Revista Colombiana de Estadística - Applied Statistics
January 2024, volume 47, issue 1, pp. 135–150
http://doi.org/10.15446/rce.v47n1.109069

Addressing Misidentification in Noninvasive DNA Sampling Using Bayesian Approach and Simulations

La identificación errónea en el muestreo de ADN no invasivo abordada mediante modelos bayesianos y simulaciones

Paula Bran[1,a], Leon Escobar[1,b]

[1]Mathematics Department, Universidad del Valle, Cali, Colombia

Abstract

Noninvasive DNA sampling has become increasingly popular in wildlife research and conservation because it allows scientists to gather valuable genetic information without disturbing or harming the animals. However, the correct identification of the species or individuals in the sample is virtually impossible when using this kind of sampling. Consequently, it becomes essential to consider the errors hidding true identities in order to improve the quality of the data. Errors, if left unaddressed, can have a considerable impact on the accuracy of statistical inferences drawn from the data. This paper endeavours to review some research about misidentification problems and how Bayesian models and Markov Chain Monte Carlo (MCMC) methods can be applied. In addition, a hypothetical scenario is presented to illustrate how genetic material can serve as unique identifier of individuals, and to highlight the potential difficulties that may arise if a proposal distribution for the MCMC simulations is inappropriately chosen.

***Key words*:** Noninvasive DNA sampling; Misidentification; Latent individual; MCMC; Reversibility.

Resumen

El muestreo de ADN no invasivo se ha vuelto cada vez más popular en la investigación y conservación de vida silvestre, ya que permite a los científicos recopilar información genética valiosa sin perturbar ni lesionar a los animales. Sin embargo, la correcta identificación de la especie o individuos en la muestra es prácticamente imposible cuando se utiliza este tipo de muestreo. En consecuencia, es fundamental considerar los errores que

[a] Profesora Asociada. E-mail: paula.bran@correounivalle.edu.co
[b] Profesor Asistente. E-mail: leon.escobar@correounivalle.edu.co

ocultan las verdaderas identidades con el fin de mejorar la calidad de los datos. Si los errores no se abordan, pueden tener un impacto considerable en la precisión de las inferencias estadísticas obtenidas a partir de los datos. Este artículo se propone revisar algunas investigaciones sobre problemas de identificación errónea y cómo se pueden aplicar los modelos bayesianos y los métodos de Monte Carlo basados en cadenas de Markov (MCMC). Además, se presenta un escenario hipotético para ilustrar cómo el material genético puede servir como identificador único de los individuos, y resaltar las dificultades potenciales que pueden surgir si se elige inapropiadamente una distribución de propuestas para las simulaciones de MCMC.

***Palabras clave*:** Muestreo de ADN no invasivo; Identificación errónea; Individuo latente; Métodos MCMC; Reversibilidad.

1. Introduction

In research involving data, there is always a risk that data may be compromised by several sources of corruption, causing errors that may be both subtle and unavoidable. Common issues include duplicated or incorrectly reported observations, as well as missing data. For instance, when conducting laboratory experiments, data can be affected by environmental conditions, equipment malfunctions, and other factors beyond the researcher's control. Surveys can collect also contaminated observations, for example, responses may contain false information of the participants due to unintentional misspelling, respondent mistakes, either intentional or unintentional, lack of interviewer impartiality, and inconsistencies with the questionnaire design, as in Berg & Lien (2009).

Noninvasive techniques are employed in wildlife research to minimize disturbance to animals both physically and psychologically. It can be carry out by means of sensing technologies, noncontact monitoring, and passive observation. One specific noninvasive technique is DNA sampling, which involves collecting genetic material from organisms without capturing or handling them. This novel sampling method is limited when it comes to accurately identifying individuals, especially when compared to traditional invasive methods.

In order to estimate the population size accurately, the observed (corrupted) data needs to be matched with the latent (true) data. This matching process helps to determine the actual sample size, which is then used for population size estimation. The majority of existing methods for estimating population size, while accounting for uncertainty, treat matching and size estimation as distinct and independent steps. See for example the work of Yoshizaki et al. (2011), Wright et al. (2009) and Lukacs & Burnham (2005*a*), where the issue of genotype misidentification is incorporated into multiple mark-recapture models for estimating animal abundance using DNA samples.

This paper aims to examine existing research on misidentification issues and explore how Bayesian models and Markov Chain Monte Carlo (MCMC) methods can help to address these challenges. It is important to emphasize that this article does not introduce a new method but rather discusses an existing approach. The aim is draw attention in the topic to highlight potential limitations and areas for improvement in the current method. By demonstrating through a specific toy example that not only illustrates how genetic material can be utilized as distinctive identifier for individuals, but also serves to underscore the importance of revisiting certain aspects of the approach to ensure accurate simulation of the posterior distribution. More precisely, the toy example to emphasizes the importance of selecting an appropriate proposal distribution for MCMC simulations to mitigate potential difficulties with no convergence of the chain.

