DNA mixture analysis

Master's thesis
University Utrecht
Mathematical sciences
Supervisor: Martin Bootsma, second reader: Tobias Müller

Mark Dirksen 3245756

August 10, 2014

A model for peak height distribution using the bivariate normal distribution is proposed to deal with some flaws of the already established model using the gamma distribution. This new model is compared to the "Gamma model" via artificially generated test data using a model that simulates the PCR process. In addition the performance of these models is compared using data from an actual case. To efficiently execute the computations necessary, techniques using bayesian networks are utilized.

Contents

1 Introduction

When multiple individuals contribute DNA to a sample the interpretation of the resulting DNA profile is not a straightforward matter. In addition to unknown contributors and uncertainty regarding proportions of DNA contributions between individuals, the process of obtaining a DNA profile from a sample is of a stochastic nature and hence subject to uncertainty.

In a recent paper, [6, Cowell et al. (2013)] presented a statistical model for the peak height distribution using the gamma distribution. This makes it possible to evaluate a likelihood function subject to certain parameters. For efficient computation of this likelihood function a technique using bayesian networks is utilized. This peak height model, termed the Gamma model, is presented in section 4.1. The computational methods utilized in this paper are reported in section 5.

This thesis will start with an overview of forensic DNA typing in section 2. Next a simulation model from [3, Gill et al.(2005)] with which realistic test data can be generated will be presented in section 3. Section 4 will then describe the Gamma model, as well as a competing model using the bivariate normal distribution termed the Bivariate Normal model. In section 5 the computational methods of [6, Cowell et al. (2013)] are discussed, as well as how these need to be adapted to the Bivariate Normal model. A method for including some marker dependence is introduced as well. The results, in section 6, are split into two parts. The first attempts to compare the performance of the Gamma- and Bivariate Normal model through simulated test data. The second utilizes both methods on an example case. Finally this thesis closes with some points of discussion in section 7.

2 Overview forensic DNA typing

Variation in human DNA is a result of variation in the nucleobases of the DNA molecule. The DNA alphabet is composed of four characters that represent the four different nucleobases: A (adenine), T(thymine),C (cytosine) and G (guanine). Human DNA can be read in order just like written language is read from left to right. The combinations of these four letters, known as nucleotides or bases, make up the biological differences among humans. We all have approximately three billion nucleotide positions, which gives a lot of potential sequences.

A marker is a small portion of this sequence used in DNA typing. The position of a DNA marker is referred to as a locus. The markers are chosen because they exhibit a great amount of variation among the human population, as well as being greatly removed from each other. Either they are on different chromosomes, or a large sequence of nucleotides separate the two markers. This last fact is then used to make the assumption that DNA frequencies between these markers are independent. Variation in the forensic markers used for DNA analysis is measured in 'repeat numbers'. As the name suggests, a repeat number is the number of times a specific sequence of nucleobases is repeated. The precise composition of this sequence differs between markers. These sequences can include a partial repeat, which corresponds to a partial repeat number. This is written

as a decimal number, with the decimal being equal to the number of nucleobases in the partial repeat. To illustrate:

Repeat number=4 $\cdots$(AATG)(AATG)(AATG)(AATG)$\cdots$

Repeat number=4.2 $\cdots$(AATG)(AATG)(AATG)(AATG)(AA)$\cdots$

The possibilities in variation at a genetic locus are termed alleles. For forensic DNA markers the alleles correspond to a (partial) repeat number. At each locus a person has a maternal and a paternal allele. If they are different they are called heterozygous and homozygous if identical. The characterization of the alleles present at a locus is called a genotype.

Before anything is measured the DNA sample in question is subjected to a process called PCR (polymerase chain reaction) amplification. This is a process in which a specific region of DNA is replicated over and over again. During each cycle, a copy of the target DNA sequence is generated for every molecule containing the target sequence. Two important parts of the DNA sequence are the short DNA sequences that flank the region to be copied. These sequences are called primers. Primers need to be added to the sample for the PCR to work and will "select" the right part to copy. These primer sequences should be short and unique to ensure good results. After approximately 30 cycles of PCR amplification sufficient copies have been created to be easily measured. Unfortunately this copying process is not always successful, which leads to variability in the amount of DNA molecules after PCR aplification. To measure the amounts of DNA after amplification a fluorescent dye is attached to a PCR primer that is incorporated into the amplified target region of DNA. Fluorescence measurements involve exciting this dye molecule and then detecting the light that is emitted from the excited dye. These measurements are then converted to an electropherogram (EPG), in which the horizontal axis gives the base pair measurement and the vertical axis the light intensity. Thus each allele of a marker corresponds to a peak size, which is a measure for the light intensity emitted. This in turn is a measure for the amount of DNA molecules of this allele type after PCR amplification. Peak size can be measured in peak height, or peak area. These are highly correlated [2, Tvedebrink et al. (2010)]. Throughout this text peak heights will be used as the standard of measurement. The collection of peak heights for all the markers used and for all the possible alleles from the EPG will be termed a DNA profile.

The PCR process is subject to several artefacts which can make analyzing DNA profiles difficult. Peak heights are commonly subject to low level noise which result in small peaks. Usually a threshold is placed below which peaks are reported as non-existent. Thus, an allele present in the DNA sample will not be recorded if the resulting peak falls below this threshold. This is called dropout. Dropout can also occur due to a complete failure to amplify. In addition to dropout, another common artefact is called stutter. These stutter products arise from the PCR process by the occasional imperfect copy. Errors in which the DNA molecule loses one repeat number is the most common, but gaining a repeat number is also possible. This usually results in a small peak (or a

contribution to an already existing peak) one repeat number below an allele peak. Other artefacts that can occur are dropin due to contamination of the sample by very small amounts of DNA, and an allele is called silent if a mutation occurs that results in the allele not being picked up at all by the PCR process.

3 A simulation method for DNA peak distributions

In [3, Gill et al.(2005)] a simulation method that closely follows the PCR process was presented. The process follows the path: *DNA sample→Extraction→Aliquot into pre-PCR reaction mixture→PCR amplification for t cycles→Visualization of alleles after electrophoresis*. The combination of these steps is responsible for the stochastic nature of the PCR process:

- *DNA sample*: N will denote the number of cells in the DNA sample.
- *Extraction*: During the process of extraction, the cells are disrupted and the DNA liberated into solution. During extraction, there is a probability $\pi_{extraction}$ (the extraction efficiency) that an individual DNA molecule will survive the process, independent of the other molecules in the sample. Thus the number of DNA molecules extracted ($N_{extracted}$) follows a binomial distribution: $N_{extracted} = \text{Bin}(N, \pi_{extraction})$.
- *Aliquot into pre-PCR reaction mixture*: A portion of the extracted sample is submitted for PCR. Therefore, there is a probability $\pi_{aliquot}$ that a given molecule will be selected. The number of DNA molecules in the aliquot ($N_{aliquot}$) also follows a binomial distribution: $N_{aliquot} = \text{Bin}(N_{extracted}, \pi_{aliquot})$.
- *PCR amplification for t cycles*: PCR is not 100% efficient. Thus, during each round there will be a probability $\pi_{PCReff} < 1$ that a DNA fragment will be amplified. Each fragment will also have a small probability $\pi_{stutter}$ of losing a repeat number. After stuttering the DNA fragment will be amplified during future cycles with the same efficiency as it was when it had its original repeat number, and will also have the same chance of losing another repeat number.
- *Visualization of alleles after electrophoresis*: The number of DNA fragments will be converted to a peak height. This step will be considered deterministic.

Each of these steps and each cycle of the PCR amplification is mutually independent. Then by the properties of the binomial distribution we can make the following simplification:

$$N_{aliquot} = \text{Bin}(N, \pi_{extraction} * \pi_{aliquot}) = \text{Bin}(N, \pi_{extraction,aliquot}). \quad (1)$$

Hereby reducing the parameter set to N, $\pi_{extraction,aliquot}$, π_{PCReff}, and $\pi_{stutter}$.

By using Monte Carlo simulation following these steps a great amount of artificial, yet realistic, test data can be generated. Note that it is possible for stutter peaks to arise at alleles two or more repeat numbers lower than the allele type of the DNA molecules of the sample.

4 Modeling DNA peak distributions

Let I be the number of (potential) contributors to the DNA mixture and M the number of markers used in the analysis of the mixture. A_m will denote the number of allelic types of marker m, with $m \in \{1, \ldots, M\}$. ϕ_i will be defined as the fraction of DNA contributed to the mixture by individual $i \in \{1, \ldots, I\}$ to the DNA mixture prior to PCR amplification. Then $\phi_i \geq 0$ and $\sum_{i=1}^{I} \phi_i = 1$. Pre-amplification DNA contributions by individuals are considered constant across markers.

Both models presented in this section incorporate the possibility of stutter and dropout. Stutter whereby the DNA molecule gains a repeat number, or loses more than one repeat number is ignored.

4.1 Gamma model for DNA peak distributions

The distribution of DNA peaks was modeled using the gamma distribution by [4, Cowell et al. (2007a)] and later refined to include artefacts in [5, Cowell et al. (2011)]. Finally some small changes were made to ease computations and to deal with artefacts more realistically in [6, Cowell et al. (2013)]. The model from the final paper will be presented here.

4.1.1 Model without artefacts

The model describes the observed peak height H_a, where a denotes the allelic type (or equivalently the repeat number). Let ρ and η be parameters, and n_{ia} denotes the number of alleles of type a carried by individual i. Then it is assumed that H_{ia}, the contribution of individual i to the observed peak height at allele a is Gamma distributed: $H_{ia} \sim \Gamma(\rho\phi_i n_{ia}, \eta)$ where $\Gamma(\alpha, \beta)$ denotes the distribution with density

$$f(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-x/\beta} \text{ (for } x > 0). \tag{2}$$

Furthermore ρ is assumed to be proportional to the total amount of DNA in the mixture prior to amplification. η determines the scale.

Using the property that the set of independent Gamma distributions with equal scale parameters η is closed under summation, we write,

$$H_a \sim \sum_i H_{ia} \sim \sum_i \Gamma(\rho\phi_i n_{ia}, \eta) \sim \Gamma(\rho \sum_i \phi_i n_{ia}, \eta). \tag{3}$$

4.1.2 Stutter

Stutter is a frequent result of PCR amplification and has to be taken into account when modeling this process. An additional parameter will be introduced for this purpose, ξ, defined as the mean stutter fraction. The contribution of individual i to the observed

peak height at allele a, H_{ia} is then decomposed into two independent Gamma distributed random variable as follows:

$$H_{ia} = H_{ia}^s + H_{ia}^0. \tag{4}$$

Here H_{ia}^s is the fraction that stutters to allele $a-1$ and H_{ia}^0 is the remaining contribution. The components are assumed to be distributed as

$$H_{ia}^s \sim \Gamma(\rho\xi\phi_i n_{ia}, \eta), \quad H_{ia}^0 \sim \Gamma(\rho(1-\xi)\phi_i n_{ia}, \eta). \tag{5}$$

The total peak height observed at allele a is then

$$H_a = \sum_i H_{ia}^0 + \sum_i H_{i,a+1}^s = H_a^0 + H_{a+1}^s, \tag{6}$$

which is Gamma distributed as well

$$H_a \sim \sum_i \Gamma(\rho\xi\phi_i n_{ia}, \eta) + \sum_i \Gamma(\rho(1-\xi)\phi_i n_{i,a+1}, \eta) \sim \Gamma(\rho(1-\xi)\sum_i \phi_i n_{ia} + \rho\xi \sum_i \phi_i n_{i,a+1}, \eta). \tag{7}$$

4.1.3 Dropout

Another common artefact due to PCR amplification is dropout. A peak is considered "dropped out" if it falls below a predetermined threshold C. This can be due to a complete failure to amplify, or the peak can simply fail to amplify sufficiently to cross the threshold C. If we ignore the case where there is a complete lack of amplification we can simply account for dropout by defining

$$Z_a = \begin{cases} H_a & \text{if } H_a \geq C \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

and then equating dropout with Z_a being equal to 0. With G denoting the cumulative distribution function of the Gamma distribution:

$$\mathbb{P}(\text{Dropout}) = \mathbb{P}(Z_a = 0) = G(C; \rho(1-\xi)\sum_i \phi_i n_{ia} + \rho\xi \sum_i \phi_i n_{i,a+1}, \eta). \tag{9}$$

4.2 Bivariate Normal model for DNA peak distributions

Quite similar to the Gamma model a bivariate normal distribution can be used to describe variability in peak heights. In this case the stutter peak of allele $a-1$ (the contribution to the peak height at allele $a-1$ due to stutter) and the allelic peak at allele a (the contribution to the peak height at allele a with stutter contributions from allele $a+1$ omitted) will not be independent in general. Specifically it is assumed that (H_{ia}^s, H_{ia}^0), the contribution of individual i to the stutter peak height at allele $a-1$ and the allelic peak height at allele a is bivariate normally distributed. Let n_{ia} again

denote the number of alleles of type a carried by individual i and additionally let r, ξ, m, $sd_{stutter}$ and $sd_{allelic}$ be parameters. With $(X_1, X_1) \sim N(r, m_1, m_2, sd_1, sd_2)$ denoting a bivariate normal distribution of random variables X_1 and X_2 with $r = \frac{\mathrm{Cov}(X_1,X_2)}{sd_1 sd_2}$, $m_i = \mathbb{E}(X_i)$ and $sd_i^2 = \mathrm{Var}(X_i)$ for $i \in \{1, 2\}$:

$$(H_{ia}^s, H_{ia}^0) \sim N(r, \xi m \phi_i n_{ia}, (1-\xi) m \phi_i n_{ia}, sd_{stutter}\sqrt{\phi_i n_{ia}}, sd_{allelic}\sqrt{\phi_i n_{ia}}). \quad (10)$$

Thus the parameters r, ξ, m, $sd_{stutter}$ and $sd_{allelic}$ control the correlation between H_{ia}^s and H_{ia}^0, the stutter percentage, mean peak heights, variability of stutter peaks and variability of allelic peaks respectively. m, like ρ in the gamma model, is assumed to be proportional to the total amount of DNA in the mixture prior to amplification.

Sums of independent bivariate normal distributions remain bivariate normal. Then $(H_a^s, H_a^0) = (\sum_i H_{ia}^s, \sum_i H_{ia}^0)$ is bivariate normal with:

$$\mathbb{E}(H_a^s) = \sum_i \mathrm{E}(H_{ia}^s) = \sum_i \xi m \phi_i n_{ia} = \xi m \sum_i \phi_i n_{ia}, \quad (11a)$$

$$\mathbb{E}(H_a^0) = \sum_i \mathrm{E}(H_{ia}^0) = \sum_i (1-\xi) m \phi_i n_{ia} = (1-\xi) m \sum_i \phi_i n_{ia}, \quad (11b)$$

$$\mathrm{Var}(H_a^s) = \sum_i \mathrm{Var}(H_{ia}^0) = \sum_i sd_{stutter}^2 \phi_i n_{ia} = sd_{stutter}^2 \sum_i \phi_i n_{ia}, \quad (11c)$$

$$\mathrm{Var}(H_a^0) = \sum_i \mathrm{Var}(H_{ia}^s) = \sum_i sd_{allelic}^2 \phi_i n_{ia} = sd_{allelic}^2 \sum_i \phi_i n_{ia}, \quad (11d)$$

$$\begin{aligned} \mathrm{Cov}(H_a^0, H_a^s) &= \mathrm{Cov}(\sum_i H_{ia}^0, \sum_j H_{ja}^s) = \sum_{i,j} \mathrm{Cov}(H_{ia}^0, H_{ja}^s) \\ &= \sum_i \mathrm{Cov}(H_{ia}^0, H_{ia}^s) + \sum_{i \neq j} \mathrm{Cov}(H_{ia}^0, H_{ja}^s) \\ &= \sum_i \mathrm{Cov}(H_{ia}^0, H_{ia}^s) = sd_{stutter} sd_{allelic} r \sum_i \phi_i n_{ia}. \end{aligned} \quad (11e)$$

Conveniently the correlation coefficient reduces to r:

$$\frac{\mathrm{Cov}(H_a^0, H_a^s)}{\sqrt{\mathrm{Var}(H_a^0)\mathrm{Var}(H_a^s)}} = \frac{sd_{stutter} sd_{allelic} r \sum_i \phi_i n_{ia}}{sd_{stutter}\sqrt{\sum_i \phi_i n_{ia}} sd_{allelic}\sqrt{\sum_i \phi_i n_{ia}}} = r. \quad (12)$$

In short:

$$(H_a^s, H_a^0) \sim N(r, \xi m \sum_i \phi_i n_{ia}, (1-\xi) m \sum_i \phi_i n_{ia}, sd_{stutter}\sqrt{\sum_i \phi_i n_{ia}}, sd_{allelic}\sqrt{\sum_i \phi_i n_{ia}}). \quad (13)$$

It is assumed that H_a^s or H_a^0 being smaller than zero results in a peak height of zero.

Let θ be the parameter set $\{r, \xi, m, sd_{stutter}, sd_{allele}, \{\phi_i : i \in \{1, \ldots, I\}\}\}$ and $\{x_a : a \in \{1, \ldots, A_m\}\}$ a collection of peak heights obtained from a DNA profile. Using

the fact that the random variables $\{H_i^s + H_{i+1}^0 : 1 \leq i \leq a-1\}$ are conditionally independent of $H_a^0 + H_{a+1}^s$ (or H_a^0 if $a = A_m$) given H_a^s, the overall likelihood of a set of observed DNA heights will be iteratively decomposed (with marker dependence repressed in the notation):

$$\begin{aligned} L(\theta|x) &\propto f_{H_A^0, H_A^s + H_{A-1}^0, \ldots, H_2^s + H_1^0}(x_A, x_{A-1}, \ldots, x_1) \\ &= f_{H_A^s + H_{A-1}^0, \ldots, H_2^s + H_1^0 | H_A^0 = x_A}(x_{A-1}, \ldots, x_1) \cdot f_{H_A^0}(x_A) \\ &= f_{H_A^{s*} + H_{A-1}^0, \ldots, H_2^s + H_1^0}(x_{A-1}, \ldots, x_1) \cdot f_{H_A^0}(x_A) \\ &= f_{H_{A-1}^{s*} + H_{A-2}^0, \ldots, H_2^s + H_1^0}(x_{A-2}, \ldots, x_1) \cdot f_{H_A^{s*} + H_{A-1}^0}(x_{A-1}) \cdot f_{H_A^0}(x_A) \ldots \\ \ldots &= f_{H_A^0}(x_A) \cdot f_{H_A^{s*} + H_{A-1}^0}(x_{A-1}) \cdot \ldots \cdot f_{H_2^{s*} + H_1^0}(x_1). \end{aligned} \tag{14}$$

Here H_a^{s*} represents the updated random variable defined as follows:

$$H_a^{s*} = \begin{cases} H_{a-1}^s | (H_a^0 + H_{a+1}^{s*}) = x_a & \text{if } a \leq A-1, \\ H_a^s | (H_a^0 = x_a) & \text{if } a = A. \end{cases} \tag{15}$$

Notice that this definition is of an iterative nature.

