Probabilistic Modelling of mRNA Electropherograms in Fluid Mixtures

Roberto Schinina, Andrei Secuiu and Dewi Batista

June 20, 2025

Problem definition

Ultimate goal

Determine body fluid presence in cases of sexual assault and violent crime.

How? mRNA profiling - infer presence of body fluids by analysing the expression of fluid-specific markers in a sample taken from the scene, victim(s), accused, etc.

Simplified example:

HBB	MUC4	PRM1	Blood	Saliva	Vaginal mucosa
5715	1750	3918	1	0	1

Ideally: Use marker values (HBB, MUC4 and PRM1) to infer presence of fluids (blood, saliva and vaginal mucosa).

Note: For us, 6 fluids and 15 markers are of interest.

Problem definition

Ultimate goal

Determine body fluid presence in cases of sexual assault and violent crime.

How? mRNA profiling - infer presence of body fluids by analysing the expression of fluid-specific markers in a sample taken from the scene, victim(s), accused, etc.

Simplified example:

HBB	MUC4	PRM1	Blood	Saliva	Vaginal mucosa
5715	1750	3918	1	0	1

Ideally: Use marker values (HBB, MUC4 and PRM1) to infer presence of fluids (blood, saliva and vaginal mucosa).

Note: For us, 6 fluids and 15 markers are of interest.

Problem definition

Ultimate goal

Determine body fluid presence in cases of sexual assault and violent crime.

How? mRNA profiling - infer presence of body fluids by analysing the expression of fluid-specific markers in a sample taken from the scene, victim(s), accused, etc.

Simplified example:

HBB	MUC4	PRM1	Blood	Saliva	Vaginal mucosa
5715	1750	3918	1	0	1

Ideally: Use marker values (HBB, MUC4 and PRM1) to infer presence of fluids (blood, saliva and vaginal mucosa).

Note: For us, 6 fluids and 15 markers are of interest.

Original work

Previous model [3]

Perform a likelihood ratio (LR) test to evaluate the hypothesis:

- H_0: at least one fluid of interest is present in sample.
- H_1: no fluids of interest are present.

Limitation: Which fluids are present?

Generative model desirable!

Fit distribution of markers conditioned on fluids present, i.e. fit

$$p(\text{markers}|\text{fluids present})$$

for each fluid combination present in given dataset.

Our work^{TM}

Dataset: 350 data points, ~ 50 per fluid combination $(\mathbf{f}_i, \mathbf{f}_j)$.

$\mathbf{m}_1$	$\mathbf{m}_2$	...	$\mathbf{m}_{15}$	$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$	$\mathbf{f}_4$	$\mathbf{f}_5$	$\mathbf{f}_6$
561	1105	...	2465	1	0	1	0	0	0
2821	6601	...	0	1	0	0	0	1	0
8911	0	...	1729	0	0	0	1	1	0

Table: Mixtures of (precisely two) fluids and their corresponding marker values.

Goal: Fit

$$p(\mathbf{m}_1, \ldots, \mathbf{m}_{15} | \mathbf{f}_i, \mathbf{f}_j)$$

for all fluid combinations $(\mathbf{f}_i, \mathbf{f}_j)$ in dataset.

How? Take inspiration from similar work in DNA profiling, e.g. assume independence of markers conditioned on fluids present [1], i.e.

$$p(\mathbf{m}_1, \ldots, \mathbf{m}_{15} | \mathbf{f}_i, \mathbf{f}_j) = p(\mathbf{m}_1 | \mathbf{f}_i, \mathbf{f}_j) \cdot \ldots \cdot p(\mathbf{m}_{15} | \mathbf{f}_i, \mathbf{f}_j).$$

Our work^{TM}

How do the individual markers conditioned on fluid combinations look?

