university of groningen / faculty of science and engineering

Probabilistic Modelling of mRNA Electropherograms in Fluid Mixtures

Andrei Secuiu s4260732 | Roberto Schinina s4612299 | Dewi Batista s3313093

1 Introduction

In [6], the authors detail a model capable of evaluating a class of hypotheses related to the presence or absence of bodily fluids in fluid mixtures taken from scenes of alleged sexual assault. The relevant bodily fluids include vaginal mucosa, menstrual secretion, blood, nasal mucosa, saliva, semen (fertile and sterile) and skin (penile and non-penile). More precisely, given a fluid mixture, the hypothesis evaluated is of the form:

H_1: At least one of vaginal mucosa or menstrual secretion is present.
H_2: Neither vaginal mucosa nor menstrual secretion is present.

The choice of vaginal mucosa and menstrual secretion as fluids of interest in the above hypothesis is due to their relevance to cases of alleged sexual assault. To evaluate such a hypothesis, one extracts the values $m_1, \ldots, m_K$ of K pre-chosen genetic markers from an mRNA profile of the fluid mixture and computes a likelihood ratio of the form

$$\mathrm{LR} = \frac{p(m_1, \ldots, m_K | H_1)}{p(m_1, \ldots, m_K | H_2)}.$$

The strength of evidence in favour of one hypothesis over the other is assessed by comparing LR to an appropriate threshold value LR^*. If $\mathrm{LR} > \mathrm{LR}^*$ then the model evidence favours H_1 and otherwise favours H_2. As to what genetic markers are precisely, such an understanding is unneeded for this work. From here, one need only understand that genetic marker values are extractable and quantifiable given a fluid mixture via modern mRNA profiling and are statistically informative measurements for assessing the presence or absence of bodily fluids in a given fluid mixture. A detailed explanation of the biology relevant to mRNA profiling is given in Appendix A.

While the authors demonstrate the effectiveness of their model, a natural limitation is that it offers no concrete way of concluding precisely *which* of the fluids of interest are present or absent in a given fluid mixture. In line with this, a generative model fit to the distribution of marker values conditioned on fluid combinations is of interest. Such a model would allow the computation of evidence under arbitrary hypotheses of fluid presence(s)/absence(s). As such, the research question

that our work intends to answer is to what extent one can model the distribution of marker values conditioned on fluid combinations. That is, how well can one model

$$p(m_1, \ldots, m_K | f),$$

where $f = (f_1, \ldots, f_L)$ denotes the presence of L fluids of interest, in a way that allows for efficient sampling.

2 Method

In this section, we offer a brief overview of the dataset used to fit the desired distributions followed by a motivation of the use of mixture models for this purpose. Details pertaining to the implementation of said mixture models are then given.

2.1 Data

The dataset used to fit our models can be found under `./data/mixtures.csv` from the root directory of our repository, a link to which is given at the beginning of the appendices. It consists of 350 data points each of which correspond to a fluid mixture concocted in a laboratory and was taken from the repository of the original work summarised in the introduction (`./Datasets/Dataset_NFI_rv.xlsx` from their root directory). Each data point consists of measurements of the 15 genetic markers of interest[1] as well as the bodily fluids present in the concocted fluid mixture. Within, only six distinct fluids are present in the overall dataset: blood, menstrual secretion, nasal mucosa, saliva, semen (fertile) and vaginal mucosa. Additionally, upon inspection, the number of bodily fluids present in each concocted fluid mixture, and correspondingly each data point of the dataset, is precisely two. In line with this, we simplify our research objective by instead fitting the distribution of marker values given fluid combinations consisting of precisely two fluids. That is, we seek to fit

$$p(m_1, \ldots, m_{15} | f_i, f_j)$$

where f_i and f_j are two of the six fluids of interest, denoted $f_1, \ldots, f_6$.

Upon further inspection, one observes that precisely six distinct fluid pairs are present in `mixtures.csv`: semen (fertile) and vaginal mucosa, saliva and vaginal mucosa, blood and nasal mucosa, nasal mucosa and saliva, blood and vaginal mucosa and, finally, blood and menstrual secretion.

Pre-processing. All pre-processing applied to `mixtures.csv` was done with the intention of improving its usability in fitting our models. The pre-processed version can be found under `./data/preproc_mixtures.csv` from the root directory of our repository and the program used to apply the pre-processing can be found under `./data_preprocessing.py` within the root directory.

The pre-processing applied includes the removal of the columns pertaining to marker values which are not of interest to our work. Following this, we transform the single column consisting of the fluids present in each fluid mixture to their one-hot encodings. For example, the first data point of `mixtures.csv` corresponds to a fluid mixture made up of semen (fertile) and vaginal mucosa

[1]The 15 genetic markers of interest in our work are HBB, ALAS2, CD93, HTN3, STATH, BPIFA1, MUC4, MYOZ1, CYP2B7P1, MMP10, MMP7, MMP11, SEMG1, KLK3 and PRM1.

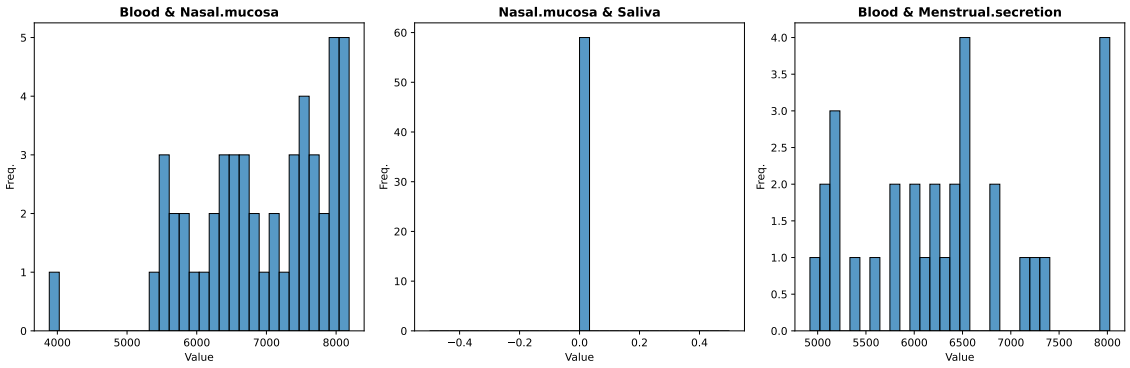

Figure 1: Histograms of the values of the marker HBB conditioned on three fluid combinations.

