is a guess on the location of θ with variance $2|\tilde{\theta}|$. The variance of θ accounts for the difficulty in eliciting this prior according to the magnitude of the parameter spanning in $(-\alpha, \infty)$. This choice attempts to reduce the effect of a largely inaccurate prior specification, which more easily occurs for a distribution on θ posing the bulk of the probability around high values of the variable. Furthermore, our choice requires to set only one hyperparameter, whereas other existing approaches involve two parameters: for the prior on θ, Jara et al. (2010) specify the mean and the variance of a truncated normal, while Carmona et al. (2018) rely on a shifted Gamma distribution.

2. A proposal distribution

$$q(\alpha^*, \theta^*|\alpha_{t-1}, \theta_{t-1}) = q(\alpha^*|\alpha_{t-1})q(\theta^*|\theta_{t-1}, \alpha^*)$$

made of:

- A reflecting random walk for $\alpha^*|\alpha_{t-1}$ (see Hoff (2009)) assuring that $\alpha^* \in (0, 1)$ by sampling

$$\alpha'|\alpha_{t-1} \sim \text{Unif}(\alpha_{t-1} - \delta, \alpha_{t-1} + \delta),$$

$$\alpha^*|\alpha' = \begin{cases} |\alpha'|, & -1 < \alpha' \leq 1 \\ 2 - \alpha', & 1 < \alpha' \leq 2 \end{cases}$$

where $0 < \delta < 1$ is the tuning parameter, here set as $\delta = 0.1$, so that we move tightly around the previous accepted α. This seems a good compromise between a reasonable speed in exploration and the need to avoid to search too far from an already accepted proposal.

- A truncated Gaussian for $\theta^*|\theta_{t-1}, \alpha^*$:

$$\theta^*|\theta_{t-1}, \alpha^* \sim \overline{N}(\theta_{t-1}, |\theta_{t-1}|, (-\alpha^*, +\infty)).$$

Also Jara et al. (2010) employ a truncated Gaussian distribution, while Carmona et al. (2018) propose $\theta^*|\theta_{t-1} \sim U(\theta_{t-1} - \phi, \theta_{t-1} + \phi)$ which has the advantage of providing a symmetric transition kernel at the cost of specifying a further tuning parameter (ϕ).

4.2 | Approximate Bayesian computation

ABC denotes a class of methods for Bayesian inference on model parameters, ruling out either the analytical or numerical evaluation of the likelihood whenever they are unfeasible. A comprehensive overview on the ABC methods can be found in Sisson et al. (2018).

The main requirement of ABC is the possibility to simulate synthetic-data from a *generative model* (i.e., a computer program reproducing the generative data process) once fed with an instance of the model parameters. The primal rejection ABC converts samples from the prior into samples from the posterior distribution of the parameters by retaining only those values that when provided to the generative model produce synthetic data (y) identical to those observed (x). In this

way, the likelihood is empirically evaluated through a Monte Carlo estimate of the probability that each value of the parameter(s) could lead to simulations identical to the observed data. The event $y = x$ occurs with positive probability only when observed data live in a discrete space: otherwise, the generative model has no chance to reproduce exactly the original data. In addition, the event $y = x$ is very rare in high-dimensional discrete spaces, hence leading to discarding many simulations. The solution to this problem is to accept a certain degree of approximation by relaxing the $y = x$ requirement, and introducing a discrepancy function $d(x, y)$ and a tolerance threshold, ϵ, so that a parameter proposal is accepted if $d(x, y) < \epsilon$. For $\epsilon \to 0$, the primal rejection ABC is recovered, and there is no approximation. The same holds if the comparison between the observed and the simulated data is realized by means of summary statistics, s, sufficient for the parameters (see Sisson et al. (2018, Ch 5)). If a (set of) sufficient statistics is not known, then some others, heuristically chosen, may be proposed by introducing a further source of approximation. Several improvements on the ABC sampling scheme have been proposed (e.g., see Beaumont (2019), and Lintusaari et al. (2017) for a review). Nevertheless, here we use a simple rejection scheme as the end-of-scale of the potentiality of ABC.

ABC for PD(α, θ). Concerning the implementation of ABC for the parameters α and θ of the PD distribution, we denote by $\pi_{[n]}^{\text{obs}}$ the observed partition and by $\pi_{[n]}^{(t)}$ the synthetic partition simulated at iteration t. The generative model is provided by the CRP, sequentially allocating each individual, according to a probability defined by (4), either to an existing class of the partition or to a new class. The choice of the summary statistics and the distance function is crucial. Looking at (3), it is apparent that the partition of n, $\pi_n := (n_1, \dots, n_k)$, is a sufficient statistic for the two parameters. It follows that by comparing π_n^{obs} and $\pi_n^{(t)}$, we might avoid the approximation deriving from heuristically chosen statistics. However, the ABC approximation error's asymptotic expression suggests that high-dimensional summary statistics give poor results (Barber et al., 2015; Fearnhead & Prangle, 2012). Generally, the unavailability of low-dimensional sufficient summary statistics leads to the curse of dimensionality. This problem becomes more relevant in our setting by considering that the sufficient summary statistic π_n is potentially infinite-dimensional. Accordingly, by naming k^{obs} and $k^{(t)}$ the length of π_n^{obs} and $\pi_n^{(t)}$, respectively, we propose to mitigate the curse of dimensionality by comparing observed and simulated data through:

- A distance between partitions:

$$d_1(\pi_n^{(t)}, \pi_n^{\text{obs}}) = \frac{1}{k^*} \sum_{i=1}^{k^*} \log^2 \frac{n_i^{(t)}}{n_i^{\text{obs}}} \tag{11}$$

  where $k^* = \min(k^{\text{obs}}, k^{(t)})$.