Different cases have to be considered when updating the distribution of the stutter peak H_a^s considering the peak height information of higher alleles.

The case of the peak at allele a not dropping out ($x_a \geq C$)

If there is no dropout at allele a the updating process is straightforward. If H_{a+1}^{s*} is normally distributed,which is assumed to be true, $(H_a^s, H_a^0 + H_{a+1}^{s*})$ is also multivariate normal, with:

$$\mathbb{E}(H_a^s) = \xi m \sum_i \phi_i n_{ia}, \tag{16a}$$

$$\mathbb{E}(H_a^0 + H_{a+1}^{s*}) = (1-\xi) m \sum_i \phi_i n_{ia} + \mathbb{E}(H_{a+1}^{s*}), \tag{16b}$$

$$\mathrm{Var}(H_a^s) = sd_{stutter}^2 \sum_i \phi_i n_{ia}, \tag{16c}$$

$$\mathrm{Var}(H_a^0 + H_{a+1}^{s*}) = sd_{allelic}^2 \sum_i \phi_i n_{ia} + \mathrm{Var}(H_{a+1}^{s*}), \tag{16d}$$

$$\begin{aligned} \mathrm{Cov}(H_a^s, H_a^0 + H_{a+1}^{s*}) &= \mathrm{Cov}(H_a^s, H_a^0) + \mathrm{Cov}(H_a^s, H_{a+1}^{s*}) \\ &= sd_{stutter} sd_{allelic} r \sum_i \phi_i n_{ia} + 0. \end{aligned} \tag{16e}$$

For a bivariate normal random variable $(X_1, X_1) \sim N(r, m_1, m_2, sd_1, sd_2)$, the conditional distribution $X_1 | X_2 = x_2$ is a normal distribution with:

$$\mathbb{E}(X_1) = m_1 + r \frac{sd_1}{sd_2}(x_2 - m_2), \tag{17a}$$

$$\mathrm{Var}(X_1) = sd_1^2(1-r^2). \tag{17b}$$

Combining equations 16 and 17:

$$\begin{aligned} H_a^{s*} \sim N \Bigg(& \xi m \sum_i \phi_i n_{ia} + \frac{sd_{stutter} sd_{allelic} r \sum_i \phi_i n_{ia}}{sd_{allelic}^2 \sum_i \phi_i n_{ia} + \mathrm{Var}(H_{a+1}^{s*})} \cdot \\ & \cdot \left[x_{a+1} - (1-\xi) m \sum_i \phi_i n_{ia} - \mathbb{E}(H_{a+1}^{s*}) \right], \\ & sd_{stutter}^2 \sum_i \phi_i n_{ia} - \frac{sd_{stutter} \sqrt{\sum_i \phi_i n_{ia}}}{\sqrt{sd_{allelic}^2 \sum_i \phi_i n_{ia} + \mathrm{Var}(H_{a+1}^{s*})}} \cdot \\ & \cdot \left[sd_{stutter} sd_{allelic} r \sum_i \phi_i n_{ia} \right]^2 \Bigg), \\ & (\text{if } a \leq A_m - 1) \end{aligned} \tag{18a}$$

$$\begin{aligned} H_a^{s*} \sim N \Bigg(& \xi m \sum_i \phi_i n_{ia} + r \frac{sd_{stutter}}{sd_{allelic}} \cdot \left[x_{a+1} - (1-\xi) m \sum_i \phi_i n_{ia} \right], \\ & sd_{stutter}^2 \sum_i \phi_i n_{ia} - \frac{sd_{stutter}}{sd_{allelic}} \cdot \left[sd_{stutter} sd_{allelic} r \sum_i \phi_i n_{ia} \right]^2 \Bigg). \\ & (\text{if } a = A_m) \end{aligned} \tag{18b}$$

The case of the peak at allele a dropping out ($x_a < C$)

Unfortunately if dropout occurs at allele a things become a little more complicated. Dropout is defined in the same way as with the Gamma model:

$$\mathbb{P}(\text{Dropout}) = \mathbb{P}(Z_a = 0), \tag{19}$$

with Z_a defined as in equation 8. If $(X_1, X_1) \sim N(r, m_1, m_2, sd_1, sd_2)$, the conditional distribution $X_1 | X_2 < x_2$ is not normally distributed. Nonetheless, this distribution is assumed to be normally distributed, which will make computations significantly easier to perform. In most cases such a conditional distribution should be close enough to a normal distribution to suit our purposes, keeping in mind that if the allelic peak has dropped out, one cannot expect a large contribution from its stutter peak. Specifically we will assume:

$$H_a^{s*} = \begin{cases} N\left(\mathbb{E}(H_a^s | H_a^0 + H_{a+1}^{s*} < C), \sqrt{\mathrm{Var}(H_a^s | H_a^0 + H_{a+1}^{s*} < C)} \right) & \text{if } a \leq A_m - 1, \\ N\left(\mathbb{E}(H_a^s | H_a^0 < C), \sqrt{\mathrm{Var}(H_a^s | H_a^0 < C)} \right) & \text{if } a = A_m. \end{cases} \tag{20}$$

If $(X_1, X_1) \sim N(r, m_1, m_2, sd_1, sd_2)$:

$$\begin{aligned}
\mathbb{E}\left[g(X_1)|X_2 < C\right] &= \int_{-\infty}^{\infty} f_{X_1|X_2<C}(x_1) \cdot g(x_1) dx_1 = \int_{-\infty}^{\infty} \frac{f_{X_1,X_2<C}(x_1)}{\mathbb{P}(X_2 < C)} \cdot g(x_1) dx_1 \\
&= \int_{-\infty}^{\infty} \frac{g(x_1)}{\mathbb{P}(X_2 < C)} \cdot \left(\int_{-\infty}^{C} f_{X_1,X_2}(x_1, x_2) dx_2 \right) dx_1 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{C} \frac{g(x_1)}{\mathbb{P}(X_2 < C)} \cdot f_{X_1,X_2}(x_1, x_2) dx_2 dx_1 \\
&\overset{(1)}{=} \int_{-\infty}^{C} \frac{1}{\mathbb{P}(X_2 < C)} \left(\int_{-\infty}^{\infty} g(x_1) \cdot f_{X_1,X_2}(x_1, x_2) dx_1 \right) dx_2 \\
&= \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2 < C)} \left(\int_{-\infty}^{\infty} g(x_1) \cdot \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} dx_1 \right) dx_2 \\
&= \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2 < C)} \left(\int_{-\infty}^{\infty} g(x_1) \cdot f_{X_1|X2}(x_1) dx_1 \right) dx_2 \\
&= \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2 < C)} \mathbb{E}\left[g(X_1)|X_2\right](x_2) dx_2.
\end{aligned} \tag{21}$$

Using this the following can be easily evaluated:

$$\begin{aligned}
\mathbb{E}\left[X_1|X_2 < C\right] &= \int_{-\infty}^{C} \mathbb{E}\left[X_1|X_2 = x_2\right] \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2 < C)} dx_2 \\
&= \frac{1}{\mathbb{P}(X_2 < C)} \int_{-\infty}^{C} \left(m_1 + r \frac{sd_1}{sd_2}(x_2 - m_2) \right) \cdot \\
&\cdot \frac{1}{sd_2\sqrt{2\pi}} e^{-(x_2-m_2)^2/(2sd_2^2)} dx_2 \\
&= \frac{1}{\mathbb{P}(X_2 < C)} \cdot m_1 \cdot \mathbb{P}(X_2 < C) \\
&+ \frac{1}{\mathbb{P}(X_2 < C)} r \frac{sd_1}{sd_2} \int_{-\infty}^{C-m_2} x_2 \frac{1}{sd_2\sqrt{2\pi}} e^{-x_2^2/(2sd_2^2)} dx_2 \\
&= m_1 + \frac{1}{\mathbb{P}(X_2 < C)} r \frac{sd_1}{sd_2} \left[-\frac{2sd_2^2}{2} \frac{1}{sd_2\sqrt{2\pi}} e^{-x_2^2/(2sd_2^2)} \right]_{-\infty}^{C-m_2} \\
&= m_1 - \frac{1}{\mathbb{P}(X_2 < C)} r sd_1 \frac{1}{\sqrt{2\pi}} e^{-(C-m_2)^2/(2sd_2^2)}.
\end{aligned} \tag{22}$$

To calculate $\mathrm{Var}\left[g(X_1)|X_2 < C\right]$ some legwork is required:

$$\begin{aligned}
\mathbb{E}\left[X_1^2|X_2 = x_2\right] &= \mathrm{Var}\left[X_1|X_2 = x_2\right] + \mathbb{E}\left[X_1|X_2 = x_2\right]^2 \\
&= sd_1^2(1 - r^2) + m_1^2 + 2m_1 r \frac{sd_1}{sd_2}(x_2 - m_2) + r^2 \frac{sd_1^2}{sd_2^2}(x_2 - m_2)^2,
\end{aligned} \tag{23}$$

(1) If $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |f_{X_1,X_2}(x_1, x_2)| \, dx_1 dx_2 < \infty$.

$$
\begin{aligned}
\mathbb{E}\left[X_1^2|X_2<C\right] &= \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2<C)}\mathbb{E}\left[X_1^2|X_2=x_2\right]dx_2 \\
&= \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2<C)}(sd_1^2(1-r^2)+m_1^2)dx_2 \\
&+ \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2<C)}2m_1 r\frac{sd_1}{sd_2}(x_2-m_2)dx_2 \\
&+ \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2<C)}r^2\frac{sd_1^2}{sd_2^2}(x_2-m_2)^2dx_2 \\
&= I_1+I_2+I_3.
\end{aligned} \tag{24}
$$

With:

$$
\begin{aligned}
I_1 &= \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2<C)}(sd_1^2(1-r^2)+m_1^2)dx_2 \\
&= \frac{1}{\mathbb{P}(X_2<C)}\cdot(sd_1^2(1-r^2)+m_1^2)\cdot\mathbb{P}(X_2<C) = sd_1^2(1-r^2)+m_1^2,
\end{aligned} \tag{25a}
$$

$$
\begin{aligned}
I_2 &= \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2<C)}2m_1 r\frac{sd_1}{sd_2}(x_2-m_2)dx_2 \\
&= \int_{-\infty}^{C} \frac{1}{sd_2\sqrt{2\pi}}e^{-(x_2-m_2)^2/(2sd_2^2)}\cdot\frac{1}{\mathbb{P}(X_2<C)}\cdot 2m_1 r\frac{sd_1}{sd_2}(x_2-m_2)dx_2 \\
&= 2m_1 r\frac{sd_1}{sd_2^2}\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\mathbb{P}(X_2<C)}\cdot\int_{-\infty}^{C-m_2} x_2 e^{-x_2^2/(2sd_2^2)}dx_2 \\
&= 2m_1 r\frac{sd_1}{sd_2^2}\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\mathbb{P}(X_2<C)}\cdot\left[-sd_2^2 e^{-x_2^2/(2sd_2^2)}\right]_{-\infty}^{C-m_2} \\
&= -2m_1 r sd_1\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\mathbb{P}(X_2<C)}\cdot e^{-(C-m_2)^2/(2sd_2^2)},
\end{aligned} \tag{25b}
$$

$$
\begin{aligned}
I_3 &= \int_{-\infty}^{C} \frac{f_{X_2}(x_2)}{\mathbb{P}(X_2 < C)} r^2 \frac{sd_1^2}{sd_2^2}(x_2 - m_2)^2 dx_2 \\
&= r^2 \frac{sd_1^2}{sd_2^2} \frac{1}{\mathbb{P}(X_2 < C)} \cdot \int_{-\infty}^{C} (x_2 - m_2)^2 \frac{1}{sd_2\sqrt{2\pi}} e^{-(x_2-m_2)^2/(2sd_2^2)} dx_2 \\
&= r^2 \frac{sd_1^2}{sd_2^2} \frac{1}{\mathbb{P}(X_2 < C)} \cdot \int_{-\infty}^{C-m_2} x_2^2 \frac{1}{sd_2\sqrt{2\pi}} e^{-x_2^2/(2sd_2^2)} dx_2 \\
&= r^2 \frac{sd_1^2}{sd_2^2} \frac{1}{\mathbb{P}(X_2 < C)} \frac{1}{sd_2\sqrt{2\pi}} \cdot \left(\int_{-\infty}^{C-m_2} (x_2^2 - sd_2^2) e^{-x_2^2/(2sd_2^2)} dx_2 \right. \\
&\left. + \int_{-\infty}^{C-m_2} sd_2^2 e^{-x_2^2/(2sd_2^2)} dx_2 \right) \\
&= r^2 \frac{sd_1^2}{sd_2^2} \frac{1}{\mathbb{P}(X_2 < C)} \frac{1}{sd_2\sqrt{2\pi}} \cdot \left(\left[-x_2 sd_2^2 e^{-x_2^2/(2sd_2^2)} \right]_{-\infty}^{C-m_2} \right. \\
&\left. + \int_{-\infty}^{C-m_2} sd_2^2 e^{-x_2^2/(2sd_2^2)} dx_2 \right) \\
&= -r^2 \frac{sd_1^2}{sd_2\sqrt{2\pi}} \frac{1}{\mathbb{P}(X_2 < C)} (C - m_2) e^{-(C-m_2)^2/(2sd_2^2)} \\
&+ r^2 sd_1^2 \frac{1}{\mathbb{P}(X_2 < C)} \int_{-\infty}^{C} \frac{1}{sd_2\sqrt{2\pi}} e^{-(x_2-m_2)^2/(2sd_2^2)} dx_2 \\
&= -r^2 \frac{sd_1^2}{sd_2\sqrt{2\pi}} \frac{1}{\mathbb{P}(X_2 < C)} (C - m_2) e^{-(C-m_2)^2/(2sd_2^2)} + r^2 sd_1^2.
\end{aligned}
\tag{25c}
$$

This finally leads us to the expression:

$$
\begin{aligned}
\mathrm{Var}\left[X_1 | X_2 < C\right] &= \mathbb{E}\left[X_1^2 | X_2 < C\right] - \mathbb{E}\left[X_1 | X_2 < C\right]^2 \\
&= I_1 + I_2 + I_3 - \mathbb{E}\left[X_1 | X_2 < C\right]^2 \\
&= sd_1^2 - r^2 \frac{sd_1^2}{sd_2\sqrt{2\pi}} \frac{1}{\mathbb{P}(X_2 < C)} (C - m_2) e^{-(C-m_2)^2/(2sd_2^2)} \\
&- r^2 sd_1^2 \frac{1}{2\pi} \frac{1}{\mathbb{P}(X_2 < C)^2} e^{-(C-m_2)^2/sd_2^2}.
\end{aligned}
\tag{26}
$$

Combining equation 16 with equations 22 and 26 gives a suitable expression (which is omitted for sake of brevity) for calculating H_a^{s*} in the case of x_a dropping out.

5 Computation

5.1 Likelihood ratio

Of interest in forensic cases is the likelihood ratio (LR):

$$
\mathrm{LR}_\theta = \frac{L(\mathrm{H}_0, \theta | \mathrm{E})}{L(\mathrm{H}_1, \theta | \mathrm{E})} = \frac{\mathbb{P}(\mathrm{E} | \theta, \mathrm{H}_0)}{\mathbb{P}(\mathrm{E} | \theta, \mathrm{H}_1)} \tag{27}
$$

with H_0 and H_1 the competing hypotheses under consideration (usually the prosecution versus the defense), E the evidence and θ a set of parameters. In this case the evidence under consideration would be the DNA samples collected from the crime scene. We will also define:

$\mathbf{H_0}$ Suspect has contributed to the collected DNA sample.

$\mathbf{H_1}$ An unknown person has contributed to the collected DNA sample. This person is assumed to be randomly drawn from the population.

Thus we will concern ourselves with calculating the likelihood function $L(\mathrm{H}_j, \theta|\mathrm{E})$, $j \in \{0, 1\}$. Likelihood ratios for multiple samples S will be considered separately.

In addition to the known contributors of the sample the possibility of DNA contributions from unknown sources will be accounted for by allowing one or multiple unknown contributors. These contributors will be assumed to be randomly drawn from the population. Consider the likelihood obtained by either the Gamma or Bivariate Normal model described in section 4. The likelihood is dependent on:

- $\mathbf{n} = (n_{iam})_{i\in\{1,\ldots,I\},a\in\{1,\ldots,A_m\},m\in\{1,\ldots,M\}}$. An object with component n_{iam} being the number of alleles of type a carried by individual i for marker m.
- $\boldsymbol{\phi}_s = (\phi_i)_{i\in\{1,\ldots,I\}}$. The vector of fractions of DNA contributed to the mixture by individual i.
- $\boldsymbol{\theta}$. The set of parameter inherent to the model used (either the Gamma or Bivariate Normal model).