which is recorded as `Semen.fertile+Vaginal.mucosa` in the first column. In one-hot encoding the fluids present for this data point, the fluids present are recorded via values of 1 under the two columns pertaining to the presence of semen (fertile) and vaginal mucosa followed by values of 0 in each other fluid-related column. Finally, in `mixtures.csv`, values of 0 for genetic markers are left unfilled and so are interpreted as `NaN` entries by some interpreters. To reduce potential issues regarding `NaN` entries, we replace each of these with 0s.

2.2 Motivating Mixture Models

Inspired by relevant literature in DNA profiling [2], we simplify our goal by assuming conditional independence of marker values given the fluids present. That is, we assume that

$$p(m_1, \ldots, m_{15}|f_i, f_j) = p(m_1|f_i, f_j) \cdot \ldots \cdot p(m_{15}|f_i, f_j)$$

for present fluids of interest f_i and f_j. A natural question when presented such a simplifying assumption is: how do the individual distributions of marker values conditioned on different fluid pairs look? The shape of the histograms for each of these distributions, plotted using the data from `preproc_mixtures.csv`, varies as illustrated in Figure 1. To clarify the behaviour of the plot in the middle of Figure 1, for a given fluid pair, one does not necessarily have that each genetic marker value is non-zero. In fact, for a given fluid pair, most of the genetic marker values are 0 across all corresponding data points. As a result, the histogram of marker values for that genetic marker consists of a single spike at 0 for the given fluid pair.

In assessing each of these histograms, of which there are 90 in total (due to 15 markers and 6 fluid pairs present in `mixtures.csv`), we observed that many consisted of what one could be convinced are clusters. For example, in Figure 2 one could be convinced that there is one wide but short cluster on the left and a tall but thin cluster on the right. In this case, fitting a mixture model seems like a reasonable approach. As this behaviour of cluster-presence was observed sufficiently often across the 90 histograms, mixture models were used to fit all individual distributions. For a brief and non-rigorous reminder of mixture models, consider Appendix B.

The classes of mixture models used in our work are Gaussian mixture models and Gamma mixture models. While the use of Gamma mixture models is motivated by relevant literature [2, 3], the

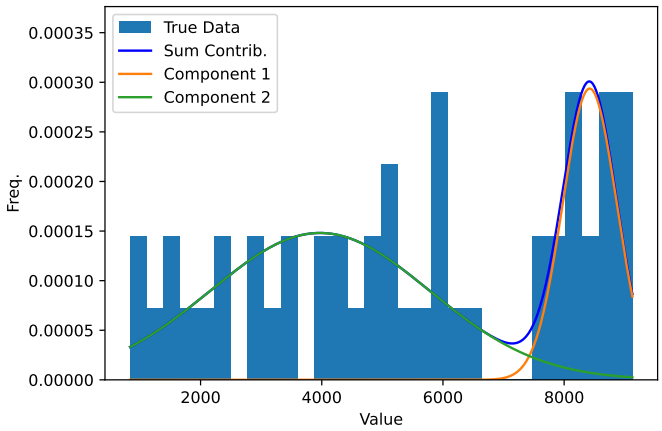

Figure 2: A two-component Gaussian mixture model fit to the marker values of ALAS2 conditioned on the presence of blood and nasal mucosa.

use of Gaussian mixture models is primarily due to the relative simplicity of their implementation.

Note: A brief investigation of our conditional independence assumption's validity is provided in Appendix C. Additionally, all 90 histogram plots of markers conditioned on fluid pairs are provided in `./figures/histograms/2D_individual_markers/` from the root directory of our repository.

2.3 Implementation

Our implementation made use of Python (for Gaussian mixtures and data pre-processing) and R (for Gamma mixtures). In particular, the fitting of a model with a known number of components made use of `sklearn.mixture.GaussianMixture` and `evmix.gammamixEM`. One should note that both implementations rely on the EM algorithm, which is known to have issues with converging to local optima instead of global optima. Thus it is important to set the hyperparameters of the above functions correctly such that multiple attempts are performed, with only the best set of parameters being kept.

In the training of a model, we created special functions to automate the process. A user could specify a priori the number of components to be used by the model, but we recommend using the functions which automate model selection. There is a maximum number of components hard-coded into each implementation in order to allow for faster runtime and to prevent overfitting: 10 for Gaussian mixtures and 5 for Gamma mixtures.

For the purpose of model selection, we utilised the Bayesian Information Criterion (BIC). An important benefit of BIC is that it allows for the direct comparison of different model families, e.g. comparing Gaussian and Gamma mixtures directly. Compared to the Akaike Information Criterion (AIC), it also further punishes a large number of model parameters. We chose BIC-based selection because it results in simpler models, which we have prioritised for the purpose of model building.

In order to determine whether a generated sample originates from the same distribution as the initial data, we used the two-sample Kolmogorov-Smirnov (KS) test. In particular, a leave-out (test

set), not used for training, is compared with an equally-sized generated dataset sampled from the model[2]. The p-value resulting from the test is recorded: the lower the p-value, the more likely it is that the two samples originate from different distributions. Obtained p-values are further-detailed in Section 4.

2.3.1 Implementation of Gaussian Mixtures

All relevant functions are given in `gaussianMixtures.py` and are provided with a detailed documentation describing the role of each function and the role of the parameters.

