A change of glasses strategy to solve the rare type match problem

Un cambio di prospettiva per risolvere il problema del rare type match

Giulia Cereda and Fabio Corradi

Abstract We propose a solution to a forensic statistics problem known as the "rare type match case". It happens when the characteristics of the crime and the suspect's traces match but they have not been observed yet in previously collected databases. The proposed solution relies on a "change-of-glasses" strategy and consists of ignoring the specific evidence characteristics thus only modeling equalities and inequalities among different types. For the rare type match case this reduces to consider the event of seeing twice a never observed type, along with a database, now coded in form of a partition, losing reference to the specific characteristics observed. We propose to use a Bayesian nonparametric approach and derive the likelihood ratio required for forensic assessment. MCMC inference is carried on and compared to MLE through a toy example.

Abstract *Si propone una soluzione al problema forense "match di tipo raro", che si verifica quando la traccia trovata sulla scena del crimine e quella di un sospetto corrispondono ma non sono mai state osservate in precedenza. La strategia proposta è un "cambio di occhiali" in cui si ignora la specifica caratteristica delle tracce osservate, modellizzando solamente uguaglianze e disuguaglianze tra le diverse caratteristiche. Per il match di tipo raro, questo significa considerare l'evento di osservare due volte una nuova caratteristica, senza riferimento a quale essa sia, insieme a un database codificato sotto forma di partizione. Proponiamo per questo problema un approccio Bayesiano non parametrico, derivando il rapporto di verosimiglianza richiesto dal protocollo forense insieme a un'inferenza MCMC e stimatori di massima verosimiglianza per i parametri.*

Key words: Rare type match problem, Forensic Statistics, Bayesian nonparametrics, two-parameter Poisson Dirichlet, MCMC methods.

Introduction

On the crime scene, a trace has been retrieved showing characteristics that match the suspect's characteristics. The forensic statistician, who is given this piece of evidence along with a database of reference containing n traces, is asked to assess the likelihood ratio (LR), in order to weight the data D under the prosecution's (identification) and the de-

Giulia Cereda
Mathematical Institute, Leiden University, e-mail: giulia.cereda7@gmail.com

Fabio Corradi
Dipartimento di Statistica, Informatica, Applicazioni Firenze, e-mail: fabio.corradi@unifi.it

fence's (no-identification) hypotheses, h_p and h_d:

$$\mathrm{LR} = \frac{\Pr(D \mid H = h_p)}{\Pr(D \mid H = h_d)}. \tag{1}$$

In the rare type match case no other traces with the same characteristics are among those contained in the database of reference and the "rarity principle" is not operational since there are no frequencies to evaluate the rarity.

We propose a "change-of-glasses strategy", which considers the event that a never observed characteristic has been observed twice regardless of its specific type. Overall, data D are made of $n+2$ observations (n from a database and two traces from the suspect and crime scene). We are thus reducing D to a partition of the set $[n+2] = \{1,...,n+2\}$, by assigning to each class the indexes of the observations with equal characteristics.

Focusing on partitions allows us to use a nonparametric Bayesian approach. More specifically, as proposed in [2], we make use of the two-parameter Poisson Dirichlet prior to model the relative ranked sizes of the classes of the partition. This choice is motivated by the power-law shape often encountered using forensic data, such as Y-STR profiles.

1 Random partitions: an example

Consider a sequence of integer-valued random variables $I_1,...,I_n$, representing units with some characteristics expressed by the value of the Is and the equivalence relation $i \sim j$ if and only if $I_i = I_j$. The equivalence classes formed by subsets of indices with identical I form a random partition of $[n]$, which will be denoted as $\Pi_{[n]}(I_1, I_2, ..., I_n)$. For instance, the following random partition

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

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. What we have retained is the composition of the classes themselves, but we have lost the information about the characteristics of each I_i. Assume partition $\pi_{[10]}$ as representing a database of 10 individuals. The rare type match case partition is obtained by augmenting $\pi_{[10]}$ by:

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

The 11-th and the 12-th observed traces constitute a new class by themselves since they are equal one another but different from those previously observed. The strategy is thus to focalise on the classes of the partition, taking only account of similarities and dissimilarities among traces.