For given $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ peak heights across different markers m are independent. Let $\mathbf{N}$ be the random variable of allele counts, which is considered independent across markers. Then the total likelihood function can be decomposed as:

$$L_s(\mathrm{H}_j, \mathbf{N}, \boldsymbol{\phi}_s, \boldsymbol{\theta}|\mathrm{E}) = \prod_{m\in\{1,\ldots,M\}} L_{ms}(\mathrm{H}_j, \mathbf{N}_m, \boldsymbol{\phi}_s, \boldsymbol{\theta}_m|\mathrm{E}_{ms}). \tag{28}$$

With L_{ms} representing the likelihood function for marker m and sample s, E_{ms} the evidence (peak heights) for the same marker and sample and $\boldsymbol{\theta}_m$ the parameters specific to marker m. Although the parameters of the model are expected to differ between markers and even between alleles, there is no practical way to incorporate this in computations as this would involve maximizing the likelihood over an additional $M - 1$ variables for each parameter that is considered to be marker specific.

Then given $\boldsymbol{\theta}_m$ and $\boldsymbol{\phi}_s$, the likelihood L_m is only dependent on $\mathbf{N}_m$. Contributors randomly drawn from the population are assumed to have a known distribution. All other contributors have a deterministic distribution. Thus the likelihood for a given marker can be calculated by a weighted sum over all possible $\mathbf{n}_m$:

$$\begin{aligned} L_{ms}(\mathrm{H}_j, \mathbf{N}_m, \boldsymbol{\phi}_s, \boldsymbol{\theta}_m|\mathrm{E}_{ms}) &= \sum_{\mathbf{n}_m} L_{ms}(\mathrm{H}_j, \mathbf{n}_m, \boldsymbol{\phi}_s, \boldsymbol{\theta}_m|\mathrm{E}_{ms}) \cdot \mathbb{P}(\mathbf{n}_m) \\ &= \mathbb{E}_{\mathbf{N}_m}\left[L_{ms}(\mathrm{H}_j, \mathbf{N}_m, \boldsymbol{\phi}_s, \boldsymbol{\theta}_m|\mathrm{E}_{ms})\right]. \end{aligned} \tag{29}$$

In addition it is assumed an additional factorization is possible:

$$L_{ms}(\mathrm{H}_j, \mathbf{N}_m, \boldsymbol{\phi}_s, \boldsymbol{\theta}_m | \mathrm{E}_{ms}) = \prod_{a=1}^{A_m} L_{msa}(\mathrm{H}_j, \mathbf{N}_{ma}, \boldsymbol{\phi}_s, \boldsymbol{\theta}_m | \mathrm{E}_{msa}), \tag{30}$$

with $\mathbf{n}_{ma}$ some subset of $\mathbf{n}_m$ and similarly E_{msa} a subset of E_{ms}. The subsets $\mathbf{n}_{ma}$ and E_{msa} can differ depending on the model used. We will assume the population to be in Hardy-Weinberg equilibrium. This means that two alleles of an individual are chosen at random with allele-frequencies $(q_1, \ldots, q_{A_m})$. Thus $\mathbf{n}_m$ follows an independent multinomial distribution with allele frequencies $(q_1, \ldots, q_{A_m})$ and $\sum_a n_{iam} = 2$.

To deal with the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ maximum likelihood estimation is performed for the competing hypotheses H_0 and H_1 separately. Thus the likelihood ratio for sample s can be expressed as:

$$\mathrm{LR}_{\theta,s} = \frac{\sup\limits_{\boldsymbol{\theta},\boldsymbol{\phi}} L_s(\mathrm{H}_0, \mathbf{n}, \boldsymbol{\phi}_s, \boldsymbol{\theta} | \mathrm{E}_s)}{\sup\limits_{\boldsymbol{\theta},\boldsymbol{\phi}} L_s(\mathrm{H}_1, \mathbf{n}, \boldsymbol{\phi}_s, \boldsymbol{\theta} | \mathrm{E}_s)}. \tag{31}$$

5.2 Bayesian network

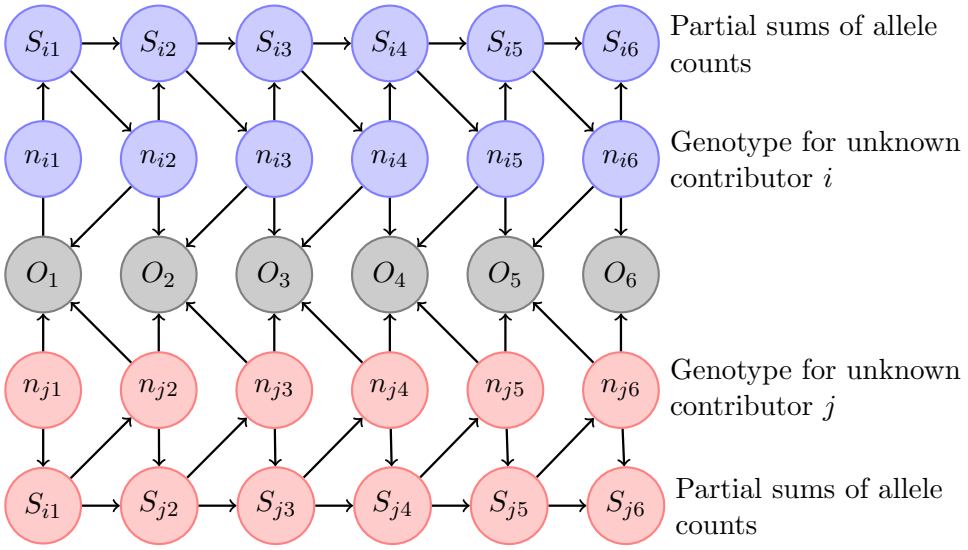

Figure 1: Bayesian network with n_{ia}-variables representing allele counts of random contributor i at allele a, S-variables a counter ensuring the allele counts are appropriately distribute and the O-variables are defined to incorporate peak height information.

To calculate the likelihood (equation 29) one quickly has to deal with a very large sum. This problem gets bigger the more random contributors there are in the calculation. To

deal with this more efficiently the method described in [6, Cowell et al. (2013)] will be utilized.

Using the properties of the multinomial distribution we can deduce the following. Without any information about allele counts $(n_{i2m}, \ldots, n_{iA_m m})$, n_{i1m} follows a binomial distribution $\text{Bin}(2, q_1)$. Introducing the counter $S_{iam} = \sum_{k \leq a} n_{ikm}$, it follows that given $S_{i(a-1)m}$, n_{iam} is distributed as:

$$n_{iam} \sim \begin{cases} \text{Bin}\left(2, \frac{q_a}{\sum_{k \geq a} q_k}\right) & \text{if } S_{i(a-1)m} = 0, \\ \text{Bin}\left(1, \frac{q_a}{\sum_{k \geq a} q_k}\right) & \text{if } S_{i(a-1)m} = 1, \\ 0 & \text{if } S_{i(a-1)m} = 2. \end{cases} \tag{32}$$

Adding these counters S allows the representation of allele counts in figure 1. Disregarding the O-variables for the moment, we will focus on the variables representing the genotypes; the n-variables. With the vector of variables $\mathbf{S}_i$ and $\mathbf{S}_j$ being the counter introduced above, the variables $(n_{i1}, \ldots, n_{i6})$ and $(n_{j1}, \ldots, n_{j6})$ follow an independent multinomial distribution with allele frequencies $(q_1, \ldots, q_6)$ and $\sum_{k=1}^{6} n_{i/jk} = 2$. This Markov representation of the genotypes allows for lower computation time than more straightforward representations. For a formal proof of the validity of this representation for the allele count distribution see [8, Graversen, T. (2013b)].

Peak height information can be incorporated by appropriately specifying the conditional distribution of the O-variables. We will define these variables as having two possible states, 0 and 1. Let k_m be a constant such that the equation below defines a probability distribution, then we will define:

$$\mathbb{P}(O_{ma} = 1 | \mathbf{N}_{ma} = \mathbf{n}_{ma}) = L_{msa}(H_j, \mathbf{n}_{ma}, \boldsymbol{\phi}_s, \boldsymbol{\theta} | \mathrm{E}_{msa}) / k_m. \tag{33}$$

Then this distribution is dependent on the occurrence or non-occurrence of dropout, as well as the model used (the Gamma model or the Bivariate normal model). The structure of the bayesian network is dependent on the model used as well. In both cases the O-variables will be conditionally independent of the S-variables given the n-variables. This is reflected in the network by the fact that for any i, j and k, the set of all n-variables d-separates S_{ij} from O_k. This is apparent since the set of n-variables separates the set of O-variables from the set of S-variables in the moralized graph. In addition it is important to note the mutual conditional independence of the O-variables given the rest of the network (the n- and S-variables), which agrees with the distribution defined in equation 33.

Then we can rewrite the likelihood function (equation 29):

$$\begin{aligned}
\mathbb{E}\left[\prod_{a=1}^{A_m} L_{ms}(\mathrm{H}_j, \mathbf{N}_{ma}, \boldsymbol{\phi}_s, \boldsymbol{\theta}_m | \mathrm{E}_{msa})\right] &= \mathbb{E}\left[\prod_{a=1}^{A_m} \mathbb{P}(O_{ma}=1|\mathbf{N}_{ma}) k_m\right] \\
&= \mathbb{E}\left[\prod_{a=1}^{A_m} \mathbb{P}(O_{ma}=1|\mathbf{N}_m)\right] \prod_{a=1}^{A_m} k_m \\
&\overset{2}{=} \mathbb{E}\left[\mathbb{P}\left(\bigcap_{a\in\{1,\ldots,A_m\}} \{O_{ma}=1\} \middle| \mathbf{N}_m\right)\right] \prod_{a=1}^{A_m} k_m \\
&= \mathbb{P}\left(\bigcap_{a\in\{1,\ldots,A_m\}} \{O_{ma}=1\}\right) \prod_{a=1}^{A_m} k_m.
\end{aligned} \tag{34}$$

Thus we can compute the likelihood function by setting all the O-variables in the network to 1 and then propagating this evidence.

Gamma Model

The stutter peak and allelic peak are assumed independent given the allele counts in this model. Thus the likelihood can be decomposed as in equation 30 and figure 1 gives the appropriate structure for the network. Due to stutter, the peak height distribution for peak a is dependent on the allele counts at allele a and $a+1$. Each O-variable has two states: 0 or 1, as described in equation 33. With $\Sigma = \rho(1-\xi)\sum_i \phi_i n_{ia} + \rho\xi \sum_i \phi_i n_{i,a+1}$, H_a the peak height at allele a and G and g the cumulative distribution function and the probability distribution function of the gamma distribution respectively, the conditional distribution of the O-variables is defined as:

$$\mathrm{P}(O_a = 1) = \begin{cases} k_m \cdot G(C; \Sigma, \eta) & \text{if } H_a < C, \\ k_m \cdot g(H_a, \Sigma, \eta) & \text{if } H_a \geq C, \end{cases} \tag{35a}$$

$$\mathrm{P}(O_a = 0) = \begin{cases} 1 - k_m \cdot G(C; \Sigma, \eta) & \text{if } H_a < C, \\ 1 - k_m \cdot g(H_a, \Sigma, \eta) & \text{if } H_a \geq C. \end{cases} \tag{35b}$$

Σ of course being dependent on the states of the n-variables a and $a+1$. k_m is a constant small enough that insures equation 35 defines a probability distribution.

One can then evaluate the likelihood function (equation 29) by setting the evidence to "1" to all the O-variables, and evaluating the "probability" of this evidence. Keeping in mind the the likelihood has been multiplied by factor k_m for each variable one has to correct this likelihood by multiplying with $\frac{1}{k_m^{A_m}}$.

(2) The auxiliary variables $\{O_{ma} : a \in \{1, \ldots, A_m\}\}$ being conditionally independent of each other given the allele counts $\mathbf{N}_m$.

Bivariate normal model

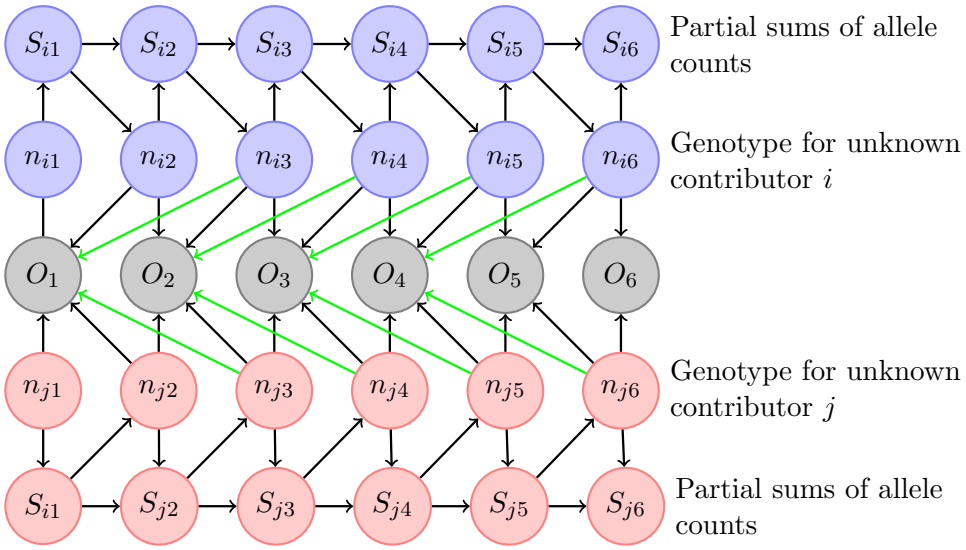

Figure 2: Bayesian network used for the Bivariate Normal model with n_{ia}-variables representing allele counts of random contributor i at allele a, S-variables a counter ensuring the allele counts are appropriately distribute and the O-variables are defined to incorporate peak height information. Black arrows are sufficient for the 0^{th}-order approximation while green arrows need to be added to be able to compute a 1^{st} order approximation.

For the bivariate normal model a network more complex then the example in figure 1 has to be used. Using the decomposition of the likelihood function defined in equations 14 and 15 we could define the O-variables as having the conditional probability:

$$\mathrm{P}(O_a = 1) = \begin{cases} k_m \cdot F_{\text{Normal}}\left(C; (2-\epsilon)m\sum_i \phi_i n_{ia} + \mathrm{E}(H^{s*}_{a+1}),\right. \\ \left.\sqrt{sd^2_{allelic}\sum_i \phi_i n_{ia} + \mathrm{Var}(H^{s*}_{a+1})}\right) & \text{if } H_a < C, \\ k_m \cdot f_{\text{Normal}}\left(H_a; (2-\epsilon)m\sum_i \phi_i n_{ia} + \mathrm{E}(H^{s*}_{a+1}),\right. \\ \left.\sqrt{sd^2_{allelic}\sum_i \phi_i n_{ia} + \mathrm{Var}(H^{s*}_{a+1})}\right) & \text{if } H_a \geq C, \end{cases} \tag{36a}$$

$$\mathrm{P}(O_a = 0) = \begin{cases} 1 - k_m \cdot F_{\text{Normal}}\left(C; (2-\epsilon)m\sum_i \phi_i n_{ia} + \mathrm{E}(H^{s*}_{a+1}),\right. \\ \left.\sqrt{sd^2_{allelic}\sum_i \phi_i n_{ia} + \mathrm{Var}(H^{s*}_{a+1})}\right) & \text{if } H_a < C, \\ 1 - k_m \cdot f_{\text{Normal}}\left(H_a; (2-\epsilon)m\sum_i \phi_i n_{ia} + \mathrm{E}(H^{s*}_{a+1}),\right. \\ \left.\sqrt{sd^2_{allelic}\sum_i \phi_i n_{ia} + \mathrm{Var}(H^{s*}_{a+1})}\right) & \text{if } H_a \geq C. \end{cases} \tag{36b}$$

With $F_{\text{Normal}}(x, m, sd)$ and $f_{\text{Normal}}(x, m, sd)$ being the cumulative distribution function and the probability distribution function respectively of a normal random variable with mean m and standard deviation sd. However, due to the iterative definition of H^{s*}_{a+1} the variable O_a would be directly dependent on the set of variables $\{n_{ik} : k \geq a, i \in \{1, \ldots, I\}\}$. Generally this would lead to highly complex bayesian networks with large cliques and thus large computation times. To combat this we introduce the following approximation:

$\mathbf{m^{th}}$ order approximation H^{0*}_{a+1+m} is assumed to be equal to H^{0}_{a+1+m}.

Thus the order of the approximation controls how many step there are in the iterative calculation like in equation 15. With order 0 defining an uncorrelated version of the normal random variables, the structure of the bayesian network used for the Gamma model (figure 1) would suffice. Using equation 36 with the approximation made above we find that the network for the m^{th} order approximation would need "arrows" from the set $\{n_{ik} : a \leq k \leq a + 1 + m, i \in \{1, \ldots, I\}\}$ to the variable O_a.

5.3 Marker dependence

In general the parameters of the model will differ between markers and even between alleles. However,it would be impractical to maximize the likelihood function over different parameters for all markers. It will be assumed that the parameters are constant across alleles. From section 3 we see that differences in parameters of the Gamma/Bivariate Normal model between markers arise from differences between the parameters $\pi_{extraction}$, $\pi_{aliquot}$, $\pi_{stutter}$, π_{PCReff} and the conversion factor arising from the visualization of the alleles after electrophoresis. Since any estimation of stutter percentage directly from the data is directly dependent on additional assumptions of the contributors to the sample we will assume the stutter parameter to be equal across all markers (and alleles) for both the Gamma and Bivariate Normal model. Luckily the mean peak height can reasonably be estimated from the data across markers. This is proportional to the parameter ρ and m for the Gamma model and the Bivariate Normal model respectively. For each marker this parameter will be scaled by a factor (x_m^{scale}) as follows:

$$x_m^{scale} = \frac{M \cdot \sum_{s=1}^{S} \sum_{a=1}^{A_m} H_{ams}}{\sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{a=1}^{A_m} H_{ams}}, \tag{37}$$

with H_{ams} the observed peak height at allele a of marker m and sample s. Note that only observed peaks will be used in the sum, there may be peaks below the threshold C that will not be counted. However, these contributions should be small enough to make little difference. Next should be considered how the variance of the peaks scale with x_m^{scale}. Looking at the simulation method for DNA peak distributions (section 3) we find that differences in peak height can arise in multiple ways. Variations in $\pi_{extraction}$, $\pi_{aliquot}$, π_{PCReff} and the conversion factor arising from the visualization of the alleles after electrophoresis all effect the mean peak height. In particular π_{PCReff} can greatly

vary the mean peak height even for small changes in its value. It is assumed the standard deviation of the peak height distribution scales with the mean peak height. To be precise:

$$\text{Gamma model} \begin{cases} \rho_m = \rho \\ \eta_m = \eta \cdot x_m^{scale} \\ \xi_m = \xi, \end{cases} \tag{38a}$$

$$\text{Bivariate Normal model} \begin{cases} m_m & = m \cdot x_m^{scale} \\ sd_{allele,m} & = sd_{allele} \cdot x_m^{scale} \\ sd_{stutter,m} & = sd_{stutter} \cdot x_m^{scale} \\ \xi_m & = \xi \\ r_m & = r. \end{cases} \tag{38b}$$

Figure 27 shows relatively constant shape parameter (proportional to ρ) with varying scale parameter (proportional to η) in the case of variable amplification parameter (π_{PCReff}). It also shows a more than linear increase in the scale parameter as a function of π_{PCReff}. Note that this is not the case with varying the pre-amplification parameter ($\pi_{extraction,aliquot}$) as seen in figure 25.