The function that performs the training of a Gaussian mixture model is `gauss_mixture_fit()`. In the beginning, it starts by splitting the available data given a marker-fluid pair into a training and a testing set, with proportions that can be set by the user[3]. The training data is used for the rest of the function, while the testing data is discarded. If a user desires a pre-defined number of components, the model is trained directly with that number of components. Otherwise 10 models are trained, each with a number of components ranging from 1 to 10; all other parameters are fixed. Then the model with the smallest BIC is selected; a user has the option to also return the relevant BIC value for the best model. Lastly, the user has the option to plot a histogram of the *entire* data (given a marker-fluid pair), together with the best fit for the training data.

The data generation is implemented in `data_generator()`. It relies entirely on a given Gaussian mixture model, thus a user can easily generate data according to any Gaussian mixture model; no training data is required for it. In the generation, a user can specify a seed for reproducibility. The user can also specify a threshold: all values below the threshold are returned as zeros. To generate data, a vector of integers corresponding to the relevant components is first generated, with probabilities given by the model weights. The number of data points to be generated for each component is aggregated. Afterwards, the relevant number of data points for each mixture component is sampled with the relevant mean and covariance matrix.

For convenience, the relevant metrics and plots for judging the adequacy of a fit are automated by the `fit_evaluator()` function. One can obtain the BIC for the best fit on the training data, the minimum and the median p-values for repeated KS tests and a histogram of the entire available data together with the generated data and the probability density function of the model fit.

2.3.2 Implementation of Gamma Mixtures

All relevant functions are given in `gammaMixtures.R` and are provided with detailed documentation describing the role of each function and the role of the parameters.

Since our implementation of Gamma mixtures is virtually identical to that of Gaussian mixtures, we limit what is stated here to a list of the differences.

- The function `gamma_mixtures()` performs model selection.
- The maximum number of components over which we do model selection is 5. The choice of lowering this last value (from 10 in the case of Gaussian Mixtures) is due to the fact that each Gamma distribution tends to not encapsulate every peak (or group of peaks) like the Gaussian does. This vastly reduces model selection runtime.

[2]One could use the two-sample KS test to compare datasets of different sizes. However, during a quick web search we saw multiple recommendations to use datasets of similar sizes; it should result in the highest power for the test.

[3]By default, the proportion is a value very close to 1. A value of exactly 1 returns an error.

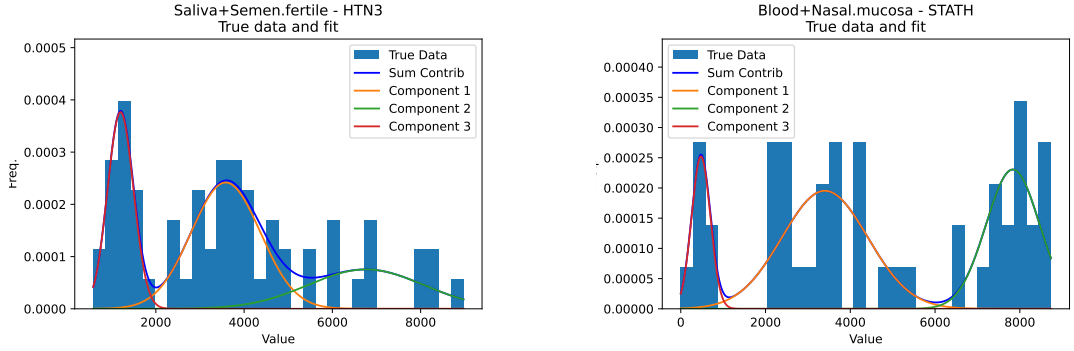

Figure 3: Example output for Gaussian mixtures, fitted on the entire dataset. The figure on the left shows an adequate fit. The figure on the right is mostly satisfactory, but the middle component does not seem to be well-modelled by a Gaussian.

- The function `p_value_statistics()` performs the two-sample KS test for 100 generated datasets.

3 Results

All programs and figures relevant to our work can be found in our GitHub repository, a link to which is given at the beginning of the appendices. Additionally, our repository contains the recorded BIC and p-value(s) attained by each selected mixture model in `NFI_Mixtures_Results.xlsx`.

3.1 Gaussian Mixtures

The performance of Gaussian mixtures was mixed. The fits for some fluid-marker combinations are adequate (as illustrated in Figure 3), but it was often the case that some combinations were not well-modelled by a Gaussian mixture (as illustrated in Figure 4). The main inadequacies of the fits can be categorised into the following:

- Some components did not have a Gaussian shape. Examples: `Blood + Nasal.mucosa - STATH` and `Menstrual.secretion + Blood - HBB`.
- Singletons (i.e. components with very small variance) are present. Examples: `Meanstrual.secretion + Blood - MMP11` and `Saliva + Semen.fertile - SEMG1`.

We noticed that the adequacy of Gaussian mixture models for the data depended mostly on the markers as opposed to the fluid pairs, but said dependence is not exclusive. Many marker-fluid pairs were modelled reasonably well, but there is also a sizeable number of pairs with inadequate fits. Due to the subjectivity of the visual assessment, we refrain from providing precise numbers.

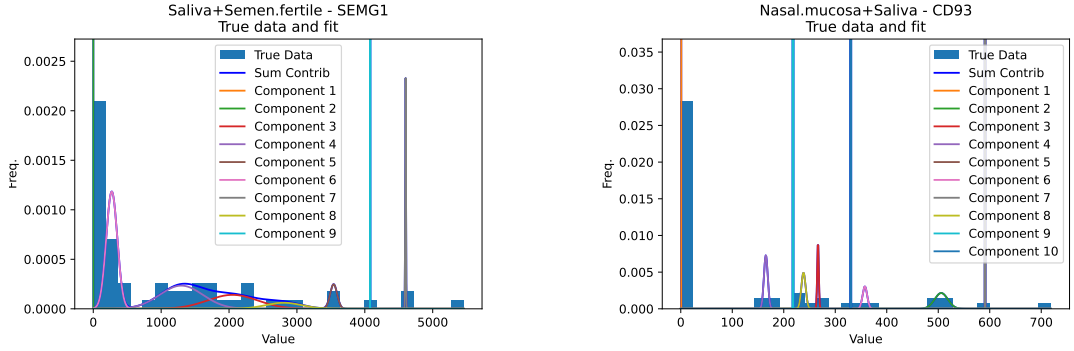

Figure 4: Another example output for Gaussian mixtures, fit on the entire dataset. The histogram on the left has anomalies pertaining to high values, indicating that robust methods could offer improvement. The figure on the right displays a distribution that is very difficult to learn by any method; accordingly, the performance of Gaussian mixtures on it is not good.