  Distance in (11) has the following properties:

  - $d_1(\cdot, \cdot) = 0$ if and only if $n_i^{\text{obs}} = n_i^{(t)}, \forall i \in \{1, \dots, \min(k^{\text{obs}}, k^{(t)})\}$;
  - $d_1(\cdot, \cdot)$ is symmetric.

- A relative distance between the number of classes in the partitions:

$$d_2(k^{(t)}, k^{\text{obs}}) = \frac{|k^{(t)} - k^{\text{obs}}|}{k^{\text{obs}}}.$$

  - $d_2(\cdot, \cdot) = 0$ iff $k^{\text{obs}} = k^{(t)}$.

Algorithm 2 summarizes the resulting ABC algorithm.

Algorithm 2. ABC

Draw $(\alpha^{(t)}, \theta^{(t)}) \sim p(\alpha, \theta)$ $\quad t \in 1, \dots, T$
Generate $\pi^{(t)}_{[n]} \sim p(\cdot|\alpha^{(t)}\theta^{(t)})$ from the CRP $\quad t \in 1, \dots, T$
Accept $(\alpha^{(t)}, \theta^{(t)})$ if $d_1(\pi^{(t)}_n, \pi^{\text{obs}}_n)<\epsilon_1$ and $d_2(k^{(t)}, k^{\text{obs}})<\epsilon_2$ $\quad t \in 1, \dots, T$

The output of Algorithm 2 is a sample from the approximate distribution of $\theta, \alpha, \pi_{[n]}|\pi^{\text{obs}}_{[n]}$.

Marginalizing by disregarding the simulated summary statistics, the output of the algorithm becomes a sample from the approximate marginal posterior distribution of θ, α given $\pi^{\text{obs}}_{[n]}$. Note that due to the sufficiency of π_n and to the acceptance criterion, because $(\epsilon_1, \epsilon_2) \to (0, 0)$ there is no approximation in the distribution of $\alpha, \theta|\pi^{\text{obs}}_{[n]}$ achieved by ABC. The thresholds (ϵ_1, ϵ_2) will be chosen through the quantile criterion (Beaumont et al., 2002), as detailed in Section 5.1.

5 | INFERENCE FROM CONTROLLED DATA

The aim of this section is to provide details on three experiments concerning the inference on PD parameters. Each experiment is based on n i.i.d. observations, simulated from a distribution $\mathbf{p}$ obtained as a realization of a PD with three pairs of known parameters α^{true} and θ^{true}, as detailed in Table 1.

5.1 | MCMC and ABC inference comparison

For each population, inference on α and θ has been obtained according to Section 4. This only requires us to provide the value of $\tilde{\theta}$, the location of the prior on θ because α is assumed distributed as a Unif$(0, 1)$. The values of $\tilde{\theta}$ in Table 1 allow us to evaluate how ABC and MCMC react to a location of the prior close to or far from θ^{true}.

In detail, we have run:

TABLE 1 Parameters $(\alpha^{\text{true}}, \theta^{\text{true}})$ characterizing three populations

P1		P2		P3	
α^{true}	θ^{true}	α^{true}	θ^{true}	α^{true}	θ^{true}
0.5	20	0.1	10	0.7	5
$E(K_n) = 247.41$		$E(K_n) = 59.36$		$E(K_n) = 291.43$	
$\tilde{\theta} = \{1, 25\}$		$\tilde{\theta} = \{1, 12\}$		$\tilde{\theta} = \{4, 30\}$	

Note: For each pair, we provide the expected number of classes, when $n = 10^3$. For each population, two different hyperparameters $\tilde{\theta}$ are proposed.

1. A CRP with $n = 10^3$ customers with parameters $(\alpha^{\text{true}}, \theta^{\text{true}})$ to simulate a partition $\pi_{[n]}$ from the corresponding population.
2. A MCMC, as detailed in Algorithm 1, for a budget of 10^6 iterations. We have operated a burn-in and a thinning as suggested in Raftery and Lewis (1992) to retain a number of almost uncorrelated simulations from the posterior.
3. An ABC algorithm, as detailed in Algorithm 2, with ϵ_1 and ϵ_2 determined as the first percentile of the empirical distributions of $d_1(\cdot, \pi_{[n]}^{\text{obs}})$ and of $d_2(\cdot, k^{\text{obs}})$.

Comparisons are made in terms of the accuracy of the density estimation, computational time and effective sample size (ESS).