This section will extensively explore the concept of noninvasive DNA sampling, its advantages, challenges, and applications. Section 2 will explain a specific genotyping error considered in Wright et al. (2009); Section 3 will present the data and model; Section 4 will present the hypothetical scenario with a more general genotyping error and the MCMC algorithm for the simulations.

1.1. Noninvasive sampling and the sample size

In particular, wildlife research can adversely affect the organisms under investigation, even when conducted with strict protocols. An illustrative case is the study conducted by Ditmer et al. (2015), where the presence of unmanned aerial vehicles (commonly known as drones) in bear habitats caused stress to the bears. The authors observed significant changes in bear behaviour when drones were present, noting elevated heart rates, even during hibernation. While drones facilitate access to natural environments for data collection, they can also cause distress and disruption to the species being studied. This example effectively highlights the potential stress experienced by certain species during research monitoring and data collection.

Noninvasive techniques utilize nonintrusive methods that do not physically or psychologically disturb the wildlife. This can involve remote sensing technologies, noncontact monitoring, or passive observation methods. In particular, noninvasive DNA sampling refers to the collection of genetic material from organisms without physically capturing or handling them. It involves extracting DNA from nonintrusive sources such as shed hair, feathers, skin cells, saliva, feces, body fluids (such as sweat, urine) or other biological samples left behind by the organism in its natural environment. In population ecology, they are often used to study elusive animals. For example, fearful wolves, shy birds, nocturnal animals or camouflaged reptiles which are virtually undetectable by eyesight.

Mark-recapture methods are often used to estimate animal abundance, which is a common problem in wildlife management. Otis et al. (1978) described mark-recapture modelling for populations that are demographically closed, that is, no individuals enter or leave the population during the study. These models assume that marks are preserved during the experiment, meaning that they do not fall off or change in a way that they could be misread, and all marks are accurately observed and registered at each trapping occasion.

Ecologists have taken advantage of the latest advances in molecular biology to obtain individual genotypes from noninvasive samples. The genotyped profiles of the individuals may then be used as marks because, in large populations, it is unlikely that two individuals will have the same genetic profile. The fact these samples are taken unobtrusively affects the reliability of the assignments of the genotypes to the individuals. The genotyped individuals may be subject to a high degree of uncertainty because the quality of the genetic information may be negatively affected by environmental factors or during DNA amplification. Because the use of noninvasive DNA data may be prone to errors, the models in Otis et al. (1978) cannot be applied as the assumption that the marks are read and recorded correctly is inadequate.

Gathering genetic data from faeces helps identify the species present in a particular area. This is especially helpful when studying elusive or dangerous species that are challenging to monitor directly. For example, Wright et al. (2009) developed a Bayesian model for estimating the population size of a nocturnal animal using faeces samples. Mondol et al. (2009) demonstrated the effectiveness of noninvasive genetic sampling methods, specifically scat collection, in estimating tiger population size. Roques et al. (2014) utilized genetic techniques to estimate population size and assess the genetic diversity and structure of jaguars in the study areas in Brazil. The results in Marucco et al. (2012) highlighted the value of long-term monitoring using noninvasive sampling, as it provided valuable information about the dynamics and persistence of wolf populations over time. By combining genetic data with spatial capture-recapture techniques, Morin et al. (2016) gained insights into coyote movements, territoriality, and population trends. Finally, the study conducted by Biffi & Williams (2017) focused on the use of noninvasive techniques to determine the population size of the marine otter in two regions of Peru. All of these used faecal DNA as natural marks for studying cryptic animals avoiding direct contact while minimizing risks to both humans and animals.

There are some difficulties inherent in the mark-recapture approach based on DNA samples. Lukacs & Burnham (2005*b*) express two concerns in these studies. First, the notion of a sampling occasion is unclear. Second, it may be virtually impossible to set out a list of marks in the population. Naturally, there is concern about these difficulties because sampling occasions and marks are dominant notions in mark-recapture studies. Both issues will be discussed separately.

First, a sampling occasion refers to the time that samples are collected from the population. This concept in a conventional mark-recapture study is considered 'as a short, discrete event' as stated by Lukacs & Burnham (2005*b*). However, in a mark-recapture study based on noninvasive DNA samples, it is a vague notion. Evidence of this is the fact that the animal shed DNA into the sample at an unknown time. Barker et al. (2014) described a general model for capture-recapture modelling of samples drawn one at a time in continuous-time. A novel aspect they included in the model is that the sampling times may be unavailable.

Second, in a standard mark-recapture study the researcher knows the list of marks in the population (for example, coloured paint, numbered tags, etc.). In mark-recapture studies using noninvasive DNA, it is difficult to know whether a previously unrecorded mark (genotype) is an error in the genotyping or a new individual, unless all the genotypes in the population are known, which is virtually impossible, because the genotypic mark is inherent to the individual. However, it is important to clarify that the marks can be misread in standard capture-recapture sampling, and they can also fall out (e.g. the paint might wash off, the tag might fall off). Then, it can face similar uncertainty issues as those in DNA-based studies.