3.2 Gamma Mixtures

Similarly, the performance of Gamma mixtures was mixed. There are two main points we would like to draw attention to. Firstly, the most common number of optimal components was one. This may lead to generalisation issues related to underfitting. As illustrated on the left of Figure 5, the learned distribution does not capture the right-most peaks even if they have the same value as most other peaks.

On the other hand, we also have instances where the optimal number of components is larger than one. In such cases, higher peaks are not ignored. However, in some cases, this may lead to overfitting, negatively impacting the model's generalisation.

4 Discussion

4.1 Comparison between Gaussian and Gamma

One advantage of using Gaussian mixtures compared to other distributions (e.g. Gamma) is that convergence towards a result is fast. Further, we have noticed that there were cases where Gaussian mixtures returned a result for which Gamma mixtures did not. Both families of distributions have the same number of parameters for a fixed number of components, thus both are equally prone to overfitting.

The main distinction between the two is that Gaussian mixtures are suitable at modelling symmetric components, while Gamma mixtures are good at capturing asymmetric distributions with e.g. a skew (such as the exponential distribution). The choice of one method over another should be done both using a relevant statistic (such as BIC) and with a visual assessment. In our investigation, we noticed that BIC favoured Gaussian mixtures over Gamma mixtures most frequently: 37 marker-fluid pairs were best modelled by Gaussian mixtures, 7 by Gamma mixtures and 5 had a result only for Gaussian mixtures. One factor that may have influenced the above result is that Gamma mixtures were forced to use at most 5 components (due to convergence issues), which

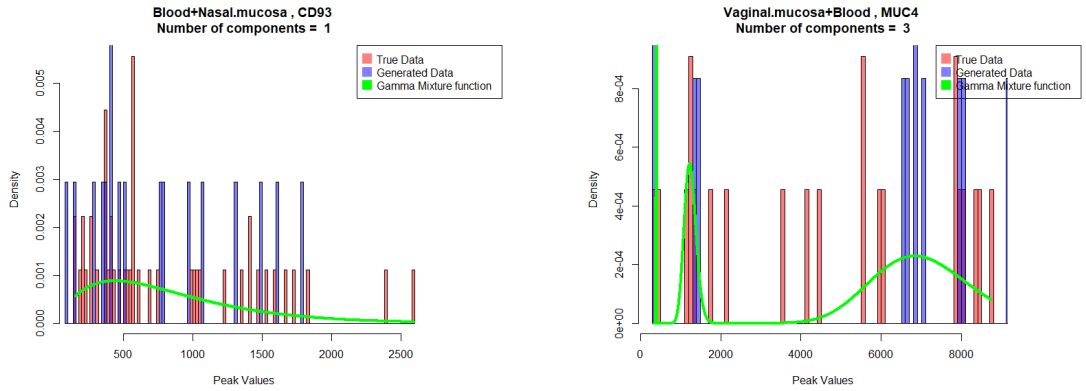

Figure 5: Example output gamma mixtures

is less than the limit of 10 for Gaussians.

4.2 About p-values

The methodology used to test the adequacy of the generated data against the initial data, found in `mixtures.csv`, is worthy of discussion. If one needs to investigate only the adequacy of the generated data against the initial data, then one could omit the creation of the leave-out set. However, this method would be vulnerable to overfitting because the parameters found are expressly designed to fit the available data in the best way the model family (e.g. Gaussian mixtures) allows. This induces a small bias which is addressed by the creation of the leave-out set. If the KS statistic is in the rejection region for data generated on a full model, this is an indicator that the model family is inadequate. However, if the KS statistic does not lead to rejection, then one cannot yet conclude that the modelling procedure is satisfactory.

The logic behind the leave-out set is borrowed from a technique called "cross-validation". The data in the leave-out set is independent from the data in the training set and, implicitly, from a model derived from the training data. As such, the test with the leave-out set is one's best estimate for the generalisation error: the expected performance of the model when encountering new (additional) data. We felt that the size of the leave-out set (one third of the total data) was large enough to obtain a meaningful figure for the KS test, while also giving the model enough data from which to learn meaningful parameter values. That said, the size was not chosen due to any rigorous criterion. A further investigation could be made conducted for the optimal proportion of the dataset used for the leave-out set. Our suggestion would be a simulation study in which around 60 data points are generated according to a Gaussian mixture which resembles the real data. Then, one would apply the methodology above and see which leave-out set size would consistently yield results closest to the truth that are also consistently not rejected.

We noticed that the p-values vary significantly with the generated data (note: the generation is controlled by the seed in the algorithm), even having fixed the marker-fluid pairs and leave-out set. This means that a large number of repetitions (e.g. at least 100) is required to obtain a distribution of p-values to have an informed idea of how well the distributions match overall. We recorded the

lowest number in the list as well as the median: the former to see the worst performance and the latter to see which p-value is a lower bound for 50% of the attempts.

It is important to remark that p-values alone are not a good indicator for how well a distribution is learned. Some inadequate distributions had very good p-values (above 80%), which may be explained by e.g. the sparsity of the particular fluid-marker pair. More precisely, there was a large emphasis on a mixture which modelled the predominant zero value (e.g. "Nasal.mucosa+Saliva - CD93"), with other components having little weight. The model learned to generate mostly values of zero, which falls in line with the actual distribution. However, the non-zero values were badly modelled. Fundamentally, the sparsity of the data for such examples makes the general task of learning a distribution very difficult. Conversely, some nice-looking models resulted in low p-values which we do not have an explanation for. Consequently, we strongly recommend the use of p-values together with a qualitative visual assessment of the goodness-of-fit.