The parameters of the fitted bivariate normal distribution as a function of the amplification parameter are shown in figure 30. The fraction of the mean and standard deviation for both the stutter and allelic peak is shown to be roughly constant in the bottom right graph. Also note the constant correlation coefficient in the upper right graph. Again a variable pre-amplification parameters shows a different picture (figure 28).

Luckily the conversion factor arising from the visualization of the alleles after electrophoresis is a deterministic step. If the peak heights are scaled by a factor k then obviously parameters m and ρ are scaled by this factor but also the standard deviation of the peak heights are scaled by this factor and the correlation coefficient remains the same. Thus equation 38 is valid for this step as well.

5.4 Posterior distribution of genotypes

For various reasons it can be desirable to know the posterior distribution of genotypes given the observed peak heights or a subset of observed peak heights. Let this (sub)set be denoted by E'$_{ms}$. If the likelihood associated with this (sub)set of peak heights can be factorized as follows:

$$\mathbb{P}(\mathrm{E}'_{ms}|\mathbf{N}_m = \mathbf{n}_m) = L'_{ms}(\mathbf{n}_m|\mathrm{E}'_{ms}) = \prod_{a \in A'} L_{msa}(\mathbf{n}_{ma}|\mathrm{E}_{msa}), \tag{39}$$

then conditioning on the auxiliary variables being equal to 1 is equivalent to conditioning on the observed peak heights E'$_{ms}$:

$$\begin{aligned}
\mathbb{P}\left(\mathbf{N}_m = \mathbf{n}_m \middle| \bigcap_{a\in A'}\{O_{ma}=1\}\right) &= \frac{1}{\mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\}\right)}\mathbb{P}\left(\mathbf{N}_m = \mathbf{n}_m, \bigcap_{a\in A'}\{O_{ma}=1\}\right) \\
&= \frac{\mathbb{P}(\mathbf{N}_m = \mathbf{n}_m)}{\mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\}\right)} \cdot \mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\} \middle| \mathbf{N}_m = \mathbf{n}_m\right) \\
&= \frac{\mathbb{P}(\mathbf{N}_m = \mathbf{n}_m)}{\mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\}\right)} \cdot \prod_{a\in A'} \mathbb{P}\left(O_{ma}=1 | \mathbf{N}_m = \mathbf{n}_m\right) \\
&= \frac{\mathbb{P}(\mathbf{N}_m = \mathbf{n}_m)}{\mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\}\right)} \cdot \prod_{a\in A'} L_{msa}(\mathbf{n}_{ma}|\mathrm{E}_{msa})/k_m \\
&= \frac{\mathbb{P}(\mathbf{N}_m = \mathbf{n}_m)}{\mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\}\right)} \cdot L'_{ms}(\mathbf{n}_m|\mathrm{E}'_{ms}) \prod_{a\in A'} \frac{1}{k_m} \\
&= \frac{\mathbb{P}(\mathbf{N}_m = \mathbf{n}_m)}{\mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\}\right)} \cdot \mathbb{P}(\mathrm{E}'_{ms}|\mathbf{N}_m = \mathbf{n}_m) \cdot \prod_{a\in A'} \frac{1}{k_m} \\
&= \frac{1}{\mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\}\right)} \cdot \mathbb{P}(\mathrm{E}'_{ms}, \mathbf{N}_m = \mathbf{n}_m) \cdot \prod_{a\in A'} \frac{1}{k_m} \\
&= \frac{\mathbb{P}(\mathrm{E}'_{ms})}{\mathbb{P}\left(\bigcap_{a\in A'}\{O_{ma}=1\}\right)} \cdot \mathbb{P}(\mathbf{N}_m = \mathbf{n}_m|\mathrm{E}'_{ms}) \cdot \prod_{a\in A'} \frac{1}{k_m} \\
&= \mathbb{P}(\mathbf{N}_m = \mathbf{n}_m|\mathrm{E}'_{ms}).
\end{aligned} \tag{40}$$

Dependence on H$_j$, $\boldsymbol{\phi}_s$ and $\boldsymbol{\theta}$ was suppressed in the above notation.

As previously discussed equation 39 is valid in the case that E'$_{ms}$ is equal to E$_{ms}$. In addition a factorization is possible for both the Gamma and Bivariate Normal model for arbitrary E'$_{ms}$.

6 Results

6.1 Validation

Validation of the Gamma model without artefacts was previously discussed in [7, Cowell(2009)]. The following assumptions made by the Gamma model were investigated:

- The peak areas follow gamma distributions.
- The scale parameter is independent of the amount of DNA in the sample.
- The mean peak height (or equivalently the shape parameter when taking into account the above assumption) is proportional to the amount of DNA in the sample.

To this effect the simulation model described in section 3 was used to validate these assumptions (section 4.1.1). The Gamma model with artefacts makes some additional assumptions that need to be validated:

- Both allelic peak heights (arising from contributions to that allelic type) and stutter peak heights follow gamma distributions.
- The scale parameter is independent of the amount of DNA in the sample for both the stutter and allelic peak heights.
- The mean peak height is proportional to the amount of DNA in the sample for both the stutter and allelic peak heights.
- The allelic peak height and stutter peak height distribution have the same value for the scale parameter.
- Stutter and allelic peak heights are independent.

Furthermore the dropout rates of the Gamma model should coincide with those of the simulation model. The bivariate normal model makes the following assumptions that need to be validated:

- Stutter and allelic peak (H_a^0, H_a^s) follow a bivariate normal distribution.
- The correlation coefficient of this distribution is independent of the starting amount of DNA.
- The mean parameters of both H_a^0 and H_a^s increases linearly with the starting amount of DNA.
- The variance of both H_a^0 and H_a^s increases linearly with the starting amount of DNA.
- The conditional distribution $H_a^s | H_a^0 + H_{a+1}^s < C$ closely resembles a normal distribution with parameters $\left(\mathbb{E}(H_a^s | H_a^0 + H_{a+1}^s < C), \sqrt{\mathrm{Var}(H_a^s | H_a^0 + H_{a+1}^s < C)}\right)$.

This in addition to correct dropout rates.

For all simulations in this section simulation model (section 3) was used with values for $\pi_{extraction}$, $\pi_{aliquot}$, $\pi_{stutter}$ and π_{PCReff} being those reported in [3, Gill et al.(2005)].

6.1.1 Gamma model

Goodness of Fit

To test if the gamma distribution is a good fit for DNA peak heights a set of Q-Q (Quantile-Quantile) plots were made. For three different starting values of DNA molecules of a certain allelic type $10,000$ simulations were run, using the method described in section 3. DNA was amplified over 28 cycles. Maximum likelihood estimation

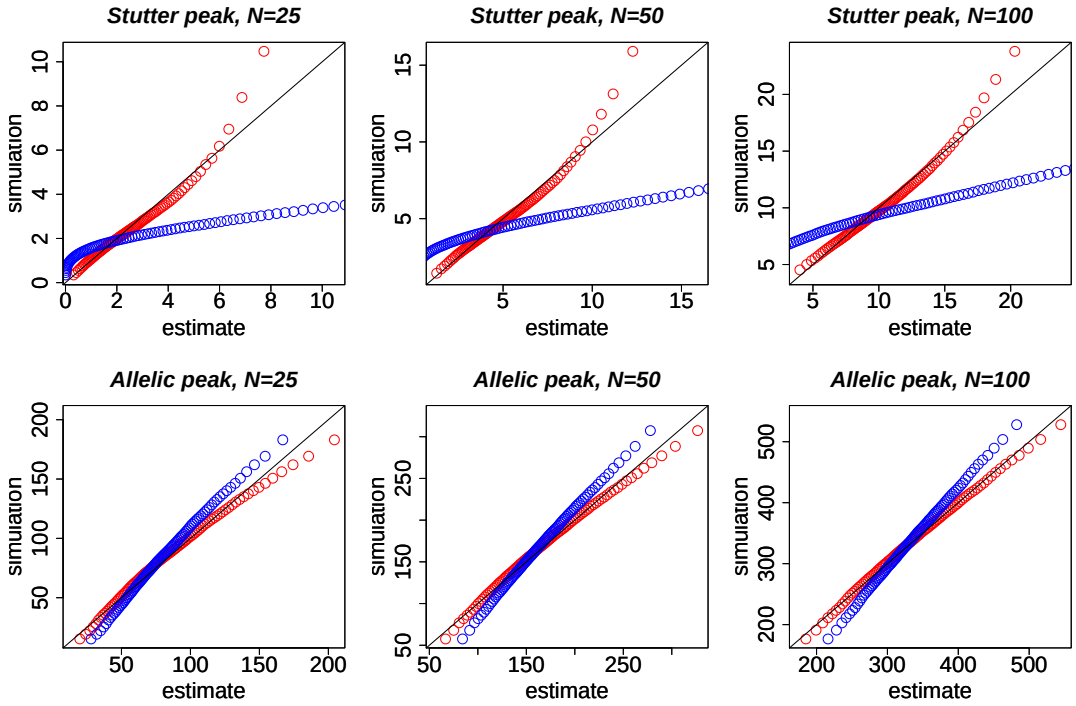

Figure 3: Quantile-quantile plots using simulations of peak height for three different starting values N of the number of an alleles. Red shows the simulation vs the fitted gamma distribution with different scale parameters, blue with equal scale parameters.

was used to estimate the parameters of the Gamma distribution. Figure 3 shows the Q-Q plots of the simulated data vs the quantiles of the fitted Gamma distribution. The red dots show the Q-Q plot when the scale parameters of the stutter and allelic peak are not assumed to be equal. The blue dots represent the assumption of equal scale parameters of the stutter and allelic peak. This clearly illustrates a significant deterioration of the fit under these more restrictive assumptions. In general the allelic peak closely resembles a gamma distribution. The same can be said of the stutter peak though to a smaller degree. This should come as no surprise as the stutter peak is generated differently from a peak without stutter. The fit also appears to improve with increasing amount of starting values of DNA molecules. Figure 4 show the histograms of the simulation as well as the estimated gamma distributions used in figure 3.

Parameters as a function of N

It is not hard to verify analytically that for the simulation model the mean peak height is proportional to the initial number of DNA molecules N. Similarly the parameters of the Gamma distribution were estimated for different values of N and the shape and

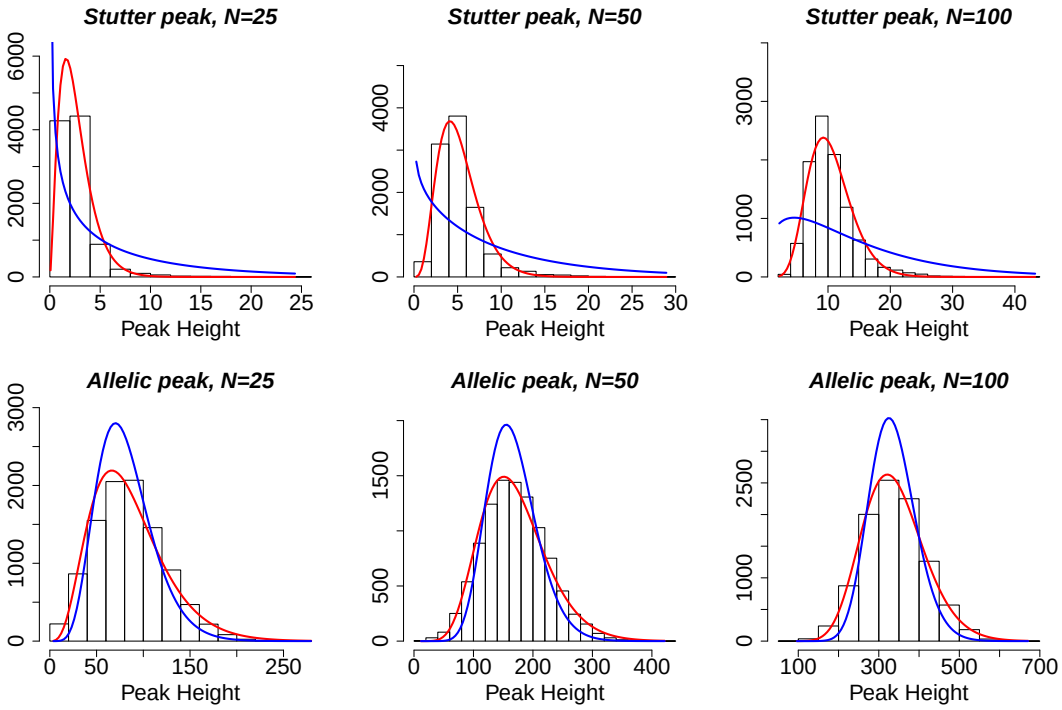

Figure 4: Histograms using simulations of peak height for three different starting values N of the number of an alleles. Red shows the the fitted gamma distribution with different scale parameters, blue with equal scale parameters.

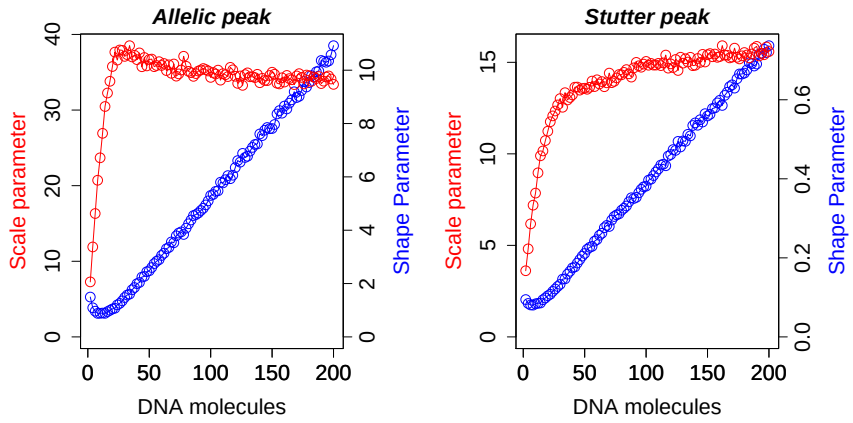

Figure 5: Estimated shape (blue) and scale (red) parameters vs the starting amount of DNA molecules N for the the stutter and allelic peak.

scale parameters are shown as a function of N in figure 5 for both the stutter and allelic peak. The assumption that the scale parameter is independent of the amount of DNA in the sample appears roughly valid for $N > 25$, with a slight slope visible for both the stutter and allelic peak. It is important to note that the parameters for the allelic peak and the stutter peak were fitted independently. A large difference in scale parameter between the stutter and allelic peaks for equal values of N is clearly visible.

Stutter and allelic peak dependence

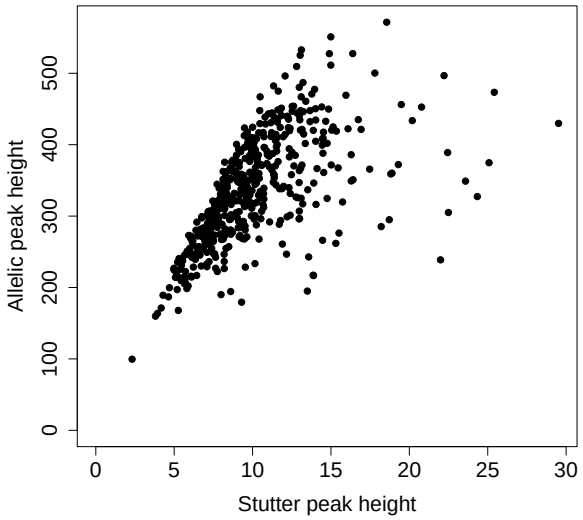

Figure 6: A sample of 500 simulations of the stutter peak height and allelic peak height.

To test the assumption that the stutter and allelic peak heights are independent an obvious first step is to see what a two-dimensional plot of the stutter peak height versus the allelic peak height looks like. This is shown in figure 6. Already it is apparent that the allelic and stutter peak heights cannot be considered independent. Low/high allelic peak height seems to correlate with low/high stutter peak height respectively. To further test the independence assumption a sample of 10.000 simulations was divided into bins (see table 1). By doing this some information is lost, but the option to use Pearson's Chi-squared test for independence is gained. Using the R function chisq.test() a p-value lower than $2.2e-16$ was found. Thus is seems safe to conclude the stutter and allelic peak height are *not* independent.

	[91, 149]	(149, 206]	(206, 264]	(264, 321]	(321, 379]	(379, 437]	(437, 494]	(494, 552]	(552, 609]	(609, 667]
[2.18, 5.95]	25	202	318	13	0	0	0	0	0	0
(5.95, 9.72]	3	63	768	1857	1366	233	0	0	0	0
(9.72, 13.5]	0	18	112	463	1150	1225	441	69	3	0
(13.5, 17.3]	0	6	45	128	246	286	233	101	27	2
(17.3, 21]	0	0	17	60	91	80	73	32	15	2
(21, 24.8]	0	1	7	25	43	41	31	9	1	1
(24.8, 28.6]	0	1	1	3	14	17	9	5	1	0
(28.6, 32.3]	0	0	0	0	3	7	2	0	0	0
(32.3, 36.1]	0	0	0	0	1	0	0	0	0	0
(36.1, 39.9]	0	0	0	0	1	0	1	2	0	0

Table 1: A sample of 10000 simulations of the stutter peak height and allelic peak height distributed into bins.