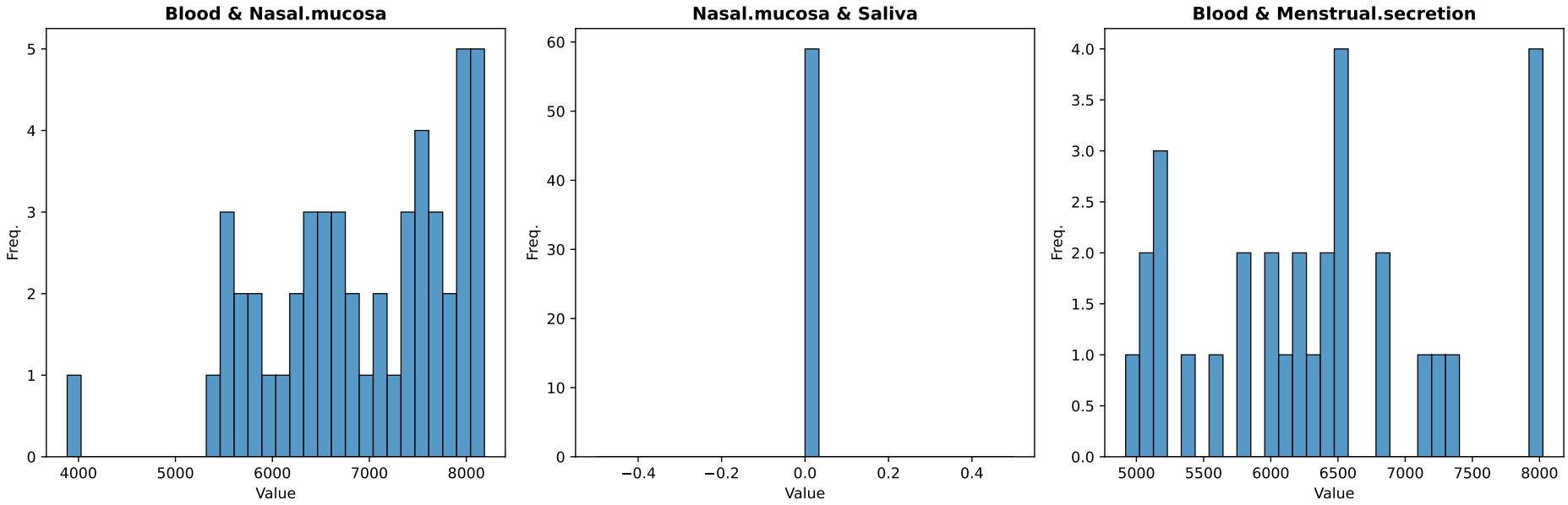

Figure: Histograms of HBB marker for three fluid combinations.