4.3 Lack of confidence intervals and sparsity of data

The implementations of mixture modelling that we have used rely on the EM algorithm to find the optimal parameters. One large drawback of the EM algorithm is that there is no inherent statistical method by which confidence intervals are returned. To address that, we recommend the use of re-sampling methods in the construction of confidence intervals (e.g. bootstrapping). However, the sparsity of the given data makes such an implementation difficult. This is because bootstrapped datasets may have very different shapes among each other, resulting in fitted mixtures which are significantly different and hard to compare. When bootstrapping e.g. the mean μ_i for the i-th component, one would fit B bootstrapped datasets, collect the resulting list of μ_i's and then extract the confidence interval. However, how does one identify which is the i-th component for the B resulting fits, when the models can differ significantly? What would one do when the optimal number of components differs across models (it may happen with data where there are many clusters of few data points)? We have no easy solution to this problem except for the recommendation of acquiring more data.

It was mentioned earlier that the data is sparse for some marker-fluid pairs. This problem is felt especially because there is no a priori probability distribution rooted in biology used in the modelling. Thus one has to both decide on a good family of distributions and to find the best model, given the earlier family. To understand the scope of the problem further and to see how much more data would be needed for a subsequent analysis, we recommend the following method. The objective is to obtain the standard error of the model parameters as a function of the number of data points. With a slight modification, one could also investigate the stability of the number of mixture components as a function of n.

- Begin by fitting the mixture model on the full initial dataset.
- Using the fitted model, generate n new data points.

  Note: The number of new data points should be sufficiently large such that the bootstrapped merged datasets (i.e. initial real data combined with the generated data) result in model fits where the number of components is constant and each component can be identified.
- Bootstrap B merged datasets to construct the bootstrap confidence intervals for each model parameter.
- Repeat the above two steps for increasing values of n.

The advantage of this method is that one can investigate n for each model parameter. In particular, it can be used to assess how many more data points would be required to confidently learn components with few data points (i.e. the hardest to learn).

4.4 Acquiring more data

Ideally, more data would be acquired from new fluid mixtures provided by new donors. However, one drawback of our methodology is that it did not use the entire dataset available: the data from individual fluids. A further study could investigate how the distributions of peak heights of individual fluids are related to the distributions of peak heights of mixtures. One possibility is to apply the above framework to both individual fluids and mixtures, and then judge whether the mixture peak heights distribution is a sum (linear combination) of peak heights from the individual fluids. One could also investigate the correlation between the peak heights of individual fluids, given a marker. Our method does not provide a framework of generating mixtures artificially from distributions of individual fluids, at least not without further study.

4.5 Bayesian methods

When the number of data points is small or the data is sparse, it is often the case that Bayesian methods are preferred over frequentist methods. In Bayesian methods one assumes a prior distribution of the peak height given marker-fluid pairs, which is updated with the available data through the likelihood. This results in the posterior distribution of the markers. Bayesian methods can perform significantly better than frequentist methods with sparse data, but it is important to choose the prior distribution well (an informed prior). With sparse data, the posterior will be defined mostly by the choice of prior. Thus, one could see the role of the data in a Bayesian study as a method to update or fine-tune an already adequate model.

The choice of an informed prior is worthy of a separate investigation in itself. We do not have a recommendation for one. The available literature could provide a suitable prior, but our investigation found only models developed for data where there are multiple peaks per marker, each corresponding to an allele length a. More about our investigation into the available literature can be found in Appendix A.

4.6 Further recommendations

In addition to our suggestions for further study, there are some aspects of our method which leave room for improvement. First and foremost, not all available data has been used in our study. More precisely, mixture modelling did not make use of the data from individual fluids. A future attempt could investigate how to incorporate that data, e.g. by generating artificial mixtures.

Secondly, in the data provided, each physical sample has been tested multiple times. This resulted in replicate data points whose values are very close (i.e. highly correlated). We chose not to aggregate the data due to the already low number of data points, but the violation of the independence assumption is clear. A future study should aggregate the replicate data points in a meaningful manner, perhaps by taking the average.

Lastly, the prevalence of situations where the maximum number of components has been reached due to singletons or outliers suggests that one could use robust fitting methods. One implementation

of Robust Gaussian Mixtures is provided in [1]. Further, some clusters did not seem to be well-modelled by Gaussian or Gamma distributions. A future study could look into incorporating mixtures of other distributions in addition to Gamma and Gaussian, e.g. the uniform distribution.

5 Conclusion

Gaussian and Gamma mixtures can constitute an effective way of modelling marker peaks. Their implementation is straightforward and the literature detailing their drawbacks (particularly drawbacks related to the use of the Expectation Maximization algorithm) is extensive. More importantly, they allow one to easily generate new data which would be beneficial if the Nederlands Forensisch Instituut decides to implement models which require a larger number of data points than the currently available datasets. On the other hand, there are clear drawbacks which need to be addressed by further research. The most pressing issues include finding a way to incorporate single fluid data (e.g. by generating artificial mixtures) using robust mixture modelling and obtaining confidence intervals for the fitted parameters.

In conclusion, we recommend using the mixture models approach for modelling genetic marker values of mRNA profiles. For a given fluid mixture and marker, we recommend consulting `NFI_Mixtures_Results.xlsx` in which we performed model selection among fluid pair mixtures.

References

[1] Package 'RGMM': Robust Mixture Model. `https://cran.r-project.org/web/packages/RGMM/RGMM.pdf`.

[2] Øyvind Bleka, Geir Storvik, and Peter Gill. EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. *Forensic Science International: Genetics*, 21:35–44, 2016.

[3] R. Cowell, Steffen Lauritzen, and Julia Mortera. A gamma model for DNA mixture analyses. *Bayesian Analysis*, 2:333–348, 06 2007.