Remark 1 (Density estimation). Figure 1 illustrates that, in almost every circumstance, the ABC and MCMC posterior distributions are close to each other, showing that ABC and MCMC lead to similar inferences. This judgment is supported by the Hellinger distances (Pardo, 2018) between ABC and MCMC distributions, whose values are displayed in Figure 1, on the top of each subfigure. Hellinger distances, which are defined in $[0, 1]$, show values on the narrow range [0.061-0.212]. The largest values occur when priors largely disagree with the information on the parameters provided by the evidence because the rejection ABC draws parameter proposals from the prior. A proposal distribution far from the target results in a high rejection rate or in a bad approximation when ϵ is chosen, as we did, as a quantile of the empirical distributions of the distances. For example, in the first population, where $\theta^{\text{true}} = 20$, the threshold ϵ_1 doubles from 0.016 ($\tilde{\theta} = 25$) to 0.031 ($\tilde{\theta} = 1$). In contrast, in the MCMC, the proposal stage does not depend on the prior and is adapted according to the last accepted parameters values.

Remark 2 (Prior distribution). For the PD model, it is not easy to elicit priors because it is difficult to figure out how α and θ jointly affect the realization of a partition $\pi_{[n]}$. A solution is to assume an Empirical Bayesian (EB) approach by specifying priors through the MLE of the parameters, as done in Lijoi et al. (2007), and Favaro et al. (2009). An alternative to plug-in the MLE solution could be to resort to a matching moment strategy. However, the drawback is that a sample far from being representative of the population would act on the inference twice. An application of the EB approach can be found in Section 5.2.

Remark 3 (Computational time and ESS). Table 2 displays the time required to obtain 10^6 simulations and the achieved ESS by using one core of an INTEL® i9 laptop. In two cases, ABC requires more than twice the time employed by MCMC. In some others, the two computing times do not differ so much. However, by using eight cores on the same machine and exploiting the ABC attitude to parallel computation, the required time becomes about one-fifth of the time required by MCMC. This latter shows some variability in the execution times because different values of α^{true} and θ^{true} generate partitions of different sizes. When the number of the classes in the partitions is larger, the complexity of a point-wise likelihood evaluation will be greater. This step is completely avoided by ABC. Thus, given a budget of iterations, ABC outperforms MCMC in terms of computational effort once parallel computing is exploited. However, we note that MCMC achieves an ESS largely greater than ABC. Even if the number of retained simulations affects the accuracy of the parameters' posterior densities, we wonder if such a large number of iterations is really required for MCMC. The answer to this question is somewhat troublesome because it implies a fictitious experiment consisting of two steps: (a) run an MCMC until convergence, operating the burn-in and the thinning until an almost uncorrelated number of simulations is retained; and (b) run an ABC until a roughly equal number of simulations as in MCMC is retained. For example, in