Mixtures

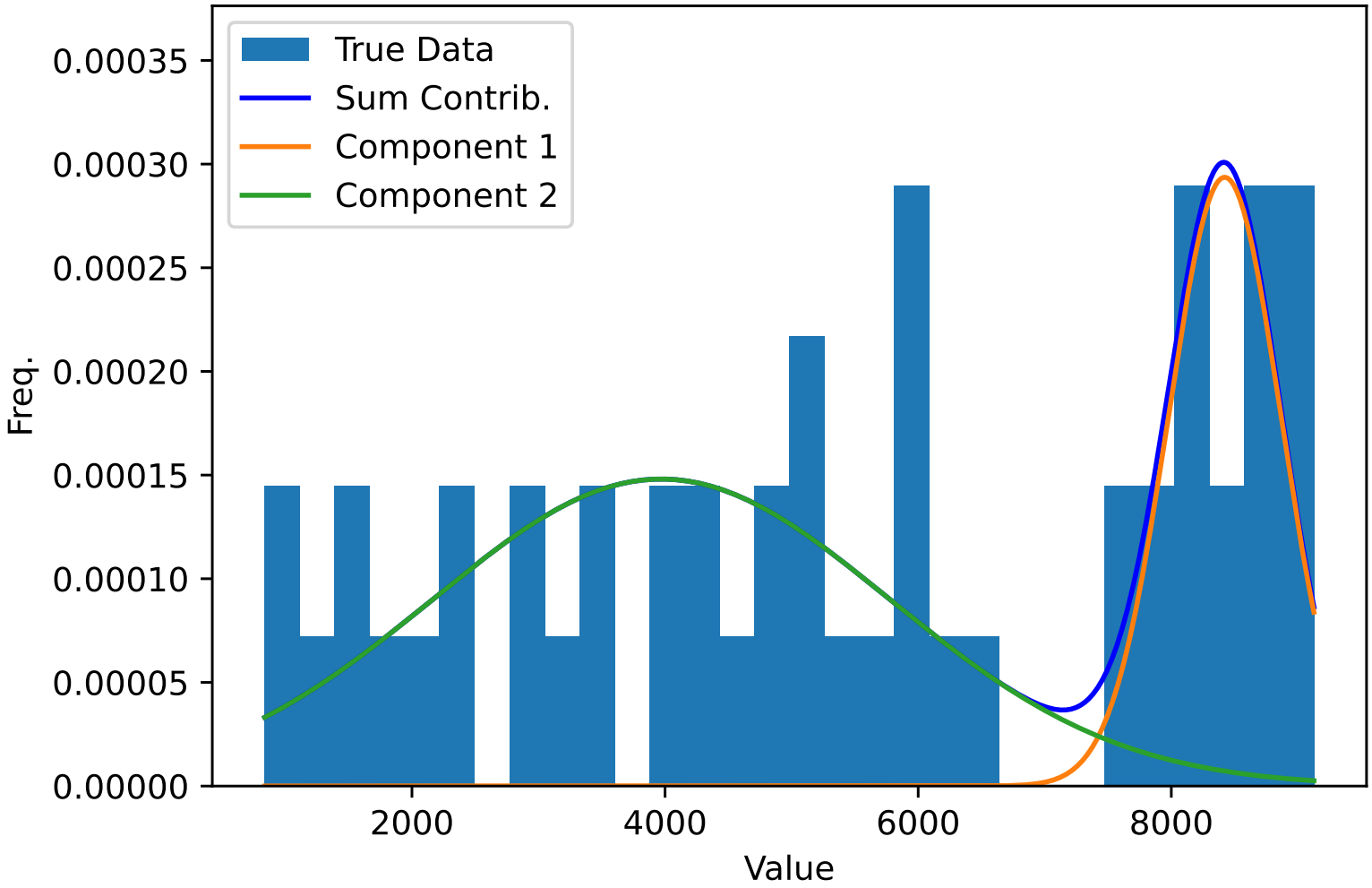

Figure: Gaussian mixture fit to ALAS2 conditioned on blood and nasal mucosa.

Mixtures

General form: For $\pi_1, \ldots, \pi_N \geq 0$ such that $\sum_{k=1}^{N} \pi_k = 1$,

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

Gaussian and **Gamma** mixtures considered:

- Inspired by literature.
- Straightforward implementation and tractable sampling.
- Trained using expectation-maximisation.

Model selection: BIC - reward good fit, punish over-complexity.

Evaluating generated data: Two-sample KS test on leave-out set.

Evaluating mixture-generated data

Test adequacy of data generation against leave-out set (1/3 of dataset).

Steps:

1. Train a model. Select the best via BIC.
2. Repeat 100 times:
   - Initialize new seed.
   - Generate new data points. Set values ≤ 150 to zero.
   - Perform two-sample KS test.
   - Record p-value.
3. Find lowest and median p-values.

Interpreting obtained p-values

Lower p-values imply the data sets come from different distributions!