[4] R. G. Cowell, T. Graversen, S. L. Lauritzen, and J. Mortera. Analysis of Forensic DNA Mixtures with Artefacts. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 64(1):1–48, 12 2014.

[5] Inc. Illumina. Methods for accurate computational decomposition of dna mixtures from contributors of unknown genotypes. `https://patents.google.com/patent/US11990208B2/en`, May 2024.

[6] RJF Ypma, PA Maaskant-van Wijk, R Gill, M Sjerps, and M Van den Berge. Calculating LRs for presence of body fluids from mRNA assay data in mixtures. *Forensic Science International: Genetics*, 52:102455, 2021.

Appendices

Our GitHub repository: `https://github.com/dewi-batista/consulting_NFI/tree/main`.

A Background - understanding the data

In order to understand what is to be modelled, it is important to understand the underlying process by which the data is acquired. This allows one to see the potential limitations of the model and study the available literature more effectively. In this endeavour, we have relied on [4].

Establishing terminology. An individual's genetic profile is encoded inside their chromosomes, which contain their DNA and mRNA profiles. An mRNA strand is essentially a very long ordered sequence of nucleic acids (or nucleobases): adenine (A), thymine/uracil (T/U)[4], cytosine (C) and guanine (G). "The DNA molecule has a double-helix structure where two strands of DNA are linked so that each nucleobase on one strand binds to a complementary nucleobase on the other strand: A is complementary to T, and C is complementary to G. Thus an allele is characterized by a sequence of letters from a single strand" [4].

The *marker* or *locus* is one specific position on the chromosome. An *allele* is a sequence of nucleotides, and it is fully characterized by the sequence of ATCG letters from a single strand. A marker can have different alleles at it. It can be said that this is essentially the genetic profile of an individual: the specific alleles that one has at various markers.

The PCR process. In a crime scene, the amount of genetic material left behind can be very small and insufficient for a proper analysis. Thankfully, there is a process that can reliably increase the amount of genetic material available through a replication procedure: the *polymerase chain reaction* process (PCR).

The process is very well explained in [4]. We quote the relevant section: "This involves adding primers and other biochemicals to the extract, and then subjecting it to a number of rapid heating and cooling cycles. Heating the extract has the effect of splitting the two complementary strands of DNA; the cooling phase then allows free-floating nucleotides to bind with these individual strands in such a way that the DNA is copied. By the action of repeated heating and cooling cycles, typically around 28 altogether, an initially small amount of DNA is amplified to an amount that is sufficiently large for quantification. [...] The amplification process is not 100% efficient, i.e. not every allele is copied in each cycle. This means that, if two distinct alleles in a marker are present in the extract in the same amount before amplification, they will occur in different amounts at the end of the PCR process".

There is a very important aspect regarding the PCR process. The allele word will have a length associated with it: the number of ATCG base pairs in the word itself plus the flanking regions[5]. It turns out that for each marker, it is not necessary to extract the exact words for each allele. To characterize the genetic profile at one marker, it suffices to look at the length of each allele: "quantifying a certain allele is equivalent to measuring how much DNA is present of a certain size" [4]. That is done by the process of *electrophoresis*.

[4] They have the same functionally. T is in DNA, U is in RNA. For simplicity, we will speak only of T.

[5] One can think of the flanking regions as sequences whose presence establishes where an allele word starts and stops.

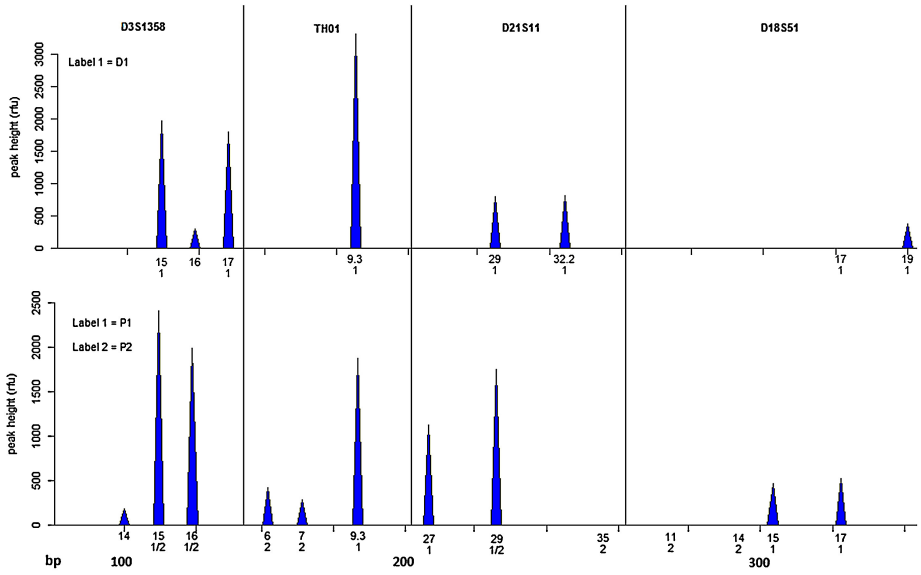

Figure 6: Electropherogram of two samples, compared at the same markers. The rectangles delimit different markers (e.g. TH01). On the horizontal axis, one can see the presence of alleles of various lengths. The vertical axis denotes the light intensity emitted, measured in rfu. Taken from [2]

For the purpose of this report, it is not necessary to explain electrophoresis in detail. Essentially, each flanking region can be attached a fluorescent dye; multiple colours can be used to distinguish across markers. It is possible to look at the light intensity (measured in *relative fluorescent units* - rfu) contribution from the alleles grouped by allele length. The higher the peak, the bigger the amount of DNA material of that specific length. The result is an electropherogram. An example can be seen in Figure 6. Overall, an electropherogram can be thought to encode the DNA profile of an individual.