Lukacs & Burnham (2005*b*) established that because it is impossible to know the genetic identities of every individual in the population two problems can result. First, the misidentification of individuals can occur which is better known as *genotyping error*. In traditional studies of mark-recapture, if a mark does not match with a mark from the known list, the observation is eliminated or, otherwise, corrected by the researcher. In DNA-based mark-recapture, if an incorrect genotype is logged, it is recorded as a new individual in the population. As a consequence, the size of the population will most likely be overestimated. Second, the authors point out that the marks may not be unique. In small and inbred populations, some animals may have the same genotypic profile. In this case, it is impossible to know if samples with identical genotypes are the same animal or close relatives. Consequently, the exclusion of individuals may underestimate the population.

Few models for estimating abundance incorporate the genotyping error. For instance, Lukacs & Burnham (2005*a*) extended the likelihood model of Otis et al. (1978) by considering the case of misidentification of individuals. They incorporated into the model the probability that a genotype (observed for the first time) is identified correctly for estimating the size of a closed population. Yoshizaki et al. (2011) further developed this model to improve the bias and precision of estimators. Wright (2011) and Bran (2018) modelled the uncertainty in the assignment of genotypes to faecal pellets of badgers to estimate abundance of this species. They implemented MCMC algorithms for simulating the posterior density involving the sample size. The former considered a Gibbs sampling, and the latter designed a reversible jump MCMC (Green, 1995). The uncertainty was due to a failure produced during the process of DNA amplification, called allelic dropout. The next section describes this genotyping error.

2. A genotyping Error

A *gene* is a sequence of DNA that codes for a heritable trait. Genes occur at specific positions on chromosomes, called *loci*. Humans and many other organisms are diploid, meaning that they inherit one set of chromosomes from each parent. Thus, for every gene, there are two possible DNA sequences called *alleles*. When two alleles have the same DNA sequence, they are *homozygous*. Otherwise, they are *heterozygous*. An individual's genotype constitutes allelic combinations at loci of interest.

Polymerase chain reaction (PCR) is a technique widely used to amplify specific regions of DNA. It is relevant because researchers often want to amplify small amounts of DNA collected from the field. A common error during PCR is *allelic dropout* which means that one allele is preferentially amplified over the other, thus erroneously genotyping the sample. For a heterozygous genotype, allelic dropout can produce a false homozygote, but this failure does not occur for homozygous genotypes. For example, if an individual has a true heterozygous genotype AB at a particular locus, but the PCR amplification is only successful for allele A, then the individual will be incorrectly genotyped as an AA homozygote. Figure 1 shows how the true genotypes may be observed and recorded when allelic dropout is present and the respective conditional probabilities. This figure was first drawn in Bran (2018).

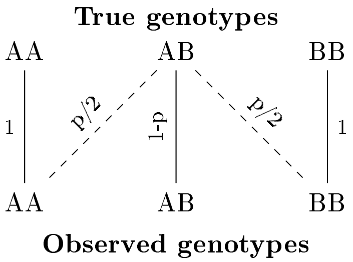

Figure 1: Dashed lines indicate allelic dropout. The true heterozygote AB is erroneously genotyped as AA or BB. The conditional probabilities are for the observed genotypes given the true genotype, Pr(Observed|True).

Notice that true homozygous genotypes are free of genotyping error. For true heterozygous, there are three possibilities. If AB is the true genotype:

- It may be wrongly observed as AA, that is, a failure to detect the allele B, with probability $p/2$.
- It may be wrongly observed as BB, that is, a failure to detect the allele A, with probability $p/2$.
- It may be correctly observed as AB with probability $1-p$.

In large populations, allelic profiles should be unique for the sampled individuals (considering that allelic profiles consist of numerous genotyped loci). However, given the procedures and conditions for amplifying DNA, genotyping errors can be introduced which may artificially increase or decrease the variation in the population and confound individual genotypes. In particular, as shown above, the use of PCR to obtain genotypes from noninvasive DNA samples complicates the identification of individuals, because the latent (actual) identities must be determined while taking into account the uncertainty of the genetic assignments.

3. Data and Model

The sample consists of $S = 47$ droppings of badgers collected from latrines in Woodchester Park, Gloucestershire, England. The dataset is originally taken from Frantz et al. (2003). The DNA extracted from each should help to determine the identity of the individuals present in the sample. Also, appendix in Wright (2011) has the information about the badger microsatellite sequences used to create the dataset, and the badger data for two PCR replicates.

A set of $L = 7$ microsatellite marker loci was considered. They were Mel102, Mel105, Mel106, Mel109, Mel111, Mel113, and Mel117 (the abbreviation Mel refers to the scientific name for the Eurasian badger, *Meles meles*). For example, 199/199 at locus Mel105 means that both the mother and the father had a common allele and so the offspring inherited the same allele from both of the parents (i.e. homozygote at that specific locus). Alternatively, 138/142 at Mel102 means that the offspring inherited one sequence from the mother which was different from the sequence from the father (i.e. heterozygote at that specific locus). The numbers in the genotypes indicate the sizes of the alleles (in base pairs). So, at Mel102, one sequence is 138 base pairs long while the other sequence is 142 base pairs long.