Dropout

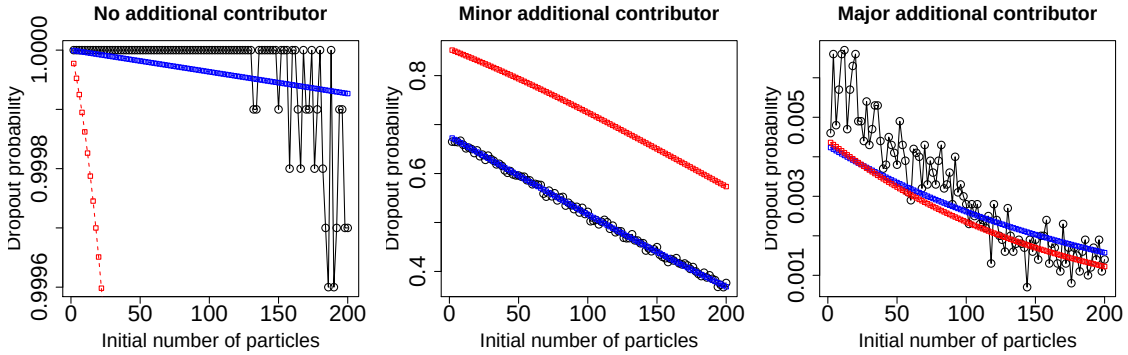

Figure 7: Dropout rates of the stutter peak as a function of the starting amount of DNA N. Black are the dropout estimates from the data, red and blue from maximum likelihood estimation of the Gamma model with red having $\rho \cdot \eta$ and ξ estimated directly from the data.

Dropout rates were tested in three different situations:

- **No additional contributor** A contributor with an initial number of DNA molecules of a single allele varying over $0 - 200$ with no other contributors.
- **Minor additional contributor** A contributor with an initial number of DNA molecules of a single allele varying over $0 - 200$ and a small contributor one repeat number lower of 12 DNA molecules.
- **Major additional contributor** A contributor with an initial number of DNA molecules of a single allele varying over $0 - 200$ and a large contributor one repeat number lower of 50 DNA molecules.

These correspond to situations with very high probability of dropout, varying probability of dropout and very low probability of dropout for the stutter peak. $10,000$ simulations were run for all all initial number of DNA molecules N of the variable contributor in the set $N \in \{2, 4, \ldots, 198, 200\}$. The threshold C was set to 50. From this dropout rates were estimated for the allelic and stutter peak. These were compared to the theoretical dropout rates of the Gamma model. Parameters for the Gamma model were estimated by maximum likelihood, only taking dropout rates into account. Information like peak heights was not included in the maximum likelihood estimation. Let G be the cumulative distribution function of the Gamma distribution, n the number of DNA molecules of the additional contributor (either 0,12, or 50), $D_{allele}(N)$ the number of times the allelic peak dropped out during simulation out of a possible $10,000$ times and $D_{stutter}(N)$ the same for the stutter peak. Then the following is maximized over ρ, ξ and η:

$$\prod_{N \in \{2,4,\ldots,198,200\}} L(N), \tag{41}$$

with

$$\begin{aligned} L(N) =& \mathrm{G}(50; \rho\xi N + \rho(1-\xi)n), \eta)^{10,000 - D_{stutter}(N)} \cdot \mathrm{G}(50; \rho(1-\xi)N, \eta)^{10,000 - D_{allele}(N)} \cdot \\ & \cdot (1 - \mathrm{G}(50; \rho\xi N + \rho(1-\xi)n), \eta))^{D_{stutter}(N)} \cdot (1 - \mathrm{G}(50; \rho(1-\xi)N, \eta))^{D_{allele}(N)}. \end{aligned} \tag{42}$$

For comparison the likelihood was maximized with the mean peak height contribution of a single DNA molecule ($\rho \cdot \eta$) and stutter (ξ) directly estimated from the data. Figure 7 shows estimated dropout rates from the data (black) compared to the dropout rates from maximum likelihood estimation of the Gamma model. Estimation with variable ξ, ρ and η (blue) shows dropout rates closely resembling those estimated from the data. However, the stutter parameter varies widely for the three different situation. In a DNA profile a wide range of scenario's likely occurs. Estimating $\rho \cdot \eta$ and ξ from the data thus represents a more realistic approach, as this forces some uniformity across the three different situations described above. Unfortunately since the scale parameters are assumed equal for the stutter and allelic peak it is not clear how to estimate the remaining parameter. Thus there is one remaining degree of freedom to optimize over. This situation (red) is also plotted. Clearly the dropout rates are less similar to the estimated dropout rates. Allelic dropout rates are very close to each other for all three situations such that they are hard to distinguish graphically and are not plotted here, but can be found in the section **Additional Figures**.

Estimated parameter values are shown in table 2. In the case of three degrees of freedom the stutter percentage varies widely, while in the case of one degree of freedom the remaining parameter that controls the variance of the peak heights varies significantly. Realistically this should be constant for these different scenarios and would result in worse fits for the dropout rates.

	Maximum likelihood estimation			Direct estimation		
Parameter	No contributor	Minor contributor	Major contributor	No contributor	Minor contributor	Major contributor
ρ	0.15	.15	.14	.36	.21	.15
η	23.4	24.5	25.8	9.7	16.3	23.1
ξ	$6.0e-05$	$3.6e-02$	$2.5e-02$	$3.1e-02$	$3.1e-02$	$3.1e-02$

Table 2: Estimated parameters of the gamma model from dropout rates with "Maximum likelihood estimation" denoting the situation in which all 3 parameters are estimated by maximizing the likelihood, whereas "Direct estimation" denotes the situation in which two parameters are estimated directly from the data.

6.1.2 Bivariate Normal model

Goodness of Fit

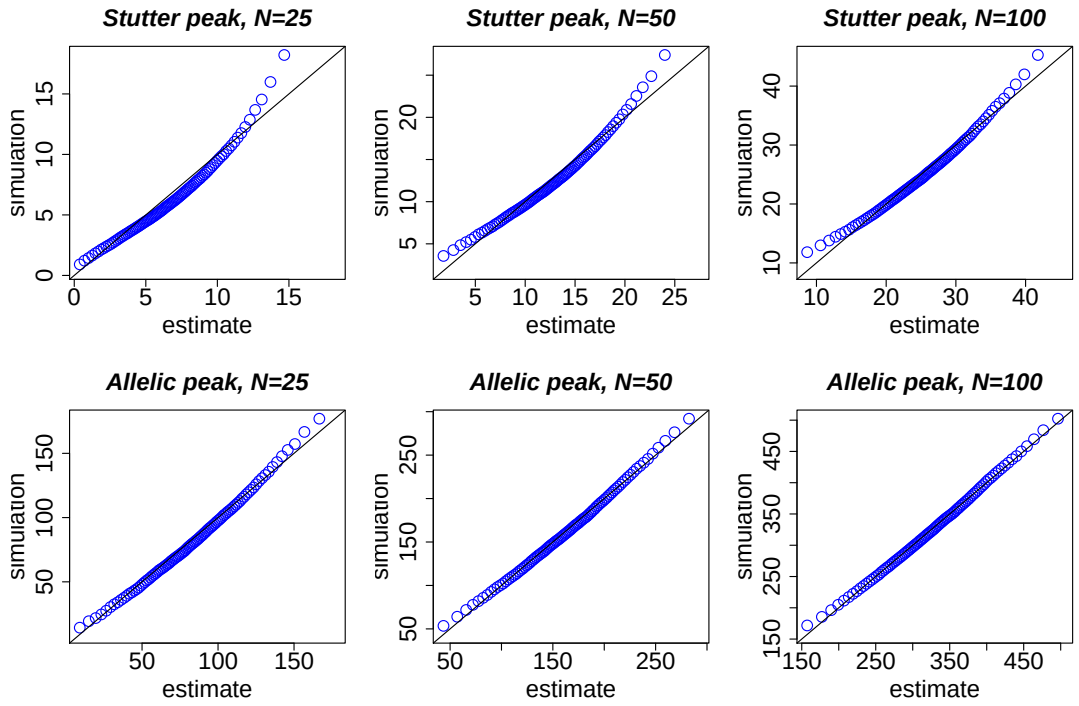

Figure 8: Quantile-quantile plots of simulation data of peak height for three different starting values N of the number of an alleles vs the fitted normal distribution

Similarly to section 6.1.1, Q-Q plots were made for different starting values of N using samples of $10,000$ simulations to evaluate the goodness of fit of the bivariate normal distribution for the stutter and allelic peak individually. Again the DNA was amplified over 28 cycles. This is shown in figure 8. The normal distribution appears to fit excellently with the allelic peak height distribution, especially for higher starting amounts of DNA N. A slight curve is evident in the QQ-plot of the stutter peaks, more

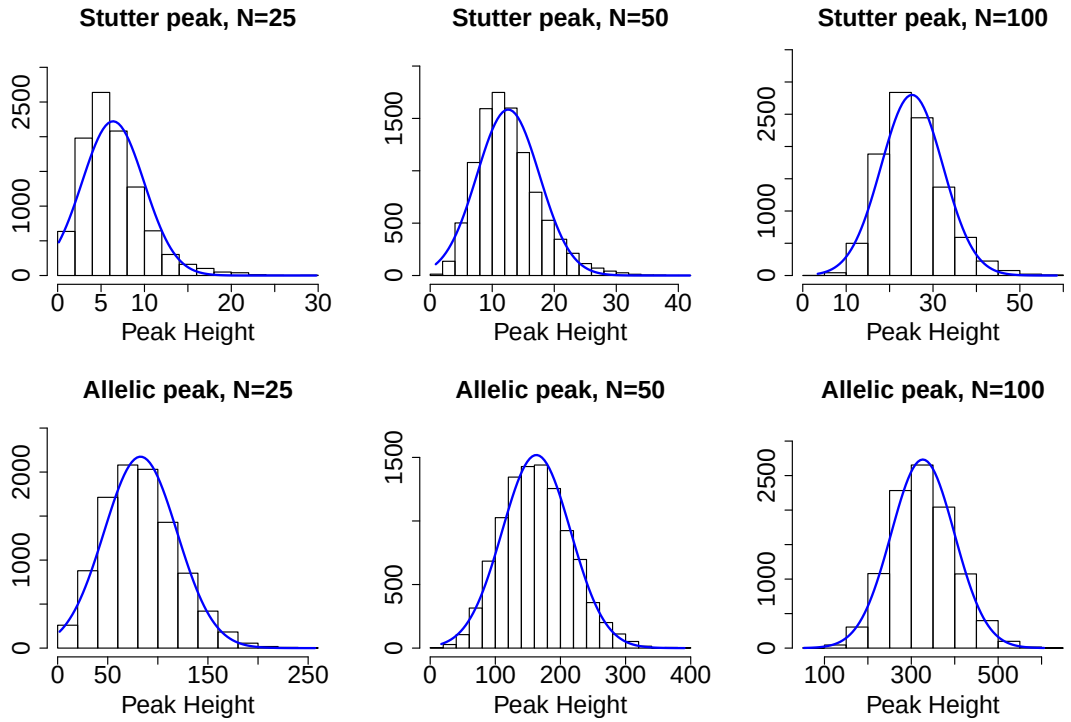

Figure 9: Histograms using simulations of peak height for three different starting values N of the number of an alleles with the fitted normal distribution plotted in blue.

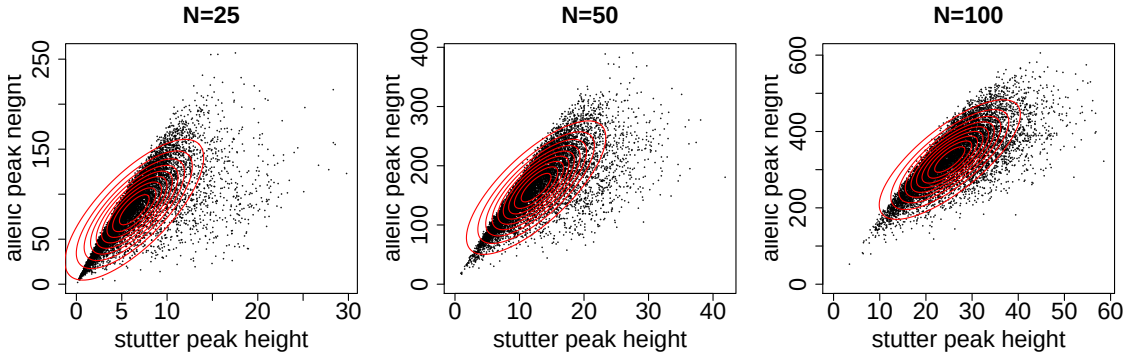

Figure 10: Combined plot of stutter and allelic peak height with simulations using three different starting values N of the number of an alleles with confidence ellipses of the fitted bivariate normal distribution.

so for lower values of N. The same simulations with accompanying estimates are plotted in figure 9 in the form of histograms with the appropriately scaled density function of

the estimate.

Two dimensional plots are shown in figure 10. Each black dot represents 1 simulation outcome, i.e. the allelic and stutter peak height. Confidence ellipses of the estimated bivariate normal distribution are plotted in red for values of 10% to 90% with steps of 10%. Visually it is clearly not an exact fit, though it does tend to improve with increasing number of starting DNA N.

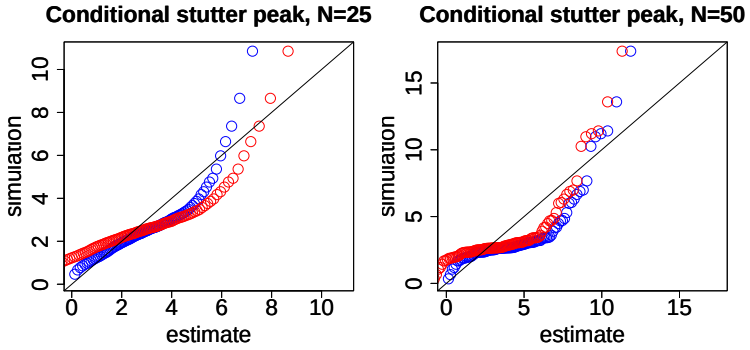

Figure 11: Q-Q plots of simulation data of stutter peak height conditional on the allelic peak dropping out for two different starting values N of the number of an alleles vs the (fitted) normal distribution. Blue corresponds to a normal distribution with estimated parameters by maximum likelihood, while red corresponds to a normal distribution with parameters as defined in equation 20.

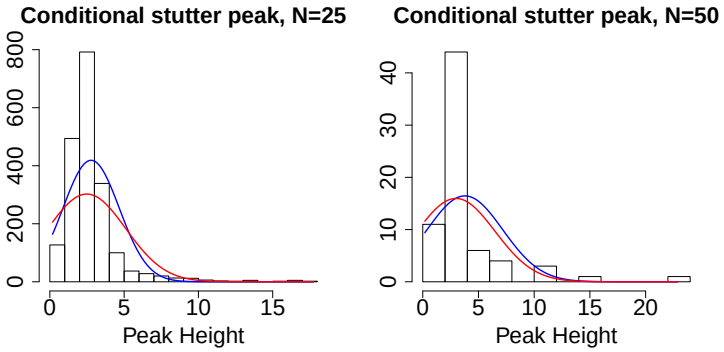

Figure 12: Histograms using simulation data of stutter peak height conditional on the allelic peak dropping out for two different starting values N of the number of an alleles with the (fitted) normal distribution plotted. Blue corresponds to a normal distribution with estimated parameters by maximum likelihood, while red corresponds to a normal distribution with parameters as defined in equation 20.

In addition it is assumed the stutter peak height given the allelic peak dropping out follows a normal distribution as in equation 20. Figure 11 shows Q-Q plots for the quantiles from the simulation vs the quantiles of the normal distribution. Blue corresponds to a normal distribution with estimated parameters by maximum likelihood, while red corresponds to a normal distribution with parameters as defined in equation 20. Given the fact that these stutter peaks are only responsible for very small contributions, the goodness of fit should be sufficient. Figure 12 shows the accompanying histograms with probability density functions.

Parameters as a function of N

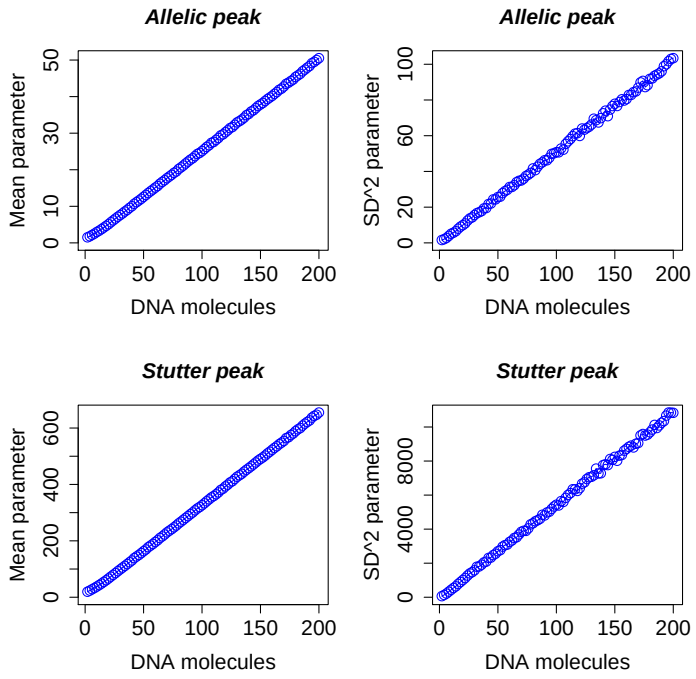

Figure 13: Estimated mean and variance of the bivariate normal distribution for both the stutter and allelic peak as a function of the starting amount of DNA N.