2 The two-parameter Poisson-Dirichlet distribution

The two-parameter Poisson Dirichlet distribution [8], is a distribution over the infinite simplex of the form $\nabla_\infty = \{(p_1, p_2, ...) \mid p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$. It has two parameters, $\alpha \in [0,1)$, and $\theta > -\alpha$ and operatively can be constructed in two steps by sorting the well-known GEM(α, θ) [5] also known as 'stick breaking prior'. One of the interesting feature of PD is its ability to represent different power-law distributions. By assuming that there is an infinite number of different characteristics, that their ordered frequencies follow a two-parameter PD distribution, and that the database is an i.i.d. sample from $\mathbf{p}$,

$$\mathbf{P} \mid \alpha, \theta \sim \mathrm{PD}(\alpha, \theta), \quad I_1, \ldots, I_n \mid \mathbf{P} = \mathbf{p} \sim_{\text{i.i.d}} \mathbf{p} \tag{3}$$

then, for all $n \in \mathbb{N}$, the random partition $\Pi_{[n]} = \Pi_{[n]}(I_1, ..., I_n)$ has the following distribution:

$$\Pr(\Pi_{[n]} = \pi_{[n]} | \alpha, \theta) = \frac{[\theta + \alpha]_{k-1;\alpha}}{[\theta + 1]_{n-1;1}} \prod_{i=1}^{k} [1 - \alpha]_{n_i - 1;1}, \tag{4}$$

where $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$, $[x]_{a;b} := \begin{cases} \prod_{i=0}^{a-1}(x + ib) & \text{if } a \in \mathbb{N}\backslash\{0\} \\ 1 & \text{if } a = 0 \end{cases}$, and n_i is the size of the ith class of $\pi_{[n]}$.

Relation (4), known as the *Pitman sampling formula* [6], will be used as likelihood for deriving inference for α and θ by using an MCMC scheme.

There is an alternative characterization of this model, called "Chinese restaurant process", studied in detail in [7]. It is defined as follows: consider a restaurant with infinitely many tables, each one infinitely large. Let $S_1, S_2, ...$ be integer-valued random variables representing the seating plan of the restaurant: tables are ranked in order of first occupancy: $S_i = j$ means that the ith customer seats at the jth table. The process is described by the following conditional probabilities:

$$S_1 = 1, \qquad \Pr(S_{n+1} = j | S_1, ..., S_n) = \begin{cases} \dfrac{\theta + k\alpha}{n + \theta} & \text{if } j = k + 1 \\ \dfrac{n_j - \alpha}{n + \theta} & \text{if } 1 \leq j \leq k \end{cases} \tag{5}$$

where k is the number of tables occupied by the first n customers, and n_j is the number of customers already occupying table j. The process depends on two parameters α and θ constrained as the PD parameters. Clearly $S_1, ..., S_n$ are not i.i.d., nor exchangeable but in [7] it is shown that $\Pi_{[n]}(S_1, ..., S_n)$ is distributed as $\Pi_{[n]}(I_1, ..., I_n)$, with $I_1, ..., I_n$ defined by (3) and they are distributed according to the Pitman sampling formula (4).

3 Likelihood ratio

Reducing the data to partitions and assuming that the two-parameter Poisson Dirichlet distribution models the ordered frequencies of the (infinite) characteristics, the LR in (1) can be defined and derived as follows:

$$\begin{aligned}
\mathrm{LR} &= \frac{p(\pi_{[n+2]} \mid h_p)}{p(\pi_{[n+2]} \mid h_d)} = \frac{p(\pi_{[n+2]}, \pi_{[n+1]} \mid h_p)}{p(\pi_{[n+2]}, \pi_{[n+1]} \mid h_d)} && \text{since } \pi_{[n+1]} \subset \pi_{[n+2]} \\
&= \frac{p(\pi_{[n+2]} | \pi_{[n+1]}, h_p) p(\pi_{[n+1]} \mid h_p)}{p(\pi_{[n+2]} | \pi_{[n+1]}, h_d) p(\pi_{[n+1]} \mid h_d)} \\
&= \frac{1}{p(\pi_{[n+2]} \mid \pi_{[n+1]}, h_d)} && \text{since } \Pi_{[n+1]} \perp\!\!\!\perp H \text{ and } p(\pi_{[n+2]} | \pi_{[n+1]}, h_p) = 1 \\
&= \frac{1}{\int p(\pi_{[n+2]} | \pi_{[n+1]}, \alpha, \theta, h_d) p(\alpha, \theta | \pi_{[n+1]}, h_d) d\theta d\alpha} \\
&= \frac{1}{\int \frac{1-\alpha}{n+1+\theta}\ p(\alpha, \theta \mid \pi_{[n+1]}, h_d) d\alpha d\theta} && \text{by (5)}
\end{aligned} \tag{6}$$

Note that under h_p the $n+2$-th and the $n+1$th characteristics are equal with probability 1. Result (6) is formally reminiscent of the likelihood ratio employed in usual forensic identification, whenever the same characteristic is observed from the crime's and the suspect's trace. There, the crucial quantity – to be integrated with respect to the posterior distribution of unknown parameters – is the probability of observing the suspect's evidence given the database enlarged with the crime trace, under the defense hypothesis [3]. Changing glasses, we now integrate, with respect to the posterior of the Poisson Dirichlet parameters, the event of observing the $n+2$-th profile, identical to the $n+1$th, both never observed before. Using a prior PD, the probability of this event, conditionally to the model parameters is provided by (5) (bottom line with $n_i = 1$) and is equal to $\frac{1-\alpha}{n+1+\theta}$. Also, from (6), it is apparent the crucial role played by the posterior of α and θ, obtained by conditioning on $\pi_{[n+1]}$. This motivates our interest in studying inference on the PD parameters.