Gaussian mixtures

General form: For $\pi_1, \ldots, \pi_N \geq 0$ such that $\sum_{k=1}^{N} \pi_k = 1$,

$$p(x) = \sum_{k=1}^{N} \pi_k \mathcal{N}(x|\mu_k, \sigma_k^2).$$

- Each marker-fluid pair modelled independently. Only mixture data used.
- $3N - 1$ parameters. Constraint $N \leq 10$.
- The implementation is in Python. It makes use of `sklearn.mixture.GaussianMixture`.
- Potential weakness: EM algorithm can find local minima $\rightarrow$ fitted curves can be good, but not optimal.

Gaussian mixtures - results

Summary of results:

Fluid vs. Marker	HBB	ALAS2	CD93	HTN3	STATH	BPIFA1	MUC4	MYOZ1	CYP2B7P1	MMP10	MMP7	MMP11	SEMG1	KLK3	PRM1
Semen.fertile+Vaginal.mucosa			No. There is a si				Yes. But it does	Yes	No. The number		No. There are or		No. Component	Yes	Yes
Saliva+Semen.fertile	Yes. Robust fittir			Yes	Yes. But it does		No. Data is scatt				No. Too few poir		Yes. Robust fittir	Yes	Yes
Saliva+Vaginal.mucosa			No. Data is scatt	Yes. There are t	Yes. The maxim		Yes	Yes. Potential to	Yes		No. Too few poir				No. A single poir
Blood+Nasal.mucosa	Yes. But it does	Yes. There are t	Yes. Potential to		Yes. Two maxim	Yes. Data is scat	Yes. Data is scat		No. Few points v				No. Too few poir		
Nasal.mucosa+Saliva	No. Too few poir		Yes. Data is scat	Yes	Yes. Potential to	Yes. Data is scat	Yes. Data is scat								
Vaginal.mucosa+Blood	Yes. A mixture w	Yes. Data is moc	No. Data is scatt				No. Maximum cc	No. Too few poir	No. Two single p						
Menstrual.secretion+Blood	No. Maximum cc	Yes. Data is moc	Yes. Potential to				Yes. Data is scat	Yes. Potential to	Yes. Data is scat	No. Too many si	Yes. There are t	Yes. Data is scat			

Figure: Summary of subjective fitting assessment.

Results are mixed -

- Green: Rarely "satisfactory". Can find nitpicks not covered by the fit.
- Yellow: A common theme: outliers of 1-2 data points modelled by their own component. Robust fitting could help.
- Red: The data does not support this type of modelling.
- Blue: Not enough data to get meaningful estimates.