Thus, the genotypes are represented as a pair of positive integers, that is, the *genotype* of an individual at that locus is a pair (x, y) where $x, y \in \mathbb{Z}^+$. There is no notion of order in this definition, that is, (x, y) and (y, x) refer to the same genotype. If the numbers are equal, then the genotype is homozygous. Otherwise, it is heterozygous.

The raw numbers in a pair of microsatellite alleles is not important, but the difference between the two numbers indicates how many mutations there are between the two alleles. So, at Mel102, allele 138 has four fewer base pairs than allele 142. In medical sciences, this difference may be important for researchers looking at the association between a microsatellite sequence and a particular disease. However, in population genetics, the raw numbers and the differences between them are not directly relevant. They are used to determine whether individuals are homozygous or heterozygous at specific loci.

Frantz et al. (2003) used replication to overcome genotyping error. They replicated until either two alleles were detected or until they were confident of observing a homozygote. Replicate genotyping indicates the presence of allelic dropout when one replicate sample displays a heterozygote, and the other replicate sample displays a homozygote at the same locus. Under the presence of allelic dropout, there is no guarantee that the observed genotypes in the sample will allow the correct identification of the individuals.

The data comprises of a $S \times L \times R$ ragged array, where an element g_{jlr} is the observed genotype in the jth sample, at locus l and the rth replicate PCR amplification with $j = 1, 2, \ldots, S$, $l = 1, 2, \ldots, L$, and $r = 1, 2, \ldots, R$. For simplicity, the consensus genotype is considered which is an array of L pairs of alleles, since for every locus there are two alleles. Thus, the data is denoted by g^{obs} which comprises a $S \times L$ array, where g^{obs}_{jl} is the observed genotype in the jth sample, at locus l.

The latent information of the genotypes and the presence of individuals in the sample is stored in a $n \times L$ matrix G and a $n \times 1$ vector y, where n is the real (unknown) sample size, which are defined as, G_{ij} denotes the true genotype of the ith individual in the sample at the jth locus, for $i = 1, \ldots, n$ and $j = 1, \ldots, L$; and $y_i = k$ indicates that the ith observed genotype in the sample belongs to the kth genotype in G.

The arrays G and y together constitute the latent information about which individual was caught in each sample. They allow to define an array denoted by g^{true} which has the same dimensionality as g^{obs} but it contains the true genotypes of the respective individual sampled. The number of unique genotypes in g^{true} determines real value of n.

The unnormalized posterior density of G and y, based on the model proposed by Wright et al. (2009), is given by

$$\pi(G, y|g^{\text{obs}}) \propto \underbrace{f(g^{\text{obs}}|G, y, p)}_{\text{likelihood function}} \cdot \underbrace{f(G|N, \gamma) \cdot f(y|N)}_{\text{prior distribution}} \tag{1}$$

where γ denotes the allele frequencies, p the dropout probability p and N is the population size, which has been considered as a fixed value for simplicity in this paper. Steorts et al. (2016) and Wright et al. (2009) consider other parameters, as N, into their models. This density describes a Bayesian model for estimating the unknown parameters G and y, given the observed genotypes g^{obs}. The likelihood function accounts for the corruption process contained in the data (allelic dropout) and was explained before. It is formally defined as,

If $g = \text{AA}$ then

$$\Pr(g^{\text{obs}}|G, y, p) = \begin{cases} 1 & \text{for } \ g^{\text{obs}} = \text{AA}, \\ 0 & \text{for } \ g^{\text{obs}} = \text{AB}. \end{cases}$$

If $g = \text{AB}$ then

$$\Pr(g^{\text{obs}}|G, y, p) = \begin{cases} p/2 & \text{for } \ g^{\text{obs}} = \text{AA or BB}, \\ 1 - p & \text{for } \ g^{\text{obs}} = \text{AB}. \end{cases}$$

4. Application

This section aims to apply the algorithm implemented by Steorts et al. (2016), called (SMERED, Split and MErge REcord linkage and Deduplication) to the badgers data considered by Wright et al. (2009). Because the data considered in the latter have a genotyping error introduced by allelic dropout, which is one particular, one different and more general corruption process will be considered, that is, any corrupted genotype may be associated or linked to any other without restrictions.

4.1. New Genotyping Error

Under allelic dropout, a true homozygote AA is only linked to either AA or AB, where B $\neq$ A. However, under the new corruption process described here, it can be linked to any combination of two alleles. The cases are AA, AX, XX, and XY where X,Y $\neq$ A. For loci with two alleles, the heterozygote XY would not be a case for AA. If p denotes the probability of corruption of an allele, m the number of alleles at the locus and the corruption is independent among alleles, then the probabilities of these possible cases are as follows.