4 MCMC inference

For a budget of T simulations, Algorithm 1 summarizes the implementation of the Metropolis-Hastings inference for the parameters (α, θ) of the Poisson Dirichlet distribution, conditionally to an observed random partition $\pi_{[n]}$.

Algorithm 1 MH

Initialize $\theta_0 \sim p(\theta)$, $\alpha_0 \sim p(\alpha)$
for $t = 1, ..., T$ **do**
  Propose a θ_{t+1}, α_{t+1} from $p(\theta_{t+1}|\theta_{t^*})p(\alpha_{t+1}|\alpha_{t*})$
  Evaluate the ratio $R = \frac{\ell(\theta_{t+1},\alpha_{t+1};\pi_{[n]})}{\ell(\theta_{t^*},\alpha_{t^*};\pi_{[n]})} \frac{p(\theta_{t+1})}{p(\theta_{t^*})} \frac{p(\alpha_{t+1})}{p(\alpha_{t^*})} \frac{p(\theta_{t^*}|\theta_{t+1})}{p(\theta_{t+1}|\theta_{t^*})} \frac{p(\alpha_{t^*}|\alpha_{t+1})}{p(\alpha_{t+1}|\alpha_{t^*})}$
  Accept θ_{t+1} with probability R
end for

In particular, we propose as prior for θ: $\theta \sim U(0, \theta_{max})$ and $p(\theta_{max}|\theta_0, \tau) = \tau\theta_0^{\tau}(\theta^{-\tau-1})$, a Pareto$(\theta_0, \tau)$. This requires to express a prior opinion on θ_0, the smallest value that θ_{max} can assume. As prior for α: $\alpha \sim \text{Unif}(0,1)$.

The proposal distribution for θ is $\theta_{t+1} \mid \theta_{t^*} \sim \text{Exp}(\frac{1}{\theta_{t^*}})$, so the mean of the proposal distribution is equal to the last accepted θ_{t^*}. The proposal for α is a reflecting random walk to take into account that $\alpha \in [0,1]$.

5 A simulated example

To derive inference for the two parameters we explore two methods:

1. MLE estimators $\widehat{\alpha}_{\text{MLE}}$ and $\widehat{\theta}_{\text{MLE}}$ obtained by using the Pitman's sampling formula (4).
2. the Metropolis Hasting method described in Section 4 that provides a sample from the posterior of α and θ given an observed partition of size $\pi_{[n]}$.

The likelihood ratio of formula (6) can be obtained by plugging-in the MLE estimates for method 1, or by Montecarlo approximation using the MCMC sample, for method 2.

In order to compare the two approaches, we apply them to observed partitions obtained from two distinct populations that are distributed according to the two-parameter Poisson Dirichlet with known parameters. This allows us to concentrate on the quality of the

inference and intentionally avoid the influence of possible model mi-specification. More specifically, we create two distinct populations of size $N = 10^6$ using the Chinese Restaurant process (5), then we draw a sample of size $n = 1000$ for each population, and obtain the corresponding partition $\pi_{[1000]}$. The true parameters of each population and the MLE estimators are shown in Table 1.

	True values		MLE		MCMC specifications				Effective sample size	
	α	θ	α_{MLE}	θ_{MLE}	n. iterations	thinning	burn-in	accepted	α	θ
Example 1	0.5	20	0.48	19.37	10^6	50	75'000	18'500	44'893.88	54'257.52
Example 2	0.2	2	7×10^{-7}	3.57	10^6	80	100'000	11'250	32'075.81	56'373.49

Table 1 Values of α and θ used to simulate two populations of size $N = 10^6$ along with the MLE for α and θ, obtained from a i.i.d sample of size $n = 1000$ reduced in form of partition. The specifications of the Metropolis Hasting algorithm are also displayed, along with the effective sample size.

To stress the inferential procedures, the partition sample of the first example (from Population 1) is very close to the average partition obtained by repeatedly applying the simulation process. On the opposite, the partition sample of the second example (from Population 2) is more "pathological", since it is quite different from the average partition. In particular, it is outstanding with regard to the number of singletons observations (only two) and of duplets (four). This heterogeneity allows us to have an insight into the robustness of the two methods over "extreme" observations from the population. As expected, in the second example the MLE provides a weak inference. The Metropolis Hasting algorithm is used with $S = 10^6$ iterations: burn-in and thinning interval are assessed by using the diagnostic tests of the R package `coda` (see Table 1).