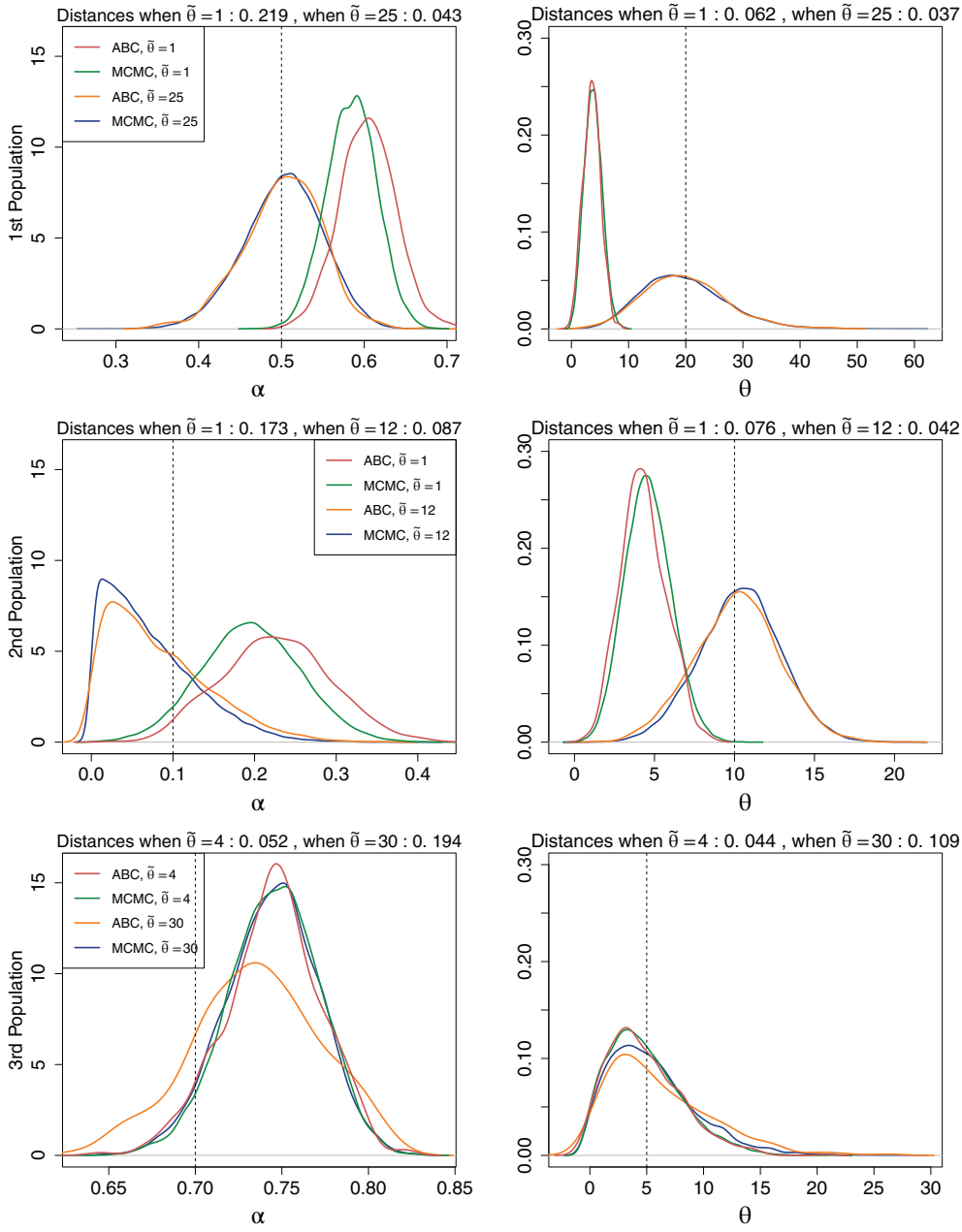

FIGURE 1 Inference on (α, θ) for populations displayed in Table 1. $\tilde{\theta}$ and Hellinger distance are specified at the top of each plot [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Running time (seconds) and effective sample size for three populations and two different hyperparameters $\tilde{\theta}$ (10^6 iterations)

		P1		P2		P3	
	$\tilde{\theta}$	**1**	**25**	**1**	**12**	**4**	**30**
Time	MCMC	233	229	82	82	254	252
	ABC	202	235	206	231	200	212
	ABC (8 cores)	43	44	42	45	38	39
ESS	MCMC	16k	42k	42k	35k	8k	9k
ABC	ABC	470	575	479	3154	687	220

the first population, setting $\tilde{\theta} = 1$, MCMC required 20,000 iterations and 5 seconds to retain 1500 almost uncorrelated simulations. To obtain a similar ESS, ABC required three million simulations and 1020 and 104 s by using one or eight cores, respectively. The same test for the other populations led to similar results. Moreover, a comprehensive comparison between the two methods would require us to quantify the effort: (a) to specify the full conditionals or the MH steps along with the activities required to assess the convergence of the chain for MCMC; and (b) to implement a valid generative model, to select summary statistics, distance functions and the tolerance threshold for ABC. This is clearly difficult to quantify but it is a relevant matter depending on the case at hand.

Remark 4 (Scalability). Looking at Table 2, it is apparent that the running times of ABC are not sensitive to changes of populations. However, the same does not hold for MCMC, whose computational times are affected by k, the number of classes in the observed partition whose expected value varies with α^{true} and θ^{true} according to the formula $E(K_n) = \frac{[\theta+\alpha]_{n:1}}{\alpha[\theta+1]_{n-1:1}} - \frac{\theta}{\alpha}$ (Pitman, 2006) (see the expected values in Table 1). In fact, in the first and third population, where the MCMC computational times are similar, the expected values of k are close. This is due to the fact that, as already noted in the previous remark, MCMC requires an evaluation of the likelihood function in (3), which is more computationally costly for higher values of k. Anyway, k depends also on the size of the dataset n. Hence, we speculate that ABC methods scale better than MCMC, also with respect to the sample size.

5.2 | Assessing the effect of the sampling variability

To evaluate the effect of sampling variability, we gained some experience by sampling 50 partitions from the three populations using the PD parameters already displayed in Table 1, from which we derived inference by using MCMC and ABC. For the sake of brevity, the results are only displayed for Population 1—the others are very similar. The inference is obtained by specifying three prior distributions:

1. We chose $\alpha \sim \text{Unif}(0, 1)$ and, according to (10), pose $\tilde{\theta} = 25$ corresponding to a small but not negligible divergence of the prior location from $\theta^{\text{true}} = 20$.
2. The parameter α is still $\sim \text{Unif}(0, 1)$ but, now, $\tilde{\theta} = 1$ corresponds to a large divergence of the prior location from θ^{true}.

3. According to the EB paradigm we assume $\theta \sim \overline{N}(\theta_{\text{MLE}}, 2|\theta_{\text{MLE}}|, [-\alpha, \infty])$. $\alpha \sim Beta(a, b)$ where $b = 2$ and a : Mode$(\alpha) = \alpha_{\text{MLE}}$.

Figure 2 shows the intervals of the CDFs (5th and the 95th percentile) over 50 experiments.