For the first case, because the two alleles are not corrupted,

$$\Pr(g^{\text{obs}} = \text{AA}|g = \text{AA}) \quad = \quad (1-p)^2.$$

For the second case,

$$\Pr(g^{\text{obs}} = \text{AX}|g = \text{AA}) \quad = \quad \frac{2p(1-p)}{m-1}$$

because there is corruption in only one allele, X is one of $m-1$ alleles distinct to A, and AX = XA. For the third case,

$$\Pr(g^{\text{obs}} = \text{XX}|g = \text{AA}) \quad = \quad \left(\frac{p}{m-1}\right)^2$$

because both alleles are corrupted and X is one of $m-1$ alleles distinct to A. For the last case,

$$\Pr(g^{\text{obs}} = \text{XY}|g = \text{AA}) \quad = \quad 2\left(\frac{p}{m-1}\right)^2$$

because both alleles are corrupted, X and Y are one of $m-1$ alleles distinct to A, and XY = YX.

For the case in which the true genotype is heterozygous, say AB, the possible outcomes for the observed genotypes are AB, AA, BB, AX, BX, XX, and XY, where $\{\text{X,Y}\} \cap \{\text{A,B}\} = \emptyset$. Their probabilities can be found following a similar process. The probabilities of the new corruption process are summarised below, where p is the probability of corruption of an allele, and m the number of alleles at the locus.

For $g = \text{AA}$,

$$\Pr(g^{\text{obs}}|G, y, p) = \begin{cases} (1-p)^2 & \text{if } g^{\text{obs}} = \text{AA}, \\ \left(\dfrac{p}{m-1}\right)^2 & \text{if } g^{\text{obs}} = \text{XX} \text{ with } \text{A} \neq \text{X}, \\ 2\left(\dfrac{p}{m-1}\right)^2 & \text{if } g^{\text{obs}} = \text{XY} \text{ with } \text{A} \neq \text{X} \text{ and } \text{A} \neq \text{Y}, \\ \dfrac{2p(1-p)}{m-1} & \text{otherwise.} \end{cases} \tag{2}$$

For $g = \text{AB}$,

$$\Pr(g^{\text{obs}}|G, y, p) = \begin{cases} (1-p)^2 + \left(\dfrac{p}{m-1}\right)^2 & \text{if } \ g^{\text{obs}} = \text{AB}, \\ \left(\dfrac{p}{m-1}\right)^2 & \text{if } \ g^{\text{obs}} = \text{XX} \ \text{ with } \ \text{X} \neq \text{A} \ \text{ and } \ \text{X} \neq \text{B}, \\ \dfrac{p(1-p)}{m-1} & \text{if } \ g^{\text{obs}} = \text{AA} \ \text{ or } \ g^{\text{obs}} = \text{BB}, \\ 2\left(\dfrac{p}{m-1}\right)^2 & \text{if } \ g^{\text{obs}} = \text{XY} \ \text{ with } \ \{\text{X,Y}\} \cap \{\text{A,B}\} = \emptyset, \\ \dfrac{p(1-p)}{m-1} + \left(\dfrac{p}{m-1}\right)^2 & \text{otherwise.} \end{cases} \tag{3}$$

The probabilities in equations 2 and 3 must add up to 1. The sums can be found in Appendix A at the end of this paper.

4.2. The Algorithm

SMERED, proposed by Steorts et al. (2016), is a hybrid MCMC algorithm. It utilizes the split-merge operations, as in Jain & Neal (2004), to jointly update G and y inside a Metropolis step. The algorithm starts by randomly choosing a pair of records. If they are associated with the same individual, then a split is proposed. Otherwise, they are merged. The algorithm is available in the supplementary material of Steorts et al. (2016). However, it is shown here as it was implemented using R software.

Algorithm 1 SMERED (Split and MErge REcord linkage and De-duplication)

```
Data: $g^{\text{obs}}, N, p$ and $\gamma$
Initializers: $G$ and $y$
Draw a pair of observations, say $i$ and $j$ for some $i \neq j$ in $\{1, \ldots, S\}$ at random.
if $y_i = y_j$ then
    Propose splitting that individual, shifting $y$ to $y^*$
else
    Propose merging the individuals who $i$ and $j$ refer to, shifting $y$ to $y^*$
end if
Update $G$ using the observations, shifting $G$ to $G^*$
Calculate $r = \min(1, \pi(G^*, y^*|g^{\text{obs}}, N, p, \gamma)/\pi(G, y|g^{\text{obs}}, N, p, \gamma))$
Set $y^{\text{new}} = y^*$ with probability $\min(1, r)$. Otherwise, set $y^{\text{new}} = y$
Update $G^*$ by using its full conditional density given $y^{\text{new}}$, shifting $G^*$ to $G^{\text{new}}$
return $G^{\text{new}}, y^{\text{new}}$
```

4.3. Implementation of the Algorithm

The following is a toy example to illustrate the performance of the algorithm proposed by Steorts et al. (2016). The purpose of the small dataset is to take advantage of the small size of the state space, which allows the inclusion of the analytical joint distribution of interest. In this way, comparisons between the simulated and the exact distributions are achievable.