Gaussian mixtures - some plots

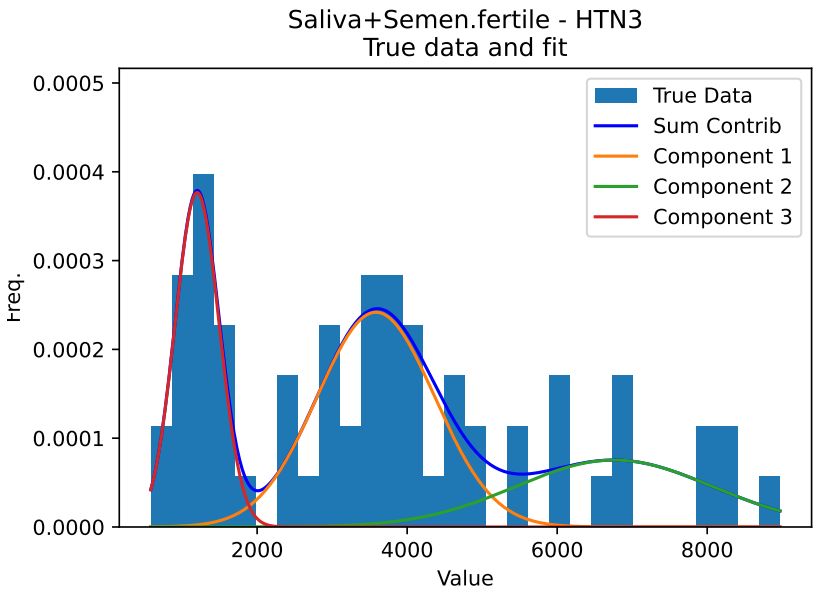

Figure: Green. A good-looking result

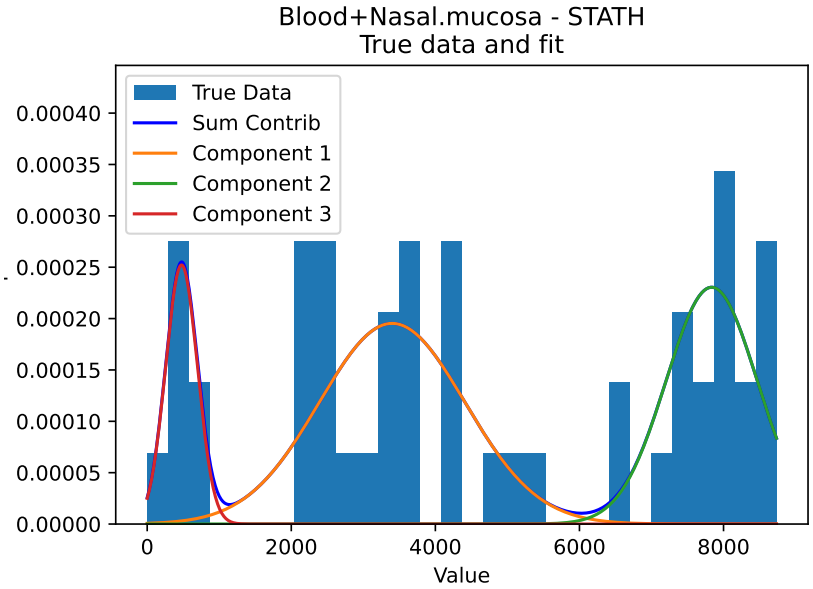

Figure: Green. Reasonable result, but component 1 does not seem to be Gaussian

Gaussian mixtures - some plots

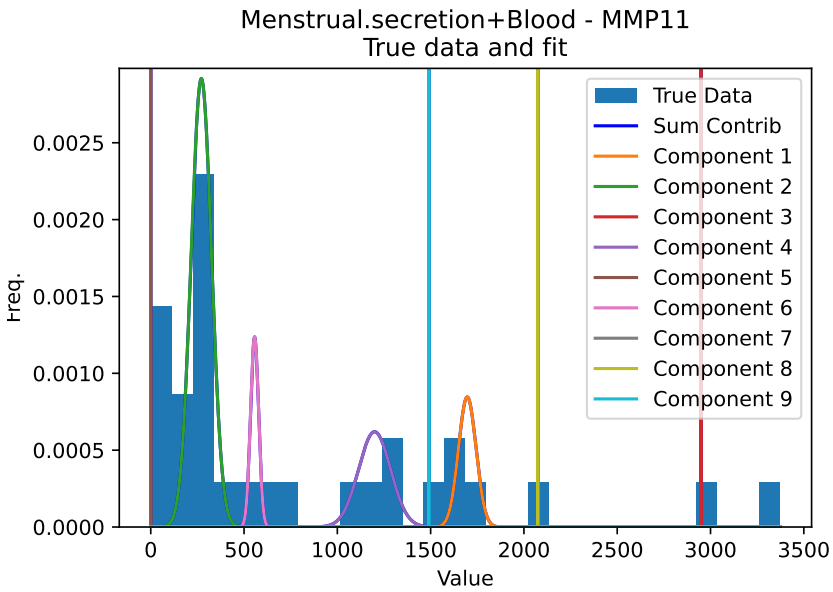

Figure: Green. Fit looks fine except for the outliers. Robust fitting?