Artifacts. PCR and electrophoresis are not perfect processes in the sense that they could suffer from the presence of artifacts. In the absence of artifacts, the presence of a peak of a certain length in the electropherogram after PCR indicates the existence of that allele in the initial sample. However, the following artifacts can complicate this analysis:

1. *Dropout*. When measuring the peak heights, there can be low level noise (e.g. due to the measuring apparatus) that could falsely signal the presence of "ghost" peaks. Because of that, a thresholding mechanism is imposed: peaks below a threshold level (typically 150 rfu) are considered noise and discarded. However, that means that alleles that are present in very small quantities (i.e. resulting in small peaks) can be falsely eliminated. This is called dropout.

2. *Dropin*. A sporadic contamination of the sample (either from the base or in the laboratory) can result in the presence of unexpected alleles. This is called dropin.

3. *Stuttering.* The PCR replication process is not perfect. An allele sequence could be inexactly copied, resulting in the copy having a different length than the initial allele. Due to the chain nature of PCR, this error will be propagated indefinitely. This is called stutter.

Samples that start with very little genetic material are especially prone to errors such as dropin or dropout.

A.1 Literature search and key ideas

From our literature survey, we found two papers that had potential to be useful for our project. The starting point was [2], provided by our client. Searching through its citations, we found [4]. While the model itself is a simplified version of [2], the merit of the paper is that it provides an in-depth explanation of electropherograms; it was crucial in our understanding of the data. The paper also explains at length the model in [2].

Afterwards we proceeded by searching for papers that have cited [2]. Most of the resulting literature was focused on applying the framework described in [2] on their data; for our purpose, they were irrelevant. One search result had caught our attention: [5]. The patent is a complex work spanning 55 pages. Due to time constraints, we could not study it in depth. However, the patent discusses different probabilistic models which could be used for future work.

Lastly, we have concentrated on searching through the main statistics journals for papers that could be relevant. The search engine used was Google Scholar and the parameters of the search were:

- Journals: relevant A* journals from this list: Annals of Applied Probability, Annals of Applied Statistics, Annals of Statistics, Annals of Probability, Biometrics, Biometrika.
- Mandatory key words: "mRNA probability forensic" or "probability forensic DNA".
- At least one word: "DNA statistical modelling PCR electrophoresis heights".

No interesting papers were found during this search.

Relevance of the literature. In the literature, the focus is on developing probabilistic models for the electropherograms introduced above. Crucially, the electropherograms model multiple peaks per marker, *identified by the allele length.* The data we were provided contains information only for a single peak per marker and it contains no information or distinction between allele lengths. As such, applying those models directly is not possible. However, there were some key ideas that we adapted for our approach.

In the model from [2] there are K contributors and M markers. At a given marker m, the observed alleles a form a vector $\boldsymbol{A}_m$. Each peak height $Y_{m,a,k}$ at marker m, from allele a and from a contributor k, is assumed to follow a Gamma distribution with parameters that depend on some shape and scale parameters μ and σ^2, but also on the proportions of total DNA amount $\pi_k \in [0, 1]$ coming from each contributor. There are many independence assumptions: contributions from different contributors are independent, genotypes (allele composition) are independent between markers, peak heights are conditionally independent given genotypes and model parameters, etc.

Based on the above, our approach relied on the following:

- We assumed independence wherever possible. In particular: of peak heights conditioned on the marker and fluid mixture type, between different measured samples (data points).

- The histogram of the peak heights conditioned on fluid type and marker are modelled as mixtures of multiple identifiable features. The contributions of each feature are independent.

  This approach is justified by the idea that a peak height for one individual is the sum of all peaks over the alleles (lengths). The total light intensity for one data point should be proportional to the total amount of alleles in the mRNA at that marker. Thus, individuals with similar genotypes should have peak heights around the same values, leading to a clustering in the histogram via genotype similarity. Mixtures are also a flexible model class.

- The distributions used for the mixture models are Gamma and Gaussian.

  Mixtures of Gaussians are well-studied, widely implemented and work well for clustering peak-like shapes. Gammas are a part of a distribution family with the same number of parameters as Gaussians (i.e. same risk of overfitting), but they can also model skewed shapes and exponential behaviour.

B Mixture Models

Mixture models are probabilistic models which help fit data generated from a combination of several underlying distributions (corresponding to different clusters). Each observation is assumed to be drawn from one of these clusters with some probability. Formally, an N-component mixture model is of the form

$$p(x) = \sum_{k=1}^{N} \pi_k f(x|\Theta_k)$$

where $\pi_1, \ldots, \pi_K$ are component weights such that $\sum_{k=1}^{N} \pi_k = 1$ and $f(x|\Theta_k)$ is the k^{th} component's density function.

For a Gaussian mixture model, the component densities $f(x|\theta_k)$ are Gaussian densities, so $\Theta_k = (\mu_k, \sigma_k^2)$ denotes the k^{th} component's mean and variance. Similarly, for a Gamma mixture model, the component densities $f(x|\Theta_k)$ are Gamma densities, so $\Theta_k = (\alpha_k, \theta_k)$ denotes the k^{th} component's shape and scale respectively.

C Conditional independence assumption

To examine the validity of the assumption of conditional independence of markers conditioned on fluid pairs, the Pearson correlations between all markers for each fluid pair were computed. As illustrated in Figure 7, relatively high correlations were found between markers. For example, one sees a correlation of 0.82 between SEMG1 and KLK3.