Consider a sample with $S = 2$ observed genotypes at a single locus with $m = 2$ alleles, as follows.

$$g^{\text{obs}} = \begin{pmatrix} 1,1 \\ 1,2 \end{pmatrix}.$$

According to the new corruption process, if an observed genotype is corrupted, then it could be any one of these true genotypes: $\{(1,1),(1,2),(2,2)\}$. The state space, denoted by $\mathcal{S}$, has $3^2 = 9$ elements, that is, g^{true} might be any of the following.

$$\mathcal{S} = \left\{ \begin{pmatrix} 1,1 \\ 1,1 \end{pmatrix}, \begin{pmatrix} 1,1 \\ 1,2 \end{pmatrix}, \begin{pmatrix} 1,1 \\ 2,2 \end{pmatrix}, \dots, \begin{pmatrix} 2,2 \\ 1,2 \end{pmatrix}, \begin{pmatrix} 2,2 \\ 2,2 \end{pmatrix} \right\}.$$

Defining $N = 3, \gamma = (0.5, 0.25, 0.25)'$ and $p = 0.25$, the SMERED algorithm is implemented to draw samples from the posterior distribution in equation (1). The density function $f(g^{\text{obs}}|G, y, p)$ has been defined above, $f(G|N,\gamma)$ is determined by γ, and $f(y|N)$ is defined as

$$f(y|N) = \frac{N!}{(N-n)!}\left(\frac{1}{N}\right)^S$$

The aim with this small example is to compare the exact posterior distribution to that simulated by SMERED. The following procedure shows how to find the exact probability for each element in $\mathcal{S}$. For example, if $g^{\text{true}} = \begin{pmatrix} 2,2 \\ 1,2 \end{pmatrix}$, the respective values of G and y are

$$G = \begin{pmatrix} 1,2 \\ 2,2 \end{pmatrix}, \quad y = (2,1).$$

So, each of the terms involved in the posterior density in (1) are calculated as follows.

$$\begin{aligned} f(g^{\text{obs}}|G,y,p) &= \left(\frac{p}{m-1}\right)^2 \left((1-p)^2 + \left(\frac{p}{m-1}\right)^2\right) = 0.0390625 \\ f(y|N) &= \frac{N!}{(N-n)!}\left(\frac{1}{N}\right)^S = \frac{3!}{(3-2)!}\left(\frac{1}{3}\right)^2 = \frac{2}{3} \\ f(G|N,\gamma) &= \gamma_{1,2} \cdot \gamma_{2,2} = 0.25 \cdot 0.25 = 0.0625 \end{aligned}$$

As follows, the posterior probability of g given $g^{\text{obs}}, \gamma, N$, and p is proportional to the product of these three terms, which is equal to 0.0016276042. Repeating this process for all nine states in $\mathcal{S}$, and finding the normalizing constant, the exact probabilities are given by,

$$(0.33123, 0.2761, 0.1656, 0.0552, 0.0920, 0.0276, 0.0184, 0.0153, 0.0184).$$

Figure 2 shows the results for 100 000 iterations of SMERED. The plot for the simulated distribution is represented by the dotted red line, while the solid black line represents the exact distribution. It seems that the Markov chain generated by the algorithm does not converge to the correct stationary distribution.

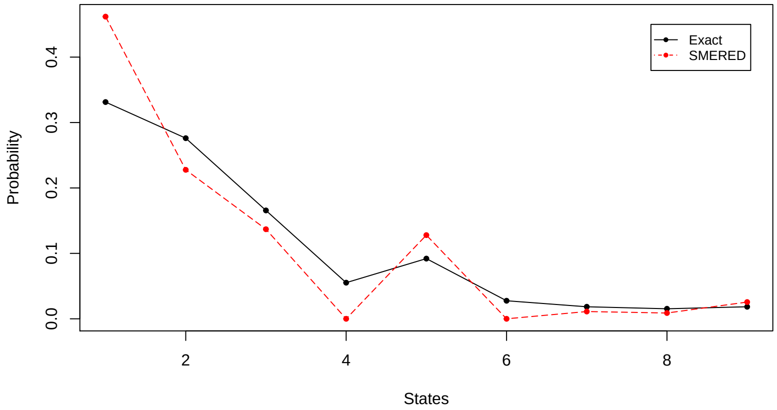

Figure 2: Exact invariant distribution vs. simulated under the new corruption process

This small example, with a manageable and available state space, shows that the Markov chain generated by the hybrid MCMC algorithm and the exact invariant distribution do not match. It casts serious doubt on the existence of the invariant distribution. The first reason might be the choice of the proposal density as symmetric, when following SMERED procedure. This is not necessarily symmetric as the sampling occurs from g^{obs} (step 9). Later, G is resampled (step 12), as explained in Steorts et al. (2016). Because reversibility might not be satisfied, the existence of a unique stationary distribution cannot be ensured.

Robert & Casella (2004) state that the existence of the invariant distribution of a Markov chain generated by a Metropolis algorithm follows by construction. The reason is that the Metropolis ratio is defined such that the transition kernel satisfies the reversibility condition, as clearly explained by Chib & Greenberg (1995). For the particular case of the SMERED algorithm, it seems that the proposal distribution for sampling a pair (G, y) was assumed to be symmetric, since the relevant ratio does not appear in the expression shown in step 10 of the algorithm 1. Besides, Gelman et al. (2004) state that asymmetric proposal distributions can be beneficial in order to speed up the evolution of the chain.