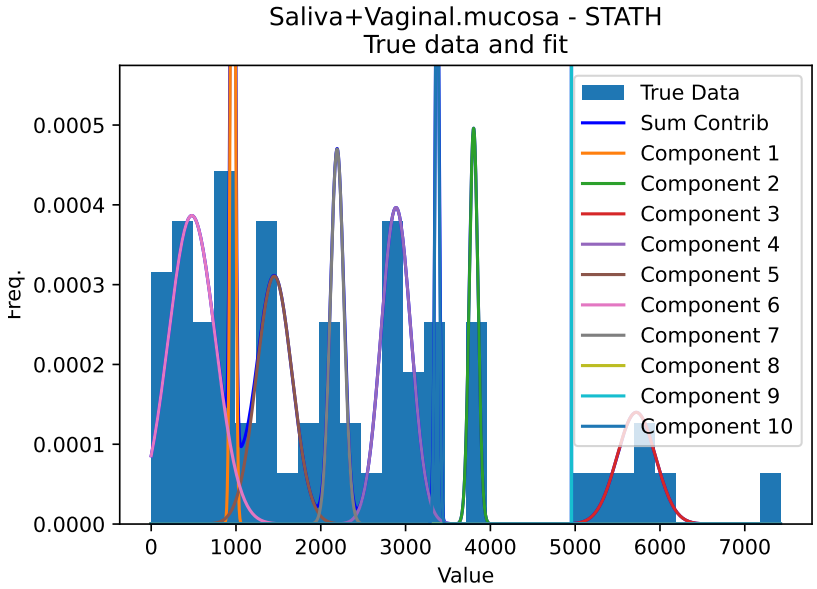

Figure: Green. Maximum number of components, but each seems to cover its base well

Gaussian mixtures - some plots

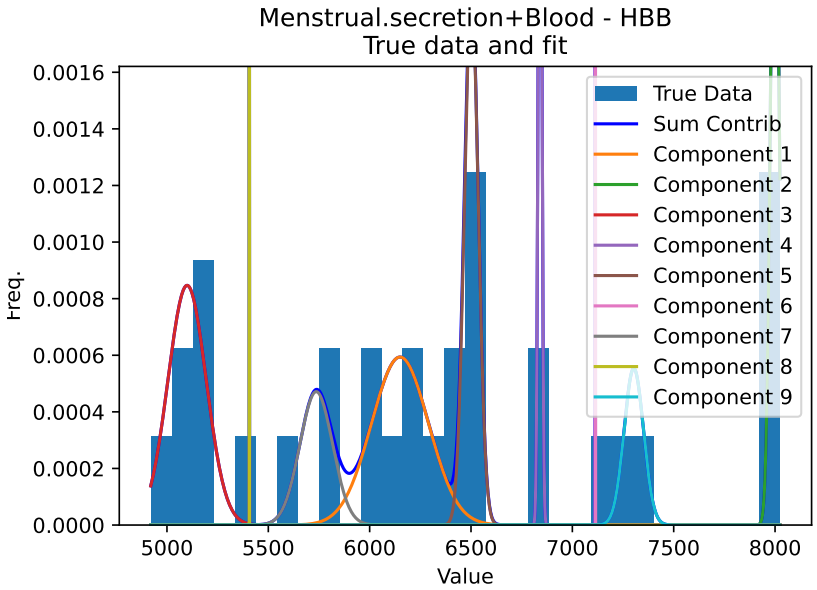

Figure: Yellow. Gaussian mixtures only seem inadequate. Robust fitting and other distributions can be considered