The most important result is that, independently of the priors, ABC and MCMC largely overlap their 90% CDF bands, showing a similar reaction to sampling variability. As a comment, the lowest overlap occurs when the prior location for θ largely disagree respect to θ^{true} (see Figure 2 in the middle). As already discussed in Section 5.1, in such conditions the rejection ABC rarely proposes candidates for θ close to the bulk of the posterior distribution, and this produces a distortion with respect to MCMC. A more sophisticated ABC version (e.g., Population Monte Carlo ABC) is expected to solve the problem. Regarding the sampling variability, the EB approach produces posteriors that are correctly located (see Figure 2 top and bottom). In addition, the sampling variability, which can be evaluated by looking at the thickness of the bands, does not vary appreciably with respect to what happens by using a prior located not very far from θ^{true}. This suggests that EB is a viable solution to specify priors in case of PD.

6 | EXPERIMENTS

In this section, we propose an experiment based on real data using the YHRD database, a collection of 18'925 7-loci (DYS19, DYS389 I, DYS389 II, DYS3904, DYS3915, DY3926, and DY3937) Y-STR profiles gathered in 51 European countries as detailed by Purps et al. (2014). The PD model shows a good fit with the ordered relative frequencies of the very large number of different

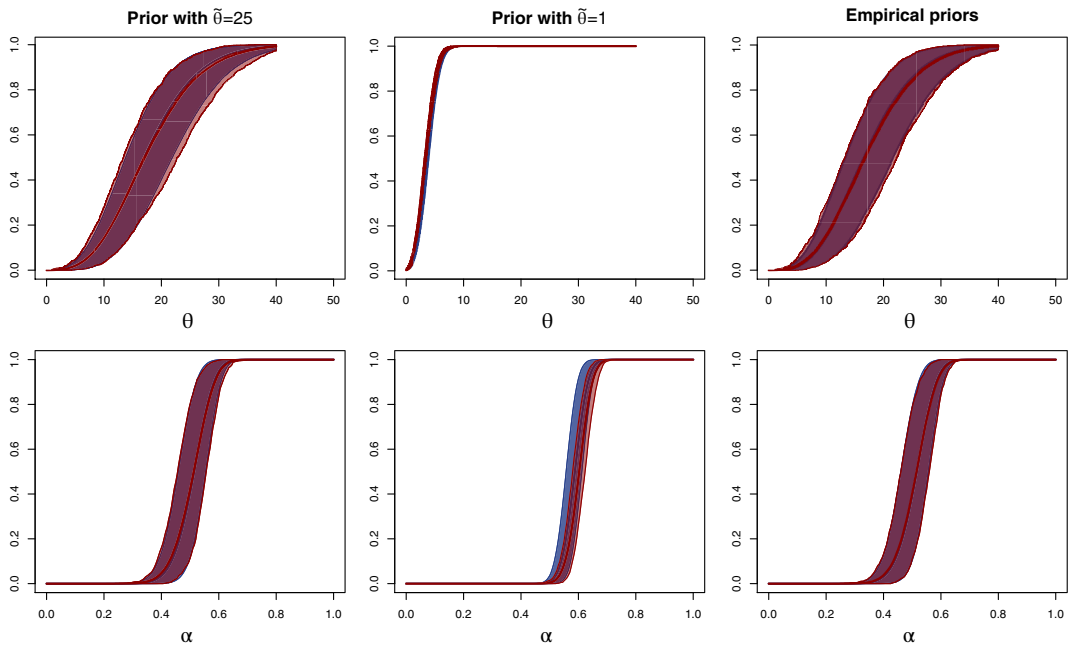

FIGURE 2 CDF 90% intervals for the PD parameters based on 50 samples from a PD(0.5, 20) population: ABC (red), MCMC (blue), overlapping CDF (brown) [Colour figure can be viewed at wileyonlinelibrary.com]

profiles in the YHRD database (an approximation of the infinite number required by the PD) (Cereda & Gill, 2020).

We planned an experiment where a forensic scientist disposes of a sample of $n = 1000$ observations from the reference population and has to deal with a rare type match case, consisting of a crime's and a suspect's identical profiles that are not present among the 1000. In a change of glass framework, the specific feature of these observations is irrelevant: what matters is that their common characteristic is not included in the database that is available.

The aim of the experiment is to investigate the differences between evaluating the LR according to (8) or to (9), and also to evaluate to what extent the sampling variability affects results. To do so, the procedure described above has been replicated 50 times. In addition, to make the analyses more comparable, we used EB priors described in Section 5.2.

Here, only the MCMC procedure has been employed because similar inference was obtained by ABC and MCMC methods (see Section 5).