Also, the failure to converge might be closely connected to the dimension of G, as a consequence of the split-merge operations applied. They are based on the procedure considered by Jain & Neal (2004) in the context of Dirichlet process mixture models. The idea is to enhance a Metropolis-Hastings algorithm regarding its efficiency for moving through the space state, and this is done by splitting or merging the mixture components. The approach in Jain & Neal (2004) takes full advantage of the conjugacy in the model to analytically integrate over the mixing proportions and component-specific parameters, leaving only the latent indicator variables. These indicators are then updated through splitting and merging steps.

In the context here, the latent indicator variables mentioned above correspond to the values in y; and the parameters for each component correspond to G. Indeed, y is updated using split-merge operations, but G is jointly updated, instead of marginalised. The problem is that there are no such conjugacy properties in the model for integrating away G. Instead, the process for updating G and y causes the dimension of G to increase (when splitting) or decrease (when merging) by one unit at each iteration of the algorithm. Thus, if the change in the dimensionality of G is taken into account, then G and y could potentially be jointly and correctly updated. Bran (2018) designed an algorithm that takes into account this dimensionality change based on a reversible jump MCMC proposed by Green (1995) by using split-merge operations.

5. Conclusion

To sum up, this paper highlighted the potential of noninvasive DNA sampling. These methods contribute to the conservation and management of wildlife while minimizing disturbances and maintaining the well-being of the studied species. By integrating these practices into research activities, scientists and conservationists have a better understanding of their population dynamics, genetic health, and habitat requirements without directly disturbing the animals. However, noninvasive DNA sampling faces some its own challenges. For example, the success rate of obtaining DNA samples might be lower compared to direct sampling methods, and sample quality can vary. In particular, genotyping error, as allelic dropout, results in uncertainty regarding the number of observed individuals (sample size).

Additionally, the implementation of a Metropolis algorithm to update the latent information of the true genotypes and their presence in the observations, generated a Markov chain that was unable to simulate the exact distribution. Indeed, the decision to utilize a toy example was driven by the fact that the entire sample space is readily available. Given its smaller scale, working with a larger sample size would not be meaningful in showcasing the limitations or failures of the approach.

Undoubtedly, the nonreversibility of the chain emerges as a critical issue that demands meticulous consideration and revision. This concern likely arises from the assumption of a symmetric proposal distribution, which, if inaccurate, can significantly impact the Metropolis ratio. Addressing this issue becomes paramount for the robustness of the SMERED algorithm.

A potential avenue for improvement involves contemplating the dimension change of the parameter space, particularly in the context of the variable G. Delving deeper into the dynamics of dimensionality changes may prove instrumental in enhancing the efficacy of split-merge operations within the algorithm. This consideration opens up possibilities for refining the performance of the algorithm and addressing challenges related to nonreversibility. As such, a comprehensive exploration of the implications of dimension changes could shed valuable insights on strategies to overcome the identified issues.

[Received: May 2023 — Accepted: December 2023]

References

Barker, R. J., Schofield, M. R., Wright, J. A., Frantz, A. C. & Stevens, C. (2014), ‘Closed-population capture-recapture modeling of samples drawn one at a time’, *Biometrics* **70**(4), 775–782.

Berg, N. & Lien, D. (2009), ‘Sexual orientation and self-reported lying’, *Review of Economics of the Household* **7**(1), 83–104.

Biffi, D. & Williams, D. A. (2017), ‘Use of non-invasive techniques to determine population size of the marine otter in two regions of Perú’, *Mammalian Biology* **84**, 12–19.

Bran, P. (2018), Properties of Gibbs samplers for inference in genetic mark-recapture models, PhD thesis, University of Otago, Dunedin, New Zealand.

Chib, S. & Greenberg, E. (1995), ‘Understanding the Metropolis-Hastings algorithm’, *The American Statistician* **49**(4), 327–335.

Ditmer, M., Vincent, J., Werden, L., Tanner, J., Laske, T., Iaizzo, P., Garshelis, D. & Fieberg, J. (2015), ‘Bears show a physiological but limited behavioral response to unmanned aerial vehicles’, *Current Biology* **25**(17), 2278–2283. http://dx.doi.org/10.1016/j.cub.2015.07.024

Frantz, A. C., Pope, L. C., Carpenter, P. J., Roper, T. J., Wilson, G. J., Delahay, R. J. & Burke, T. (2003), ‘Reliable microsatellite genotyping of the Eurasian badger (Meles meles) using faecal DNA’, *Molecular Ecology* **12**(6), 1649–1661.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL.

Green, P. J. (1995), ‘Reversible jump Markov chain Monte Carlo computation and Bayesian model determination’, *Biometrika* **82**(4), 711–732.

Jain, S. & Neal, R. M. (2004), ‘A split-merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture model’, *Journal of Computational and Graphical Statistics* **13**(1), 158–182.