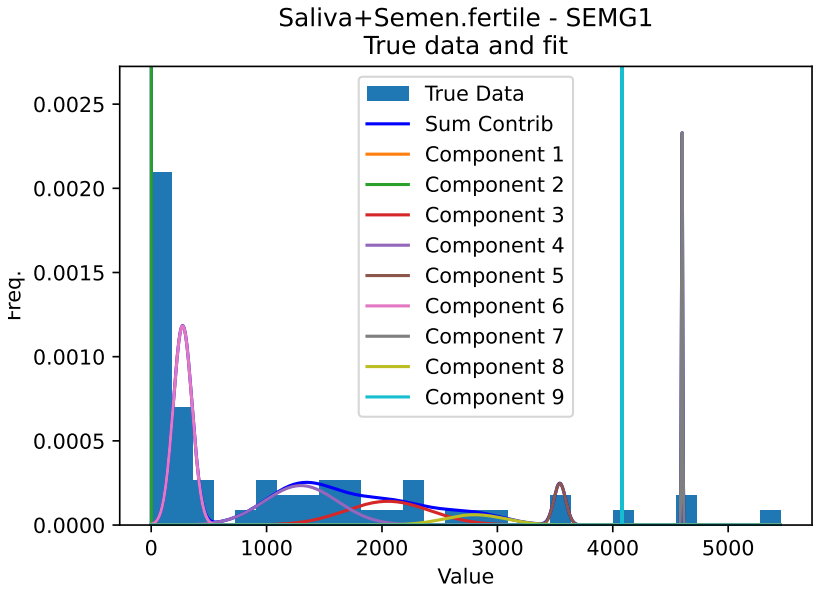

Figure: Yellow. Many singletons. Mixtures do not look perhect. EM in local minimum?

Gaussian mixtures - some plots

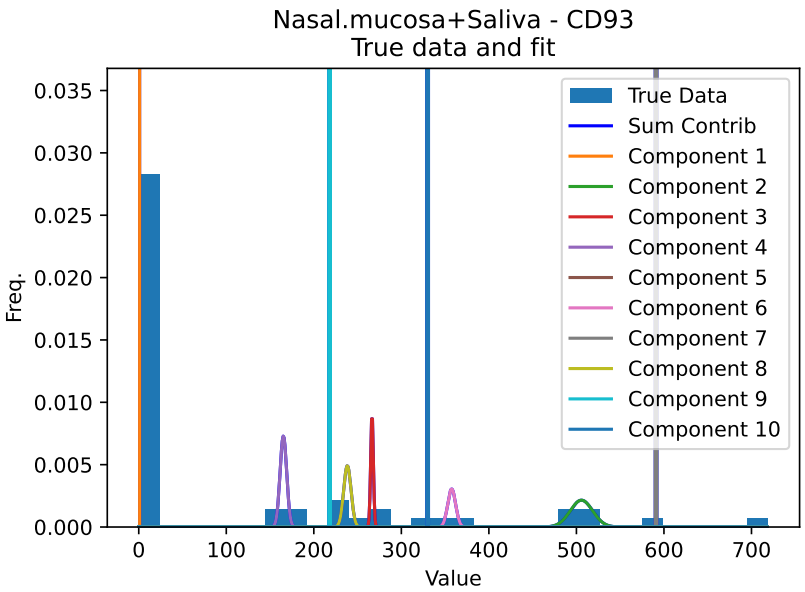

Figure: Red. Non-zero data is sparse. Hard to fit a meaningful model.