Figure 1 shows the MLE estimates and true values along with the posterior distributions of $\alpha, \theta \mid \pi_{[n+1]}$ obtained with MH with 95% credible intervals. In the first example, both methods provide good inference for α and θ, while in the second case, with the pathological partition, MLE is practically useless for α and bad for θ. The MCMC approach, even though not optimal, represents an improvement, since at least the true values for α and θ are reached by the credible intervals. The same considerations can be made regarding LR values (last column of Figure 1).

6 Conclusions

At first glance, the rare type match problem appears as an odd issue. Actually the "not yet observed" condition is very common also with the most widespread forensic identification evidence, the autosomal DNA-STR profiles. Indeed, if considered as whole profiles, they are often unique and only resorting to some independence assumptions it is possible, to consider each locus separately. In other circumstances, if such forms of independence do not hold, as it happens for the Y-STR profiles or for non-genetic characteristics, the rare type match problem remains a common and challenging issue. Our proposal is very general and only relies on a few conditions, such as the existence of a high number of different modalities for the considered characteristic and their power-law behavior. For the Y-STR rare type match problem, [1] proposed a different solution that does not reduce data to a partition but makes use of some genetic assumptions and the knowledge of some population parameters (such as mutation rates, and IBD parameters). In the future, it will be interesting to compare their approach with ours. Other areas of application concern important qualitative evidence used in forensic science for identification, such as glass

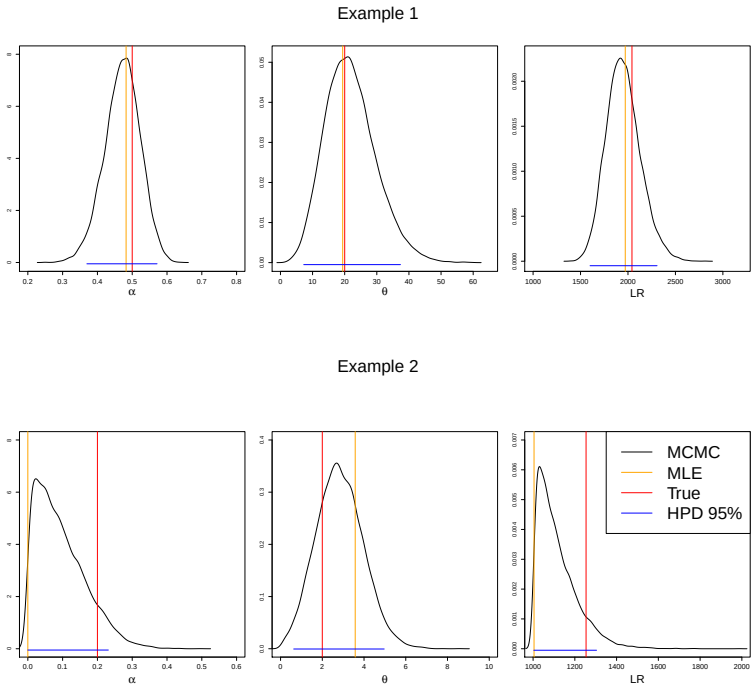

Fig. 1 Comparison of the inference provided by MCMC simulation and MLE. The first column corresponds to inference for α, the second column for θ, and the third column to the LR values obtained by plugging in MLE estimate or by Montecarlo approximation of the integral. The vertical lines represent the true value (red) and the MLE estimates (orange). The credible intervals with probability 95% of the distribution obtained through MH are also displayed through horizontal blue segments.

fragments and tire marks. Many efforts have been devoted to solve the rare type match problem in these areas, see [4] and we hope our contribution would be helpful.

References

1. M. M. Andersen and D. J. Balding. How convincing is a matching y-chromosome profile? *PLOS Genetics*, 13:1–16, 2017.
2. G. Cereda and R. D. Gill. A nonparametric Bayesian approach to the rare type match problem. *Entropy* 22(4): 439, 2020.
3. G. Cereda. Bayesian approach to LR in case of rare type match. *Statistica Neerlandica*, 71:141–164, 2017.
4. J. M. Curran, T.N. Hicks, and J.S. Buckleton, *Forensic Interpretation of Glass Evidence*. CRC Press, Boca Raton, Florida, 2000.
5. R. C. Griffiths. Exact sampling distributions from the infinite neutralalleles model. *Advances in Applied Probability*, 11:326–354, 1969.
6. J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995.
7. J. Pitman. *Combinatorial Stochastic Processes*. École D'Été de Probabilités de Saint-Flour XXXII - 2002. Springer, Berlin, 2006.
8. J. Pitman. The two-parameter generalization of Ewens' random partition structure. Technical report, Department of Statistics U.C. Berkeley CA, 2003.