Lukacs, P. M. & Burnham, K. P. (2005*a*), 'Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error', *The Journal of Wildlife Management* **69**(1), 396–403.

Lukacs, P. M. & Burnham, K. P. (2005*b*), 'Review of capture–recapture methods applicable to non-invasive genetic sampling', *Molecular Ecology* **14**(13), 3909–3919.

Marucco, F., Vucetich, L. M., Peterson, R. O., Adams, J. R. & Vucetich, J. A. (2012), 'Evaluating the efficacy of non-invasive genetic methods and estimating wolf survival during a ten-year period', *Conservation Genetics* **13**(6), 1611–1622.

Mondol, S., Ullas Karanth, K., Samba Kumar, N., Gopalaswamy, A. M., Andheria, A. & Ramakrishnan, U. (2009), 'Evaluation of non-invasive genetic sampling methods for estimating tiger population size', *Biological Conservation* **142**(10), 2350–2360.

Morin, D. J., Kelly, M. J. & Waits, L. P. (2016), 'Monitoring coyote population dynamics with fecal DNA and spatial capture-recapture', *Journal of Wildlife Management* **80**(5), 824–836.

Otis, D. L., Burnham, K. P., White, G. C. & Anderson, D. R. (1978), 'Statistical inference from capture data on closed animal populations', *Wildlife Monographs* (62), 3–135.

Robert, C. P. & Casella, G. (2004), *Monte Carlo statistical methods*, Springer, New York.

Roques, S., Furtado, M., Jácomo, A. T. A., Silveira, L., Sollmann, R., Tôrres, N. M., Godoy, J. A. & Palomares, F. (2014), 'Monitoring jaguar populations panthera onca with non-invasive genetics: a pilot study in brazilian ecosystems', **48**(3), 361–369.

Steorts, R. C., Hall, R. & Fienberg, S. E. (2016), 'A Bayesian approach to graphical record linkage and deduplication', *Journal of the American Statistical Association* **111**(516), 1660–1672.

Wright, J. A. (2011), Incorporating Genotype Uncertainty into mark-recapture-Type models for Estimating Abundance using DNA Samples, PhD thesis, University of Otago.

Wright, J. A., Barker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E. & Gleeson, D. M. (2009), 'Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples', *Biometrics* **65**(3), 833–840.

Yoshizaki, J., Brownie, C., Pollock, K. & Link, W. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture-recapture studies', *Environmental and Ecology Statistics* **18**(1), 27–55.

Appendix A. Verifying Probabilities Sum

To ensure that the probabilities in equations 2 and 3 have been correctly specified, the sum to 1.0 for each will be examined. The idea is to count how many cases hold the condition of g^{obs}. Tables A1 and A2 show the sums for both true homozygote and true heterozygote cases, respectively. The third column of each table sums 1.0.

Table A1: Counting cases for g^{obs} when $g = \text{AA}$, and $\text{X} \neq \text{A}$

g^{obs}	Counting	Counting$\cdot$ $\Pr(g^{\text{obs}} \mid G, y, p)$
AA	1	$(1-p)^2$
XX	$m-1$	$p^2/(m-1)$
XY	$(m-1)(m-2)/2$	$(m-2)p^2/(m-1)$
AX	$m-1$	$2p(1-p)$

Table A2: Counting cases for g^{obs} when $g = \text{AB}$, and $\text{X,Y} \notin \{\text{A,B}\}$

g^{obs}	Counting	Counting$\cdot$ $\Pr(g^{\text{obs}} \mid G, y, p)$
AB	1	$(1-p)^2 + \left(\dfrac{p}{m-1}\right)^2$
XX	$m-2$	$(m-2)\left(\dfrac{p}{m-1}\right)^2$
AA or BB	2	$\dfrac{2p(1-p)}{m-1}$
XY	$\dfrac{(m-3)(m-2)}{2}$	$(m-3)(m-2)\left(\dfrac{p}{m-1}\right)^2$
AX or BX	$2(m-2)$	$2(m-2)\left[\dfrac{p(1-p)}{m-1} + \left(\dfrac{p}{m-1}\right)^2\right]$

Summing the third column in Table A1,

$$\begin{aligned}\sum_{g^{\text{obs}}} \Pr(g^{\text{obs}} \mid g = \text{AA}, p) &= 1 - 2p + p^2 + \frac{p^2}{m-1} + \frac{m-2}{m-1}p^2 + 2p - 2p^2 \\ &= 1 + \left(\frac{1}{m-1} + \frac{m-2}{m-1} - 1\right)p^2 \\ &= 1.\end{aligned}$$

Similarly, summing the third column in Table A2,

$$\begin{aligned}\sum_{g^{\text{obs}}} \Pr(g^{\text{obs}} \mid g = \text{AB}, p) &= (1-p)^2 + (m-1)^2\left(\frac{p}{m-1}\right)^2 + 2(m-1)\frac{p(1-p)}{m-1} \\ &= 1 - 2p + p^2 + p^2 + 2p - 2p^2 \\ &= 1.\end{aligned}$$