The mean and variance of the bivariate normal distribution for both the stutter and allelic peak have been assumed to be proportional to the starting amount of DNA N. Figure 13 shows as nice a linear relationship for mean and variance of both the stutter and allelic peak heights as could be hoped for. In addition figure 14 shows a roughly constant correlation parameter for $N > 25$. Figure 15 shows estimates of the mean and standard deviation parameter of the normal distribution for the stutter peak when the allelic peak has dropped out as a function of N. Direct maximum likelihood estimates from the data (blue) are compared to estimates from data using the parameters in equation 20 (red). For larger N there is an increased probability of lower dropout frequencies. If there is

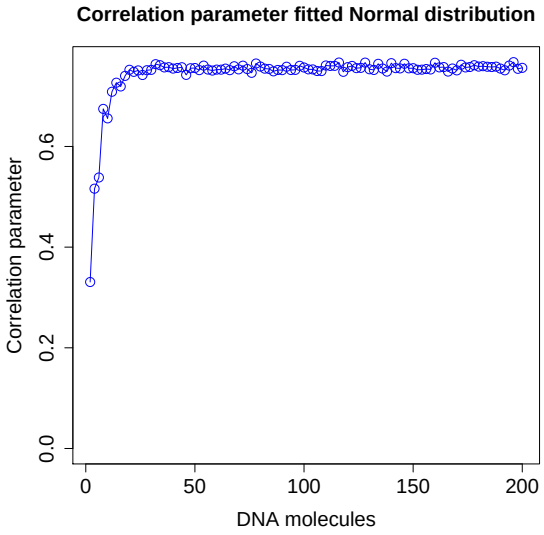

Figure 14: Estimated correlation parameter of the bivariate normal distribution as a function of the starting amount of DNA N.

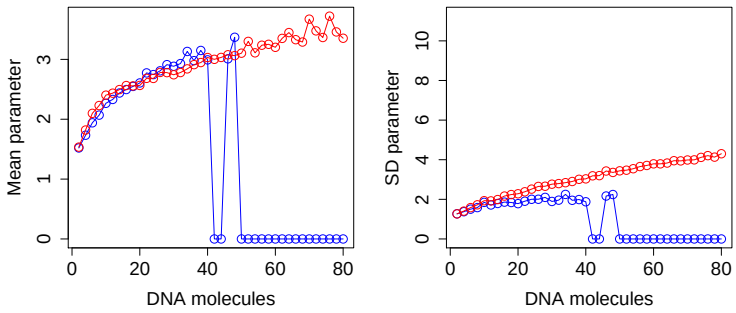

Figure 15: Estimated mean and standard deviation parameters of the normal distribution of the stutter peak conditional on the allelic peak dropping out. Points in red show estimates from data using the parameters in equation 20, while blue show maximum likelihood estimations using data directly.

no dropout the parameters are set to 0, which can be seen in the figure for $N > 40$.

Dropout

As in section 6.1.1 dropout rates were tested in three different situations:

- **No additional contributor** A contributor with an initial number of DNA molecules of a single allele varying over $0 - 200$ with no other contributors.

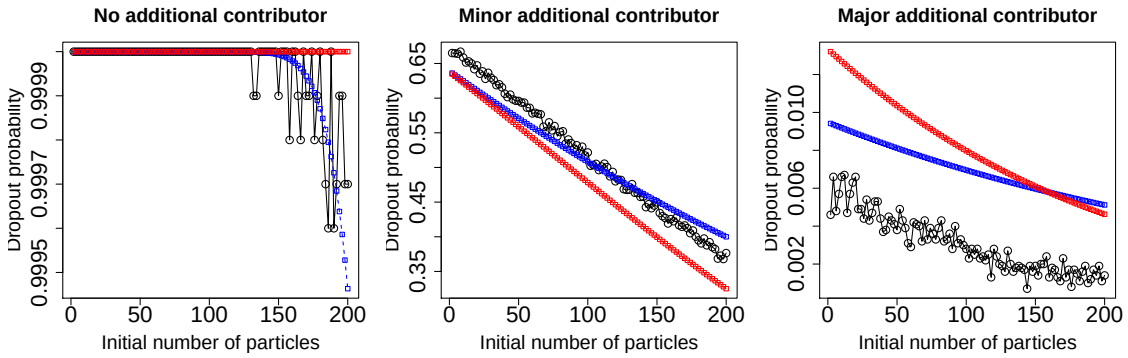

Figure 16: Dropout rates of the stutter peak as a function of the starting amount of DNA N. Black are the dropout estimates from the data, blue from maximum likelihood estimation of the Bivariate Normal model and red with parameters estimated from the data.

- **Minor additional contributor** A contributor with an initial number of DNA molecules of a single allele varying over $0 - 200$ and a small contributor one repeat number lower of 12 DNA molecules.
- **Major additional contributor** A contributor with an initial number of DNA molecules of a single allele varying over $0 - 200$ and a large contributor one repeat number lower of 50 DNA molecules.

$10,000$ simulations were run for all all initial number of DNA molecules N of the variable contributor in the set $N \in \{2, 4, \ldots, 198, 200\}$. The threshold C was set to 50. From this dropout rates were estimated for the allelic and stutter peak. These were compared to the theoretical dropout rates of the Bivariate Normal model. Parameters were estimated by maximum likelihood, only taking dropout rates into account. Peak height information was not incorporated into the estimation except to calculate the conditional distribution of the stutter peak given the allelic peak height. Let N be the cumulative distribution function of the Normal distribution, (H_i^0, H_i^s) the bivariate normal random variable associated with simulation $i \in \{1, \ldots, 10000\}$ and $(H_i^s | H_i^0)$ be as defined in section 4.2. As before (equation 41) maximum likelihood estimation was performed now

over variables r, m, ξ, $sd_{stutter}$ and sd_{allele}, and:

$$
\begin{aligned}
L(N) = & \prod_{\{H_i^0 \leq 50\} \cup \{H_i^s \leq 50\}} \mathrm{N}\left(50; m(1-\xi)N, \sqrt{N} sd_{allele}\right) \cdot \mathbb{P}\left(H_i^s \leq 50 | H_i^0 \leq 50\right) \cdot \\
& \prod_{\{H_i^0 \leq 50\} \cup \{H_i^s > 50\}} \mathrm{N}\left(50; m(1-\xi)N, \sqrt{N} sd_{allele}\right) \cdot \left(1 - \mathbb{P}\left(H_i^s \leq 50 | H_i^0 \leq 50\right)\right) \cdot \\
& \prod_{\{H_i^0 > 50\} \cup \{H_i^s \leq 50\}} \left(1 - \mathrm{N}\left(50; m(1-\xi)N, \sqrt{N} sd_{allele}\right)\right) \cdot \mathbb{P}\left(H_a^i \leq 50 | H_i^0 = h_i^0\right) \cdot \\
& \prod_{\{H_i^0 > 50\} \cup \{H_i^s > 50\}} \left(1 - \mathrm{N}\left(50; m(1-\xi)N, \sqrt{N} sd_{allele}\right)\right) \cdot \left(1 - \mathbb{P}\left(H_i^s \leq 50 | H_i^0 = h_i^0\right)\right) \\
& ,
\end{aligned}
\tag{43}
$$

In calculating $(H_i^s | H_i^0)$ care must be taken to incorporate the number of initial DNA molecules of the stutter peak n. For comparison the likelihood was maximized with all parameters directly estimated from the data. Figure 16 shows estimated dropout rates from the data (black) compared to the dropout rates from maximum likelihood estimation of the Bivariate Normal model (blue) and dropout rates from the Bivariate Normal model with parameters estimated directly (red) for the stutter peak. As expected the MLE Bivariate Normal model appears to fit better than the Bivariate Normal model with parameters estimated directly. Even so both models fit reasonably well with large differences in likelihood only occurring if the generated data was unlikely in the first place. E.g. if in the case of no additional contributors dropout occurs there is a large difference in likelihood noticeable between the red and blue "lines". Dropout rates for the allelic peak fit very well, just like with the gamma model, and can be found in the section **Additional Figures**.

	Maximum likelihood estimation			Estimated parameters
Parameter	No contributor	Minor contributor	Major contributor	
r	.30	$8.0e-02$	$4.2e-02$	.59
m	3.49	3.48	3.50	3.50
ξ	$2.8e-02$	$2.6e-02$	$2.8e-02$	$3.1e-02$
$sd_{stutter}$	.67	1.30	1.40	.38
sd_{allele}	7.69	7.56	7.21	7.63

Table 3: Estimated parameters of the bivariate normal model from dropout rates with "Maximum likelihood estimation" denoting the situation in which all parameters are estimated by maximizing the likelihood, whereas "Direct estimation" denotes the situation in which the parameters are estimated directly from the data.

Estimated parameter values are shown in table 3. Maximum likelihood estimation return relatively constant values m, ξ and sd_{allele} while r and $sd_{stutter}$ show more varia-

tion. Perhaps unsurprisingly direct estimation of these first 3 parameters closely resemble those estimated by maximum likelihood, in contrast to the latter two.

6.2 Example Case

To examine the performance of the models in practice we will test the above methodology on a case introduced in [9, Gill et al.(2008)] and also analyzed in [9, Cowell et al.(2011)] and [9, Cowell et al.(2013)]. The case concerns an incident in which the deceased had spent the evening with a group of friends. An altercation in the car park between the deceased and several others resulted in the death of the victim. The alleged offenders were then observed going into a public house to clean themselves in the lavatory. Subsequently two blood stains, MC18 and MC15, were found and were subjected to DNA analysis. This resulted in two DNA profiles indicating multiple contributors of at least three persons. The genotype of the victim, a suspect and an additional individual who likely contributed to one or more of the blood stains is known.

6.2.1 Estimated likelihoods

Both blood samples will be considered separately testing both the Gamma model and the Bivariate Normal model with 0^{th}, 1^{st} and 2^{nd} order approximation. For comparison the situation in which the suspect is present will be compared to the presence of a random individual from the population. In both cases there will be an additional random individual to account for any unknown sources of DNA. For ease of comparison between the Gamma and Bivariate Normal model the parameters $m = \rho\eta$, $sd_{allele} = \sqrt{\rho(1-\xi)}\eta$ and $sd_{stutter} = \sqrt{\rho\xi}\eta$ will be reported in stead of ρ and η. Then $\mathbb{E}(H_a^s) = \xi m \sum_i \phi_i n_{ia}$, $\mathbb{E}(H_a^0) = (1-\xi)m \sum_i \phi_i n_{ia}$, $\mathrm{Var}(H_a^s) = sd_{stutter}^2 \sum_i \phi_i n_{ia}$ and $\mathrm{Var}(H_a^0) = sd_{allele}^2 \sum_i \phi_i n_{ia}$ similar to the Bivariate Normal model (equation 11). In addition ϕ_{vic}, ϕ_K, ϕ_{R_i} and ϕ_{sus} will be the fraction of DNA contributed by the victim, the additional known contributor, the i^{th} random contributor and the suspect respectively. Let $\boldsymbol{\theta}$ be the vector of parameters the likelihood $L(\boldsymbol{\theta}, E)$ is to be maximized over, with $\boldsymbol{\theta'}$ the values found by maximization. Then standard errors σ_i for parameter θ_i are estimated as follows:

$$\sigma_i = \frac{1}{\sqrt{\frac{\partial^2 L(\boldsymbol{\theta},E)}{\partial \theta_i^2}(\boldsymbol{\theta'})}}. \tag{44}$$

The parameter ϕ_{sus} is defined as the remainder contribution $1 - \sum_{\phi_i \setminus \phi_{sus}} \phi_i$ and as such the likelihood is not a function of ϕ_{sus}. Thus there is no standard error calculated for this parameter. This is also true for the parameter $sd_{stutter}$ in case of the Gamma model, which is defined as a function of m, ξ and sd_{allele}. The second partial derivatives are found by numerical estimation using the R-package numDeriv [11, Gilbert, P. and Varadhan, R. (2012)]. For numerical maximization the R-package Rsolnp [10, Alexios Ghalanos and Stefan Theussl (2012)] was used.

Sample MC18 with the suspect assumed to have contributed (H_0)

	Gamma model	Bivariate Normal model		
		Order= 0	Order= 1	Order= 2
r	-	-	0.84 ± 0.12	0.95 ± 0.02
m	1064 ± 19	1062 ± 20	1063 ± 21	1064 ± 19
ξ	0.079 ± 0.008	0.077 ± 0.007	0.071 ± 0.004	0.071 ± 0.003
sd_{allele}	82.7 ± 8.5	87.1 ± 10.1	86.4 ± 9.1	88.5 ± 8.5
$sd_{stutter}$	24.3	18.0 ± 5.4	16.7 ± 3.9	17.5 ± 2.6
ϕ_{vic}	0.705	0.706	0.704	0.704
ϕ_K	0.085 ± 0.008	0.085 ± 0.008	0.083 ± 0.008	0.084 ± 0.008
ϕ_{R_1}	0.023 ± 0.010	0.020 ± 0.009	0.024 ± 0.007	0.024 ± 0.006
ϕ_{sus}	0.187 ± 0.009	0.189 ± 0.009	0.189 ± 0.009	0.187 ± 0.008
likelihood (log)	$-1.178e02$	$-1.146e02$	$-1.137e02$	$-1.140e02$

Table 4: Parameters of the Gamma model and the Bivariate Normal model obtained by maximum likelihood estimation for the DNA profile obtained from MC18 with the suspect assumed to have contributed.

Table 4 shows the estimated parameters with accompanying likelihoods for the different models. The different models agree roughly on all the parameters. Greatest variation between the models appears to be between the parameter $sd_{stutter}$. However, standard errors for this parameter are also relatively large. The correlation is estimated to be very significant, especially for the second order approximation. Additionally, whereas the standard error for this parameter is fairly large for the first order approximation, the parameter value shows little uncertainty for the second order approximation. Using the uncorrelated Bivariate Normal model seems to offer a better fit then the Gamma model going by the difference in the likelihood. A smaller improvement can be made by using higher order approximations, although this difference is fairly small. The first order approximation results in the highest likelihood.

Sample MC15 with the suspect assumed to have contributed (H_0)

Table 5 shows the estimated parameters with accompanying likelihoods for the different models. Again the models are mostly in agreement on all the parameters. Greatest variation between the models appears to be between the parameter $sd_{stutter}$ and also in r between the Bivariate Normal models with order$= 1, 2$. The parameter $sd_{stutter}$ shows high standard errors across the board. This is even moreso the case for r. The correlation is estimated to be fairly insignificant, especially for the first order approximation. There is very little difference in the maximized likelihood between all the different models.

	Gamma model	Bivariate Normal model		
		Order= 0	Order= 1	Order= 2
r	-	-	0.02 ± 0.33	0.46 ± 0.22
m	919 ± 19	920 ± 21	920 ± 21	914 ± 22
ξ	0.069 ± 0.008	0.071 ± 0.006	0.071 ± 0.006	0.070 ± 0.006
sd_{allele}	83.9 ± 8.9	91.9 ± 10.9	91.9 ± 10.9	91.2 ± 10.5
$sd_{stutter}$	22.8	13.2 ± 4.6	13.3 ± 4.7	17.9 ± 4.7
ϕ_{vic}	0.822	0.823	0.826	0.827
ϕ_K	0.043 ± 0.008	0.042 ± 0.008	0.042 ± 0.008	0.047 ± 0.008
ϕ_{R_1}	0.018 ± 0.011	0.021 ± 0.009	0.020 ± 0.009	0.014 ± 0.009
ϕ_{sus}	0.117 ± 0.008	0.111 ± 0.009	0.111 ± 0.009	0.112 ± 0.009
likelihood (log)	$-1.060e02$	$-1.058e02$	$-1.058e02$	$-1.054e02$

Table 5: Parameters of the Gamma model and the Bivariate Normal model obtained by maximum likelihood estimation for the DNA profile obtained from MC15 with the suspect assumed to have contributed.

Sample MC18 with a random person assumed to have contributed ($\mathbf{H_1}$)

Estimated parameters with accompanying likelihoods for the different models are shown in table 6. There is more variation between parameters than in the previous situations. The parameter $sd_{stutter}$ shows even greater variability between models than before, so much so that it cannot reasonably be explained by variance looking at the estimated standard errors. The same is true for the stutter parameters ξ and ϕ_{R_1}. The correlation is estimated to be very significant, similarly to hypothesis H_0 for the same sample. Using the uncorrelated Bivariate Normal model seems to offer a better fit then the Gamma model going by the difference in the likelihood. A smaller improvement can bet made by using higher order approximations, although this difference is smaller. The second order approximation results in the highest likelihood.

Sample MC15 with a random person assumed to have contributed ($\mathbf{H_1}$)

Table 7 shows the estimated parameters with accompanying likelihoods for the different models. As before, the models are mostly in agreement on all the parameters. The parameter $sd_{stutter}$ differs significantly between the Gamma and the Bivariate Normal model. Correlation is estimated to be practically non-existent, though the standard errors show large uncertainty for the parameter r. Likelihoods are very similar between the different models.

Likelihood ratios

The likelihood ratios obtained from the maximized likelihood functions above are shown in table 8. The different models result in very similar likelihood ratios with the second order approximated Bivariate Normal model for sample MC18 being the only large

	Gamma model	Bivariate Normal model		
		Order= 0	Order= 1	Order= 2
r	-	-	0.84 ± 0.12	0.98 ± 0.10
m	1063 ± 17	1061 ± 17	1059 ± 19	1045 ± 13
ξ	0.074 ± 0.008	0.068 ± 0.007	0.077 ± 0.005	0.085 ± 0.003
sd_{allele}	74.6 ± 8.4	75.0 ± 9.1	78.8 ± 9.1	85.2 ± 7.9
$sd_{stutter}$	21.2	14.8 ± 5.6	26.9 ± 7.2	34.9 ± 3.9
ϕ_{vic}	0.694	0.690	0.702	0.717
ϕ_K	0.085 ± 0.009	0.089 ± 0.007	0.085 ± 0.008	0.083 ± 0.007
ϕ_{R_1}	0.032 ± 0.008	0.032 ± 0.008	0.021 ± 0.011	0.005 ± 0.006
ϕ_{R_2}	0.188 ± 0.009	0.188 ± 0.009	0.192 ± 0.012	0.195 ± 0.007
likelihood (log)	$-1.306e02$	$-1.275e02$	$-1.267e02$	$-1.253e02$

Table 6: Parameters of the Gamma model and the Bivariate Normal model obtained by maximum likelihood estimation for the DNA profile obtained from MC18 with an additional random person assumed to have contributed.

outlier.