Figure 3 represents the contour plot of the parameters' posterior density obtained by using MCMC simulations (black levels); the α_{MLE} and θ_{MLE} (red dot); the value of $\phi(\alpha, \theta)^{-1}$ for pairs of α and θ (blue lines). It appears that $\phi(\alpha, \theta)^{-1}$ is much more sensitive to α than to θ, and that there is an asymmetry toward small values of α. In this case, the posterior distribution is far from being symmetric and centered around $(\alpha_{\mathrm{MLE}}, \theta_{\mathrm{MLE}})$ and has a long tail on pairs of α and θ, corresponding to high values of $\phi(\alpha, \theta)^{-1}$. Accordingly, the value of the $\mathrm{LR}_{\mathrm{MC}}$ is all the more reason to be expected

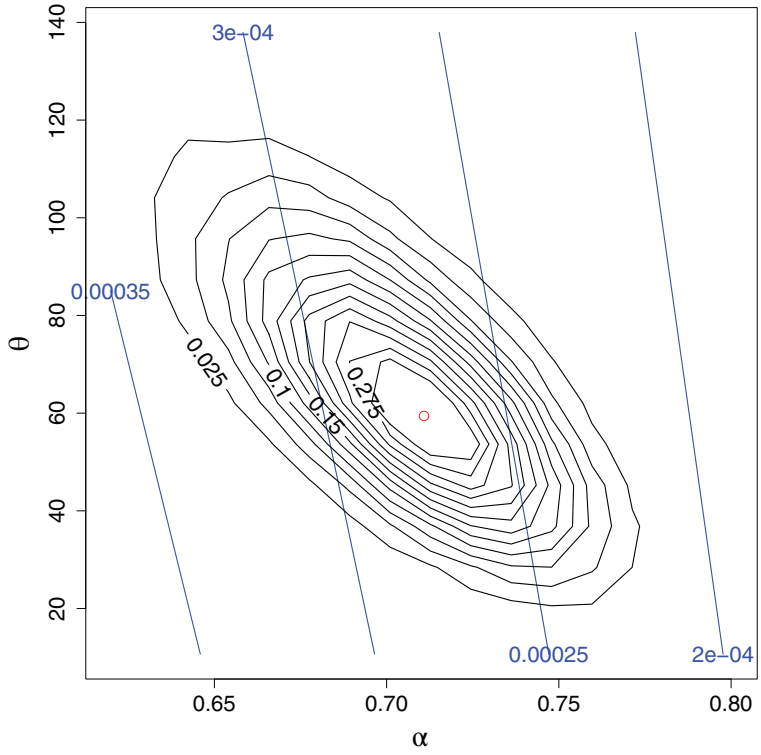

FIGURE 3 Density of the posterior distribution of $\alpha, \theta | \pi_{[n+1]}$ obtained with MCMC (black lines), and values of $\phi(\alpha, \theta)^{-1}$ (blue lines). The MLE is represented with a red point [Colour figure can be viewed at wileyonlinelibrary.com]

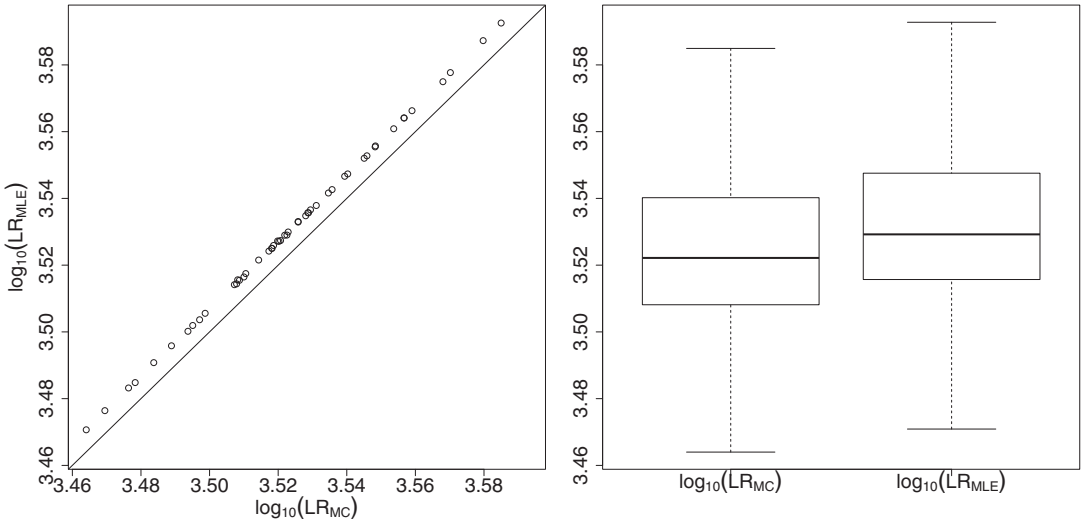

FIGURE 4 Comparison (scatterplot and boxplots) of the ($\log_{10}$) LR values obtained by using (8) or (9) over 50 experiments

to be smaller than LR_{MLE}. This explanation holds for all experiments because the contour plots, represented in Figure 3 for one experiment only, are very similar for all experiments.

As expected and discussed above, Figure 4 shows that the LR_{MC} is always smaller than the corresponding LR_{MLE}. It also shows little variability among the LRs that are computed by using 50 different samples from the population. This is due to the reduction of the data in partitions because many different databases may lead us to observe the same partition. Indeed, labels are lost and the forensic statistician does not take into account the specific characteristic observed on the crime scene, but only takes into account the equality between the crime and the suspect profiles, and the novelty that they represent with respect to the database.