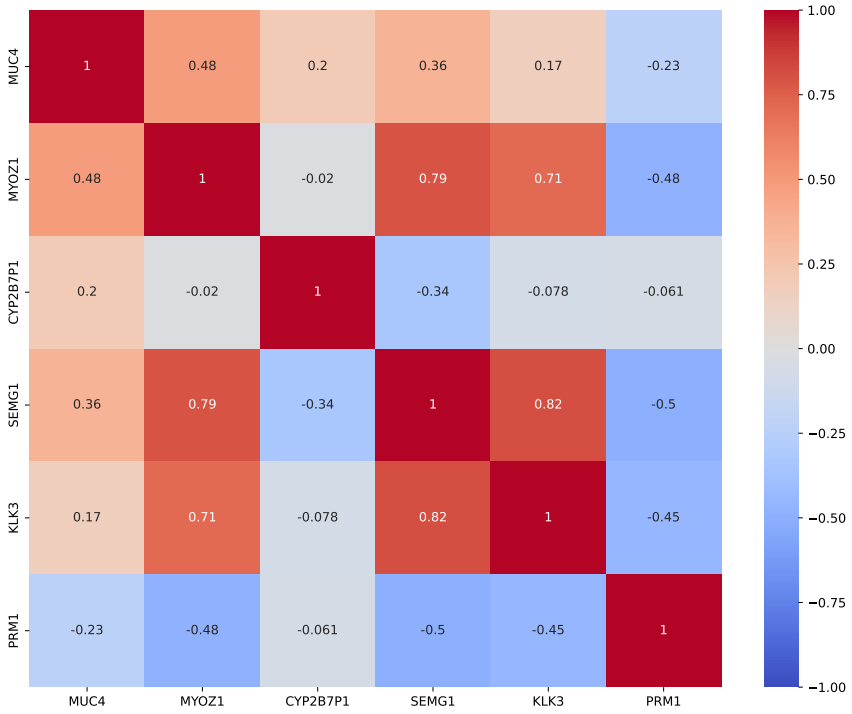

Figure 7: Pearson correlation matrix of all non-zero markers conditioned on semen (fertile) and vaginal mucosa.

In line with this finding, we plotted the 3D histograms pertaining to some of these highly-correlative markers, as illustrated in Figure 8

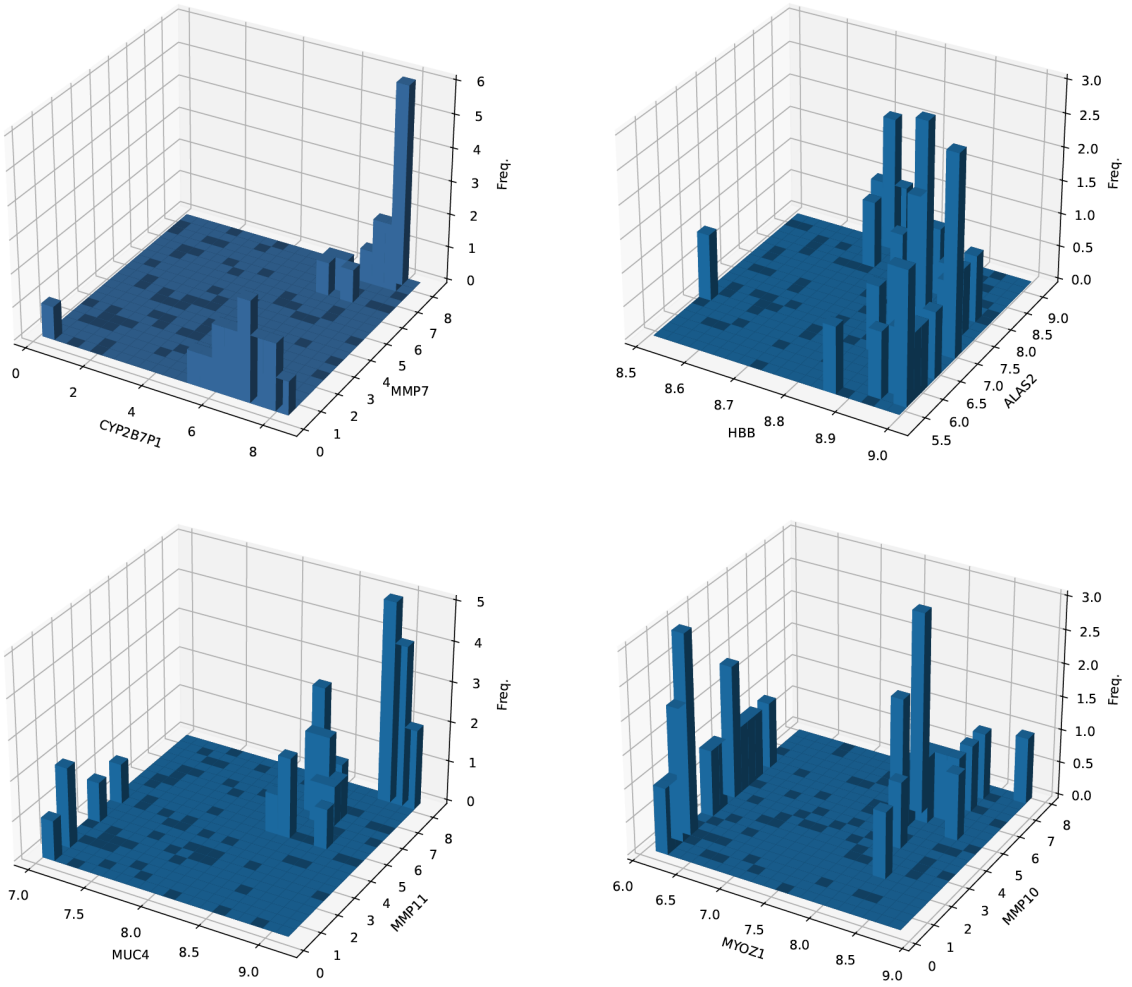

Figure 8: Histograms of highly-correlative marker pairs conditioned on blood and menstrual secretion.

What one observes in Figure 8 is generally observed throughout all such histogram plots of highly-correlative marker pairs. That is, the distributions observed look very difficult to fit. How to fit such distributions in a streamline manner is unclear to us. As such, sticking with the simplifying assumption of conditional independence is advised.

It is worth noting that this issue would perhaps be remedied by the acquisition of more data. That is, with far more data, these plots might make it immediately clear how one would fit the pairwise distribution of highly-correlative marker pairs in a streamline fashion, e.g. 2D mixture models.