Implications

The two samples MC18 and MC15 have shown mixed results. The Bivariate Normal model to be for the different order approximations has been shown to have a very similar goodness of fit (similar likelihoods) to the Gamma model. This was accompanied by very low estimated values of the correlation coefficient r. The parameter $sd_{stutter}$ did differ between the Gamma and Bivariate Normal model, although standard errors indicate significant uncertainty. In contrast, estimates from sample MC18 show a general increase in likelihood from the Gamma model to the Bivariate Normal model. Increasing the order approximation further improves the goodness of fit. Only in the case of hypothesis $matrmH_0$ does the first order approximation have a slightly higher likelihood than the second order approximation. The correlation between the stutter and allelic peak is estimated to be very significant. In addition to variation between the Gamma and Bivariate Normal model, $sd_{stutter}$ also shows a lot of variation between the different order approximations for hypothesis $mathrmH_0$.

The two extra degrees of freedom the Bivariate Normal model has over the Gamma model can be expressed by the parameters $sd_{stutter}$ and r. Unfortunately these are exactly the parameters that show difficulty in being estimated with any consistency. One sample shows high correlation with a relatively high degree of certainty, whereas the other estimates low correlation but with a great deal of uncertainty. Although in the latter case a very high correlation coefficient does seem unlikely. The parameter $sd_{stutter}$ seems to be generally lower for the Bivariate Normal model, if not for some drastically higher values found for the second order approximation. Standard errors also seem to be relatively large for this parameter. One has to question the validity of

	Gamma model	Bivariate Normal model		
		Order= 0	Order= 1	Order= 2
r	-	-	$3.46e-15 \pm 0.37$	$2.90e-07 \pm 0.49$
m	921 ± 20	918 ± 21	918 ± 21	920 ± 23
ξ	0.070 ± 0.010	0.072 ± 0.006	0.072 ± 0.006	0.072 ± 0.005
sd_{allele}	84.4 ± 10.8	91.4 ± 13.4	91.4 ± 13.4	104.5 ± 14.5
$sd_{stutter}$	23.1	13.9 ± 5.6	13.9 ± 5.6	12.2 ± 4.2
ϕ_{vic}	0.810	0.806	0.806	0.791
ϕ_K	0.033 ± 0.017	0.031 ± 0.011	0.031 ± 0.011	0.024 ± 0.011
ϕ_{R_1}	0.041 ± 0.010	0.055 ± 0.011	0.055 ± 0.011	0.092 ± 0.018
ϕ_{R_2}	0.116 ± 0.017	0.108 ± 0.032	0.108 ± 0.032	0.092 ± 0.018
likelihood (log)	$-1.197e02$	$-1.192e02$	$-1.192e02$	$-1.192e02$

Table 7: Parameters of the Gamma model and the Bivariate Normal model obtained by maximum likelihood estimation for the DNA profile obtained from MC15 with an additional random person assumed to have contributed.

	MC18	MC15
Gamma model	12.82	13.74
Bivariate Normal model, order=0	12.88	13.41
Bivariate Normal model, order=1	13.01	13.41
Bivariate Normal model, order=2	11.26	13.74

Table 8: Log-Likelihood ratios of the Gamma model and the Bivariate Normal model obtained by maximum likelihood estimation.

adding additional parameters to a model when these parameters cannot be estimated consistently from actual data. The fact that the Bivariate Normal model shows no better fit for one sample is especially disappointing considering considering the fact that the likelihood is maximized over an additional number of parameters (either one or two). Furthermore, the likelihood ratios are very similar for all models except for the second order approximated Bivariate Normal model with sample MC18. Since calculating the likelihood ratio was the primary goal, using any of these models, with one exception, would have resulted in the same conclusion.

6.2.2 Assessing absence/presence of peak distribution

To further examine the validity of the different models it is useful to examine whether the data represents a plausible outcome given the model in question. Before examining the peak height distribution in this context we will only take the absence or presence of peaks into consideration. Given a sample s and marker m this can be considered as a

sequence of independent binomial experiments B_a as follows:

$$\begin{aligned}\mathbb{P}(\mathrm{B}_a = 1) &= \mathbb{P}(H_a \geq C | \{H_i : i \in \{a+1, \ldots, A_m\}\}) \\ \mathbb{P}(\mathrm{B}_a = 0) &= \mathbb{P}(H_a < C | \{H_i : i \in \{a+1, \ldots, A_m\}\}) = 1 - \mathbb{P}(\mathrm{B}_a = 1).\end{aligned} \tag{45}$$

The distribution of the observed peak height H_{ma} of course being determined by the model (and its parameters) in question. To determine these probabilities additional auxiliary variables D_{ma} are introduced with the same parents as the variables O_{ma} and states 0 and 1 with:

$$\mathbb{P}(D_a = 1 | \mathbf{N}_a = \mathbf{n}_a) = \mathbb{P}(H_{ma} \geq C | H_j, \mathbf{n}_a, \boldsymbol{\phi}_s, \boldsymbol{\theta}). \tag{46}$$

Specifically for the Gamma model this is equal to:

$$\mathrm{P}(D_a = 1) = 1 - G(C; \Sigma, \eta), \tag{47}$$

with $\Sigma = \rho(1-\xi)\sum_i \phi_i n_{ia} + \rho\xi \sum_i \phi_i n_{i,a+1}$ and G the cumulative distribution function of the gamma distribution. And in the case of the Bivariate Normal model they are defined as:

$$\begin{aligned}\mathrm{P}(D_a = 1) = 1 - F_{\mathrm{Normal}} \Bigg(& C; (2-\epsilon)m \sum_i \phi_i n_{ia} + \mathrm{E}(H_{a+1}^{s*}), \\ & \sqrt{sd_{allelic}^2 \sum_i \phi_i n_{ia} + \mathrm{Var}(H_{a+1}^{s*})} \Bigg),\end{aligned} \tag{48}$$

with $F_{\mathrm{Normal}}(x, m, sd)$ being the cumulative distribution function of a normal random variable with mean m and standard deviation sd. The m^{th} order approximation described in section 5.2 also applies here.

Setting evidence for the variables $\{O_i : i \in \{a+1, \ldots, A_m\}\}$ is effectively equal to conditioning on the peak heights $\{H_i : i \in \{a+1, \ldots, A_m\}\}$ (section 5.4). Since no evidence will be entered for the D-variables these auxiliary variables will have no influence on the rest of the model. After propagating the evidence, $\mathbb{P}(B_a = 1)$ can readily be obtained from the corresponding D-variable.

Let p_{ma} be the probability of the outcome of the binomial experiment for allele a and marker m as described above. We will define the partial sum:

$$\mathrm{PartialSum}(m, a) = \left(\prod_{i=1}^{m-1} \prod_{j=1}^{A_m} -\log(p_{ij}) \right) \cdot \left(\prod_{j=1}^{a} -\log(p_{mj}) \right). \tag{49}$$

The progression of this partial sum through markers and alleles is plotted in figures 17 through 20 for both hypotheses ($\mathrm{H}_i : i \in \{0, 1\}$) and for both samples MC18 and MC15. Via Monte Carlo simulation of the binomial sequence with sample size $10,000$, .95 and .99 quantiles were estimated. These are plotted in blue and red respectively. All partial

sums end up below the .95 quantile although some are further removed than others. The 2^{th}-order approximated bivariate normal model appears to be an outlier with respect to the other models for sample MC18 and hypothesis H_0. The partial sum appears to flirt with crossing the over the .95 quantile. As such, one could suspect the data be somewhat unlikely given this model. Perhaps unsurprisingly, this corresponds to the one outlier among the likelihood ratios (table 8). Disregarding this case the partial sums all look very similar for equal sample and hypothesis.

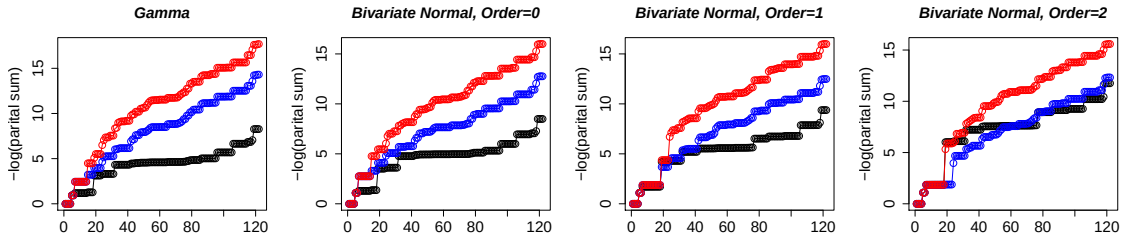

Figure 17: Progression of the partial sum (equation 49) for sample MC18 and hypothesis H_0 with estimated .95 and .99 quantiles in blue and red resp.

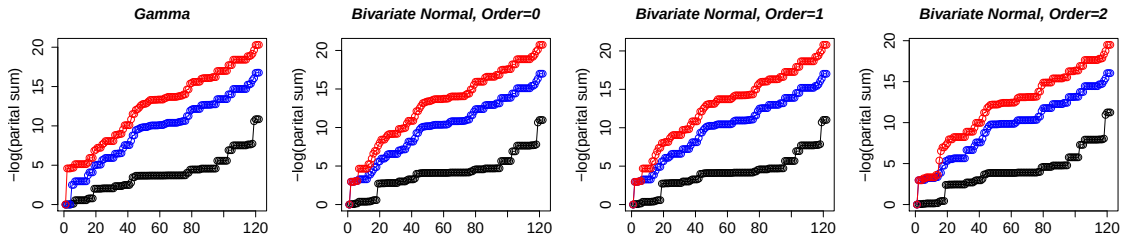

Figure 18: Progression of the partial sum (equation 49) for sample MC15 and hypothesis H_0 with estimated .95 and .99 quantiles in blue and red resp.

6.2.3 Assessing peak height distribution

Consider the peak height distribution given this peak not dropping out, $(H_a|H_a \geq C)$. Then this is a continuous distribution for $x \geq C$ and as a result the random variable $\mathbb{P}(H_a \leq x_a|H_a \geq C$ is uniformly distributed on $(0, 1)$. Let Dr_m be the set of alleles in which no peak is observed for marker m and x_{ma} the peak height at allele a and marker m. Then under the null hypothesis of the peak heights following the distribution obtained by maximum likelihood estimation of the model in question we can consider the following random variables a results of random draws from independent uniformly

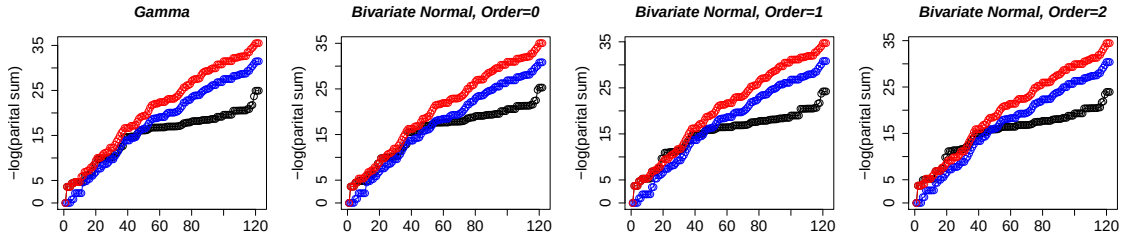

Figure 19: Progression of the partial sum (equation 49) for sample MC18 and hypothesis H_1 with estimated .95 and .99 quantiles in blue and red resp.

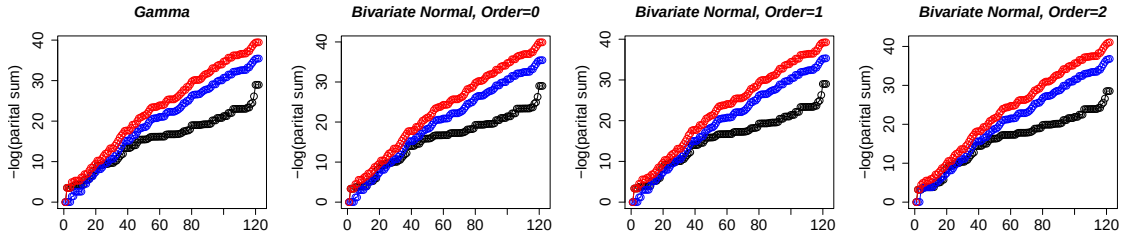

Figure 20: Progression of the partial sum (equation 49) for sample MC15 and hypothesis H_1 with estimated .95 and .99 quantiles in blue and red resp.

distributed random variables:

$$B_{ma} \sim \mathbb{P}\left(H_{am} \leq x_{ma} | H_{ma} \geq C, H_{im} = x_{mi} \ \forall \ i \in \ \{a+1, \ldots, A_m\} \setminus Dr_m, \ H_{mj} < C \ \forall \ j \in Dr_m\right). \tag{50}$$

Conditioning on $H_{mi} \ \forall \ i \in \ \{a+1, \ldots, A_m\} \setminus Dr_m$ and $H_{mj} \forall \ j \in Dr_m$ is equivalent to entering evidence into the corresponding O-variables. Finally the additional auxiliary variables Q_{ma} are introduced:

$$\mathbb{P}(Q_{ma} = 1 | \mathbf{N}_{ma} = \mathbf{n}_{ma}) = \mathbb{P}\left(H_{ma} \leq x_{ma} | \mathrm{H}_j, \mathbf{n}_{ma}, \boldsymbol{\phi}_s, \boldsymbol{\theta}_m\right). \tag{51}$$

For specificity regarding the model used the D-variables in section 6.2.2 can be taken as a template. Note that no evidence is entered for either the Q-variables or the D-variables. Then the outcome of B_{ma}; b_{ma}, is easily evaluated by propagating the evidence and evaluating:

$$b_{ma} = \frac{\mathbb{P}(Q'_{ma} = 1) - \mathbb{P}(D'_{ma} = 0)}{1 - \mathbb{P}(D'_{ma} = 0)}, \tag{52}$$

with Q' and D' the updated distributions after propagation.

Q-Q plots of the quantiles of these b_{ma} versus the theoretic quantiles of a uniform distribution are plotted in figures 21 through 24 for both hypotheses ($\mathrm{H}_i : i \in \{0, 1\}$) and for both samples MC18 and MC15. No great difference is evident between the models. The Q-Q plots of the prosecutions hypothesis H_1 do show more deviation from the expected quantiles when compared to the defense's hypothesis H_0.

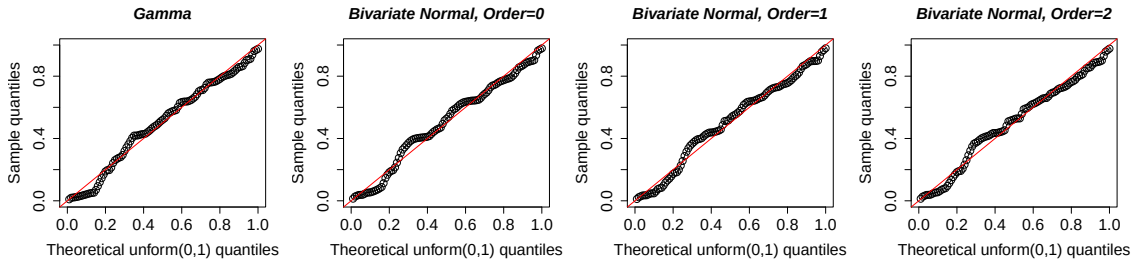

Figure 21: Q-Q plots of the quantiles of $\{b_{am} : m \in \{1, \ldots, M\}, a \in \{1, \ldots, A_m\}\}$ versus the theoretic quantiles of a uniform distribution for sample MC18 and hypothesis H_0.

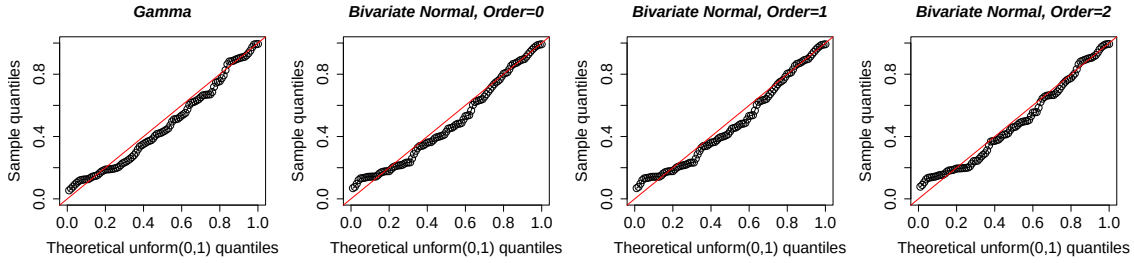

Figure 22: Q-Q plots of the quantiles of $\{b_{am} : m \in \{1, \ldots, M\}, a \in \{1, \ldots, A_m\}\}$ versus the theoretic quantiles of a uniform distribution for sample MC15 and hypothesis H_0.

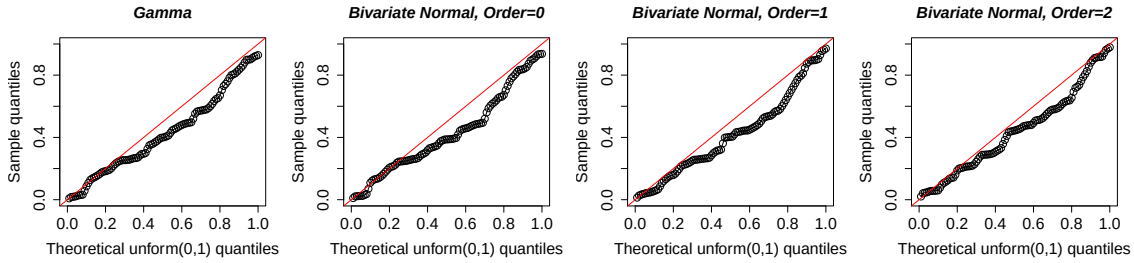

Figure 23: Q-Q plots of the quantiles of $\{b_{am} : m \in \{1, \ldots, M\}, a \in \{1, \ldots, A_m\}\}$ versus the theoretic quantiles of a uniform distribution for sample MC18 and hypothesis H_1.