7 | DISCUSSION

Since Brenner (2010) defined the problem of assigning a probative value to the rare type match as the fundamental problem of forensic mathematics, some efforts have been made to provide a solution to the problem (Cereda, 2017a; Cereda, 2017b; Dorp et al., 2020).

Recently, Cereda and Gill (2020) proposed a Bayesian nonparametric model, providing an approximate solution based on modeling the ordered relative frequencies of a highly variable characteristic through a PD prior distribution. In our opinion, switching attention from the specific characteristic to the event of observing a match of a not yet observed type is the correct way to address the rare type match problem. Moreover, we demonstrate that the hybrid approximation introduces a nonconservative bias, which is quite undesirable in the forensic context. To solve this problem, we suggest two different inferential strategies to obtain the posterior distribution of the two parameters: using the MCMC and using the ABC approach. As a by-product, this activity also results in a study of the effectiveness of ABC to derive the parameters' posterior distribution in a Bayesian nonparametric setting where, as in the present case, it is possible to implement both ABC and MCMC.

In Section 5.1 we verified that the two procedures provide a comparable inference so that they can be both employed. As concerns efficiency, ABC naturally supports parallel computing on

multi-core laptops, thus reducing the computational time. Finally, comparing Algorithms 1 and 2 and considering the machinery of convergence diagnostic for the MCMC, it appears that the implementation of ABC is more straightforward than MCMC. The adoption of the rejection ABC allows for a high degree of multi-threading. However, the same is not valid for schemes allowing for parameters' proposal following a Markov chain, such as in Marjoram et al. (2003), or when the tolerance parameters are adaptively determined, such as in Beaumont et al. (2009), and Del Moral et al. (2012).

Our forensic application considered Y-STR profiles, for which other authors proposed to evaluate the support to the identification hypothesis through the inheritance of the genetic characteristics among generations (Andersen et al., 2013; Andersen & Balding, 2017). A reduction of data similar to the one used in our application was carried out in Cereda (2017b) to solve the rare type match problem using a generalization of the Good Turing estimator.

Other kinds of evidence, such as fragments of glasses, measured according to the refractive index, the chemical composition of the glass, its thermal history, and any surface characteristics, quickly lead to a large number of different profiles and a rare match case may arise (e.g., see Vergeer et al. (2020)).

The same is true for evidence such as fibers, which are another quite ordinary piece of evidence. This circumstance suggests that our method should be successfully used in applications using PD distribution in different forensic frameworks, such as speaker recognition (Silnova et al., 2020), DNA analysis (Fernando, 2017) or in other fields, such as finance (Sosnovskiy, 2015).

ORCID

Giulia Cereda https://orcid.org/0000-0002-2913-6206
Fabio Corradi https://orcid.org/0000-0003-3949-3837
Cecilia Viscardi https://orcid.org/0000-0002-2791-7025

REFERENCES

Aldous, D. J. (1985). *Exchangeability and related topics École D'Été de Probabilités de Saint-Flour* (Vol. *1117*). Springer-Verlag.

Andersen, M. M., & Balding, D. J. (2017). How convincing is a matching Y-chromosome profile? *PLOS Genetics*, *13*, 1–16.

Andersen, M. M., Eriksen, P. S., & Morling, N. (2013). The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, *329*, 39–51.

Barber, S., Voss, J., & Webster, M. (2015). The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, *9*, 80–105.

Beaumont, M., Cornuet, J. M., Marin, J. M., & Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, *96*, 983–990.

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

Beaumont, M. A. (2019). Approximate Bayesian computation. *Annual Review of Statistics and Its Application*, *6*, 379–403.

Brenner, C. H. (2010). Fundamental problem of forensic mathematics—The evidential value of a rare haplotype. *Forensic Science International: Genetics*, *4*, 281–291.

Broderick, T., Jordan, M. I., & Pitman, J. (2012). Beta processes, stick-breaking and power laws. *Bayesian Analysis*, *7*, 439–476.

Caliebe, A., Jochens, A., Willuweit, S., Roewer, L., & Krawczak, M. (2015). No shortcut solutions to the problem of Y-STR match probability calculation. *Forensic Science International: Genetics*, *15*, 69–75.

Carlton, M. A. (1999). *Applications of the two-parameter Poisson-Dirichlet distribution* [Ph.D. thesis]. University of California, Los Angeles.

Carmona, C., Nieto-Barajas, L., & Canale, A. (2018). Model-based approach for household clustering with mixed scale variables. *Advances in Data Analysis and Classification*, *13*, 559–583.

Cereda, G. (2017a). Bayesian approach to LR in case of rare type match. *Statistica Neerlandica*, *71*, 141–164.

Cereda, G. (2017b). Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach). *Scandinavian Journal of Statistics*, *44*, 230–248.

Cereda, G., & Gill, R. D. (2020). A nonparametric Bayesian approach to the rare type match problem. *Entropy*, *22*, 439.

Dawid, A. P. (2017). Forensic likelihood ratio: Statistical problems and pitfalls. *Science & Justice*, *57*, 73–75.

Dawid, A. P., & Mortera, J. (1996). Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 425–443.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Pruenster, I., & Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, *37*, 212–229.