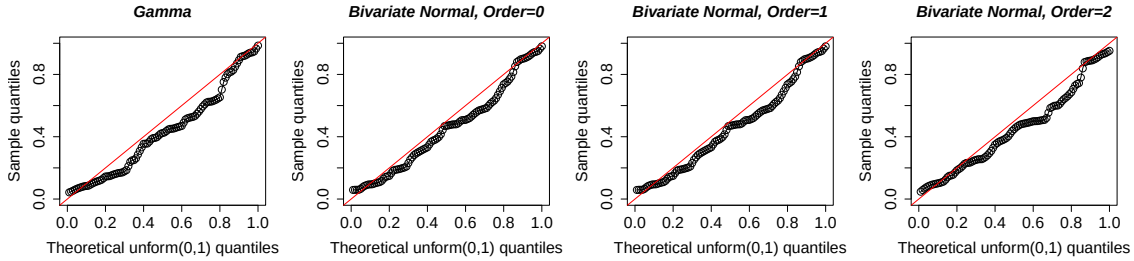

Figure 24: Q-Q plots of the quantiles of $\{b_{am} : m \in \{1, \ldots, M\}, a \in \{1, \ldots, A_m\}\}$ versus the theoretic quantiles of a uniform distribution for sample MC15 and hypothesis H_1.

7 Discussion

Validation of the Bivariate Gamma model using the simulation model of section 3 gives promising results. Problems inherent to using a gamma distribution in modeling the peak height distribution could be resolved by using a bivariate normal model. Firstly there is no restriction to one parameter to control both the mean and standard deviation of the stutter peak. Independent gamma distributions with equal scale parameters sum to be gamma distributed. For normal distributions both the mean and standard deviation can be different. In addition the normal distribution has an obvious extension to a bivariate distribution, whereas bivariate gamma distributions are more complex and prone to more restrictions. As a result the Bivariate Normal model appears to perform better when tested in this manner (section 6.1).

To see how the Bivariate Normal model compares to the Gamma model in practise both models were applied to an actual case (section 6.2). Results differed between the two samples. One showed practically no difference between the models over a range of different order approximations. In addition maximum likelihood estimation returned a correlation coefficient close to zero in many of those cases. Thus there appeared to be no benefit to using the Bivariate Normal model over the Gamma model. The

other sample *did* show significant improvement to using the Bivariate Normal model. Even the uncorrelated version of this model produced a higher maximized likelihood than the Gamma model. Using higher order approximations only further heightened the likelihood. Thus we are left with mixed results. The cause of this can only be speculated on. Perhaps the numerous flaws in the method relating to marker and allele specific parameters, the absence of consideration for certain artefacts such as silent alleles, as well as small flaws in the model, causes there to be significant variation as to which model better fits the data of a specific case. It could be that many of these problems would need to be resolved before the proposed change of model yields consistent results.

One has to consider whether the cost of using the Bivariate Normal model are outweighed by the benefits. Switching to the uncorrelated Bivariate Model only comes at the cost of needing the maximize over an extra parameter, whereas upping the order of approximation increases computation time of the likelihood function significantly. In the case of two random contributors, when reaching the point where this computation time is mostly determined by the time it takes to propagate the network, the computation times roughly increases ninefold when upping the order of approximation by one. When adding extra random contributors or considering a particular random contributor for multiple samples, the complexity of the network quickly increases and using the Bivariate Normal model with any order approximation other than zero soon becomes infeasible.

Acknowledgements

I would like to thank Richard Gill for all his help and supervision, as well as Martin Bootsma for agreeing to supervise me.

References

[1] Butler, J. M. (2005). Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Second Edition. Elsevier Academic Press.

[2] Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2010). Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures. Applied Statistics, 59(5) : 855 − 874.

[3] Gill, P., Curran, J., and Elliot, K. (2005). A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. Nucleic Acids Research, 33(2) : 632 − 643.

[4] Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2007). A gamma model for DNA mixture analyses. Bayesian Analysis, 2(2) : 333 − 348.

[5] Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2011). Probabilistic expert systems for handling artifacts in complex DNA mixtures. Forensic Science International: Genetics, 5(3) : 202 − 209.

[6] Cowell, R. G., Lauritzen, S. L.,Graversen, T. and Mortera, J. (2013). Analysis of DNA Mixtures with Artefacts. arXiv:1302.4404.

[7] Cowell, R. G. (2009). Validation of an STR peak area model. ForensicScience International: Genetics, 3(3) : 193 − 199.

[8] Graversen, T. (2013). Statistical Analysis of DNA Mixtures. PhD thesis, Department of Statistics, University of Oxford. arXiv:1307.4956.

[9] Gill, P., Curran, J., Neumann, C., Kirkham, A., Clayton, T., Whitaker, J., and Lambert, J. (2008). Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. Forensic Science International: Genetics, 2(2) : 91 − 103.

[10] Alexios Ghalanos and Stefan Theussl (2012). Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. R package version 1.14.

[11] Gilbert, P. and Varadhan, R. (2012). numDeriv: Accurate Numerical Derivatives. R package version 2012.9 − 1.

8 Additional Figures

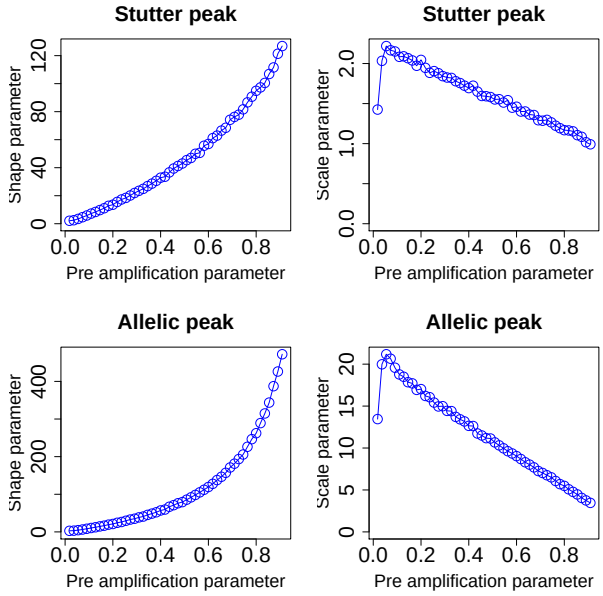

Figure 25: Parameters of the gamma distribution fitted by maximum likelihood estimation for the stutter and allelic peak with variable pre-amplification parameter ($\pi_{extraction,aliquot}$).

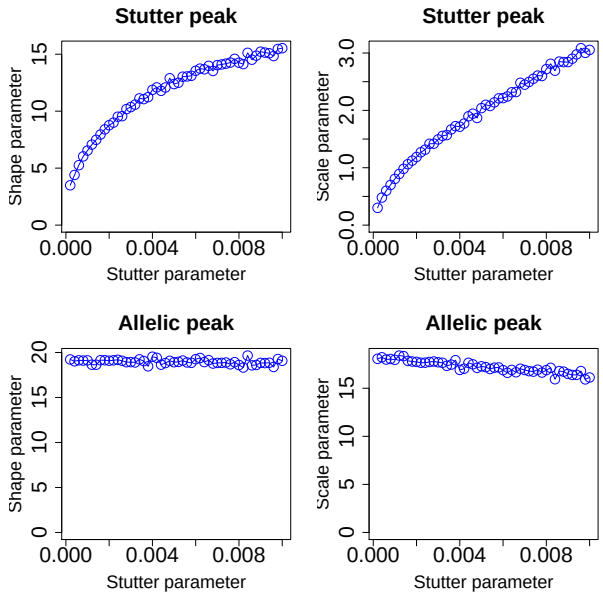

Figure 26: Parameters of the gamma distribution fitted by maximum likelihood estimation for the stutter and allelic peak with variable stutter parameter ($\pi_{stutter}$).

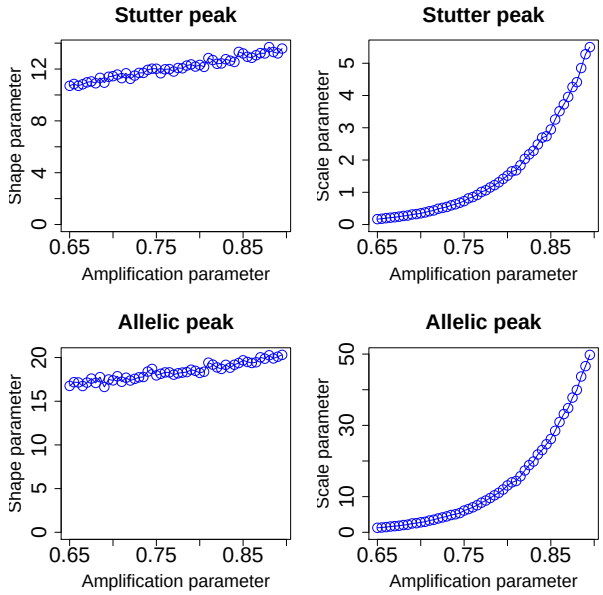

Figure 27: Parameters of the gamma distribution fitted by maximum likelihood estimation for the stutter and allelic peak with variable amplification parameter (π_{PCReff}).

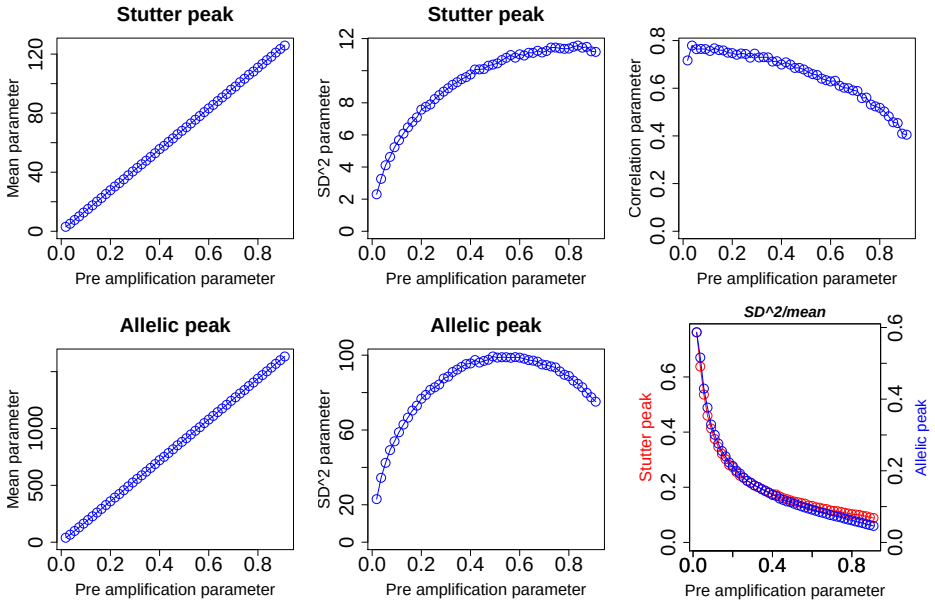

Figure 28: Parameters of the bivariate normal distribution fitted by maximum likelihood estimation for the stutter and allelic peak with variable pre-amplification parameter ($\pi_{extraction,aliquot}$).

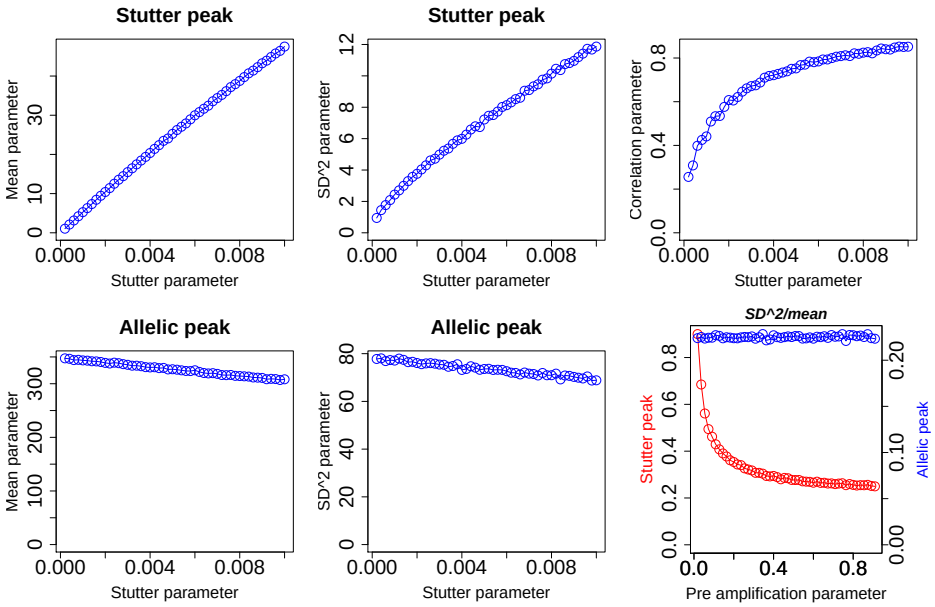

Figure 29: Parameters of the bivariate normal distribution fitted by maximum likelihood estimation for the stutter and allelic peak with variable stutter parameter ($\pi_{stutter}$).

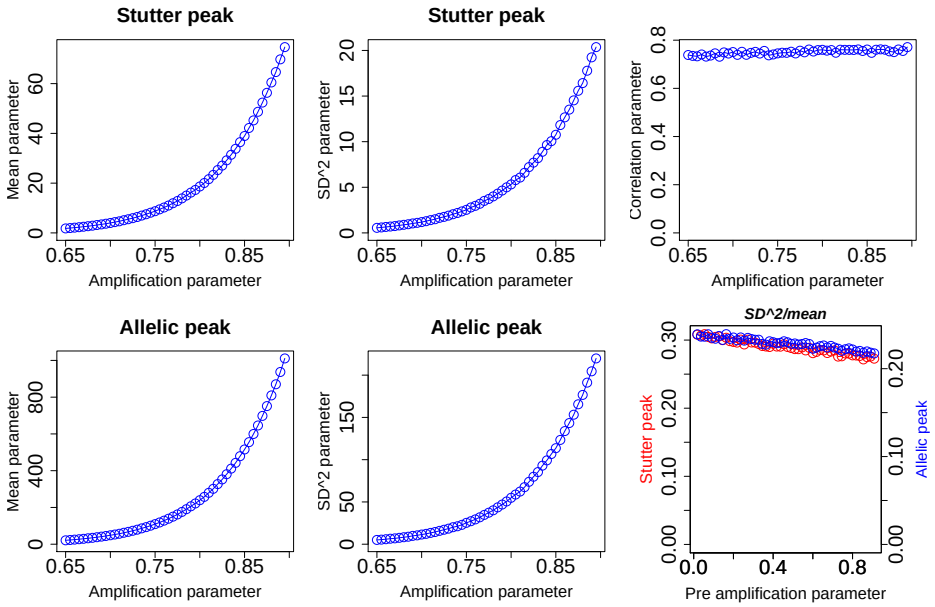

Figure 30: Parameters of the bivariate normal distribution fitted by maximum likelihood estimation for the stutter and allelic peak with variable amplification parameter (π_{PCReff}).

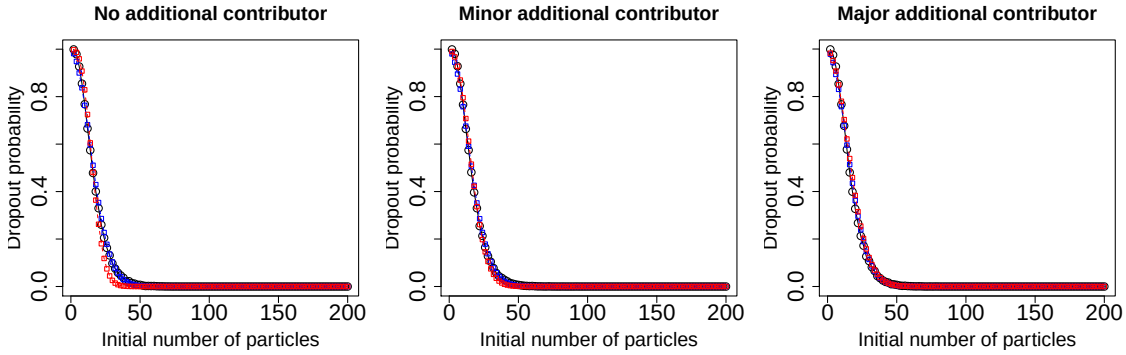

Figure 31: Dropout rates of the allelic peak as a function of the starting amount of DNA N. Black are the dropout estimates from the data, red and blue from maximum likelihood estimation of the Gamma model with red having $\rho \cdot \eta$ and ξ estimated directly from the data.

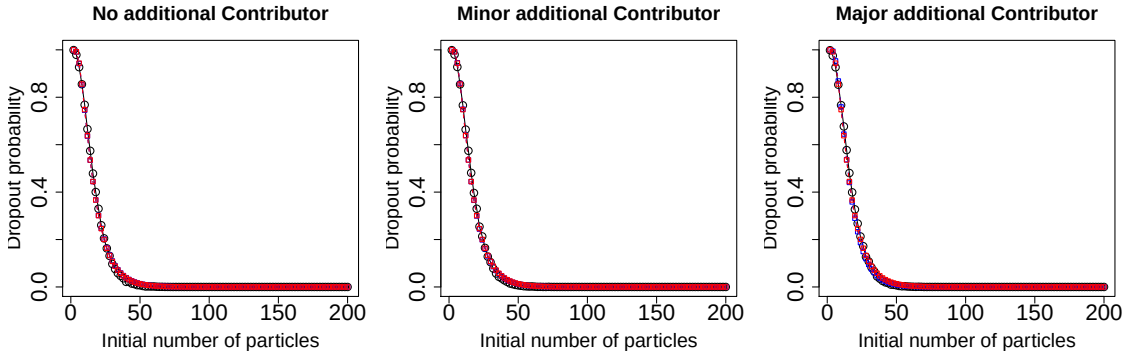

Figure 32: Dropout rates of the allelic peak as a function of the starting amount of DNA N. Black are the dropout estimates from the data, blue from maximum likelihood estimation of the Bivariate Normal model and red with parameters estimated from the data.