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

Dennis, J., & Schnabel, R. (1996). *Numerical methods for unconstrained optimization and nonlinear equations Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics.

Dorp, I., Leegwater, A. J., Alberink, I., & Jongbloed, G. (2020). Value of evidence in the rare type match problem: Common source versus specific source. *Law, Probability and Risk*, *19*, 85–98.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, *89*, 268–277.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*, 577–588.

Escobar, M. D., & West, M. (1998). Computing nonparametric hierarchical models. *Practical Nonparametric and Semiparametric Bayesian Statistics*, *133*, 1–22.

Favaro, S., Lijoi, A., Mena, R. H., & Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Methodological)*, *71*, 993–1008.

Fearnhead, P., & 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*, 419–474.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparamteric problems. *Annals of Statistics*, *1*, 209–230.

Fernando, M. (2017). *Bayesian models for PCR stutter* [Ph.D. thesis]. The University of Auckland.

Goldwater, S., Johnson, M., & Griffiths, T. (2006). *Interpolating between types and tokens by estimating power-law generators*. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems* (Vol. *18*). MIT Press.

Hoff, P. (2009). *A first course in Bayesian statistical methods Springer Texts in Statistics*. Springer.

Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, *17*, 499–520.

Jara, A., Lesaffre, E., De Iorio, M., & Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics. 4*, 2126–2149.

Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *37*, 1–22.

Lijoi, A., Mena, R. H., & Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, *94*, 769–786.

Lijoi, A., Mena, R. H., & Prünster, I. (2008). A Bayesian nonparametric approach for comparing clustering in EST analysis. *Journal of Computational Biology*, *10*, 1315–1327.

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

Marjoram, P., Molitor, J., Plagnol, V., & Tavare, S. (2003). Markov Chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 324–328.

Meester, R., & Slooten, K. (2021). *Probability and forensic evidence: Theory, philosophy, and applications*. Cambridge University Press.

Murphy, K., Viroli, C., & Gormley, C. (2019). Infinite mixtures of infinite factor analysers. *Bayesian Analysis*, *15* 937–963.

Navarrete, C., Quintana, F., & Muller, P. (2008). Some issues on nonparametric Bayesian modeling using species sampling models. *Statistical Modelling*, *8*, 3–21.

Pardo, L. (2018). *Statistical inference based on divergence measures Statistics: A Series of Textbooks and Monographs*. CRC Press.

Pitman, J. (1992). *The two-parameter generalization of Ewens' random partition structure* (Technical Report No. 345). Department of Statistics U.C, Berkeley, CA.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, *102*, 145–158.

Pitman, J. (1996). *Some developments of the Blackwell-MacQueen urn scheme Lecture Notes-Monograph Series* (Vol. *30*, pp. 245–267). Institute of Mathematical Statistics.

Pitman, J. (2002). *Combinatorial stochastic processes Lecture notes for St. Flour Summer School*. Department of Statistics, University of California at Berkeley.

Pitman, J. (2006). *Combinatorial stochastic processes École D'Été de Probabilités de Saint-Flour XXXII - 2002*. Springer.

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S. M. T., Santos, L. H., Anslinger, K., Bayer, B., Ayub, Q., Wei, W., Xue, Y., Tyler-Smith, C., Bafalluy, M. B., Martínez-Jarreta, B., Egyed, B., Balitzki, B., Tschumi, S., … Roewer, L. (2014). A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Science International: Genetics*, *12*, 12–23.

Raftery, A. E., & Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science*, *7*, 493–497.

Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Sibuya, M., & Yamato, H. (2001). Pitman's model of random partitions (Technical Report). RIMS Kokyuroku. Research Institute for Mathematical Science, Kyoto University. https://www.kurims.kyoto-u.ac.jp/~kyodo/kokyuroku/contents/pdf/1240-7.pdf

Silnova, A., Brummer, N., Rohdin, J., Stafylakis, T., & Burget, L. (2020). *Probabilistic embeddings for speaker diarization*. Proceedings of the Odyssey 2020 the Speaker and Language Recognition Workshop.

Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. Chapman & Hall/CRC Press.

Sosnovskiy, S. (2015). On financial applications of the two-parameter Poisson Dirichlet distribution. *arXiv:1501.01954*.

Taroni, F., Bozza, S., Biedermann, A., & Aitken, C. (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, *15*, 1–16.

Vergeer, P., Leegwater, A. J., & Slooten, K. (2020). Evaluation of glass evidence at activity level: A new distribution for the background population. *Forensic Science International*, *316*, 110431.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models (Technical Report). Institute of Statistics and Decision Sciences Duke University. https://www2.stat.duke.edu/~mw/MWextrapubs/West1992alphaDP.pdf

Zhou, X., Huang, J., & Wu, X. (2017). Estimation of Poisson–Dirichlet parameters with monotone missing data. *Mathematical Problems in Engineering*, *2017*, 7892507.

How to cite this article: Cereda, G., Corradi, F., & Viscardi, C. (2022). Learning the two parameters of the Poisson–Dirichlet distribution with a forensic application. *Scandinavian Journal of Statistics*, 1–21. https://doi.org/10.1111/sjos.12575