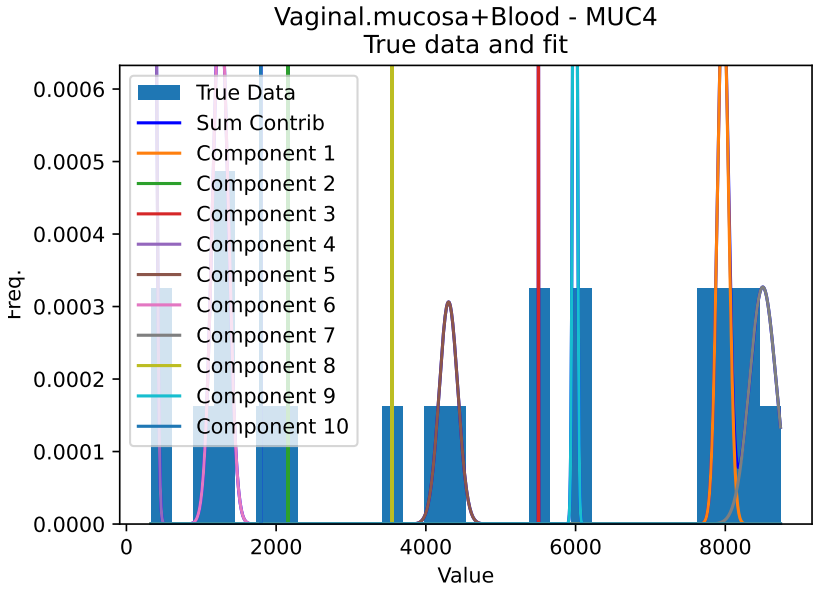

Figure: Red. Data clustered in sharp peaks. Mixtures do not seem to be an appropriate family.

Gaussian mixtures - discussion

- p-values alone are not a good indicator for goodness of fit:
  - High variation between generated data sets
  - Bad fits can have good p-values due to many zeros present
  - Good fits can have bad p-values (in some trials)
- The visual assessment is important, but also very subjective

Drawbacks:

- Data from individual fluids not used.
- Replicate data points from the same sample not aggregated.
- Many outliers present make the model overfit.
- Gaussians inappropriate for modelling some clusters.
- Literature models unused.

Gamma mixtures

General form: For $\pi_1, \ldots, \pi_N \geq 0$ such that $\sum_{k=1}^{N} \pi_k = 1$,

$$p(x) = \sum_{k=1}^{N} \pi_k \text{Gamma}(x|\alpha_k, \beta_k).$$

- Each marker-fluid pair is modeled independently. Only mixture data was used.
- $3N - 1$ degrees of freedom.
- Maximum number of components for model selection is 5.
- The implementation is in R. It makes use of `evmix.gammamixEM`

Gamma mixtures - some plots

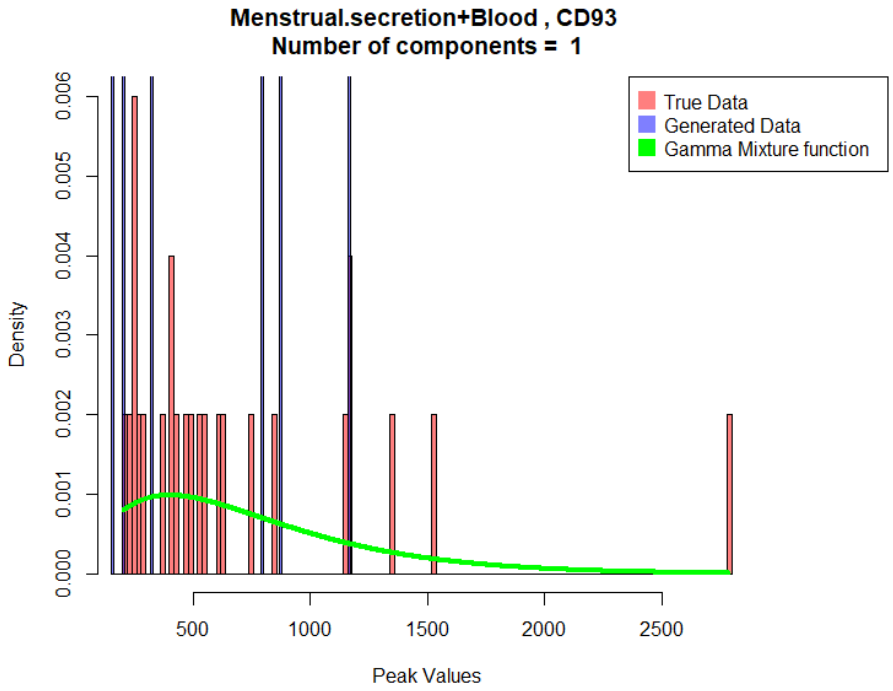

Figure: Gamma Mixture: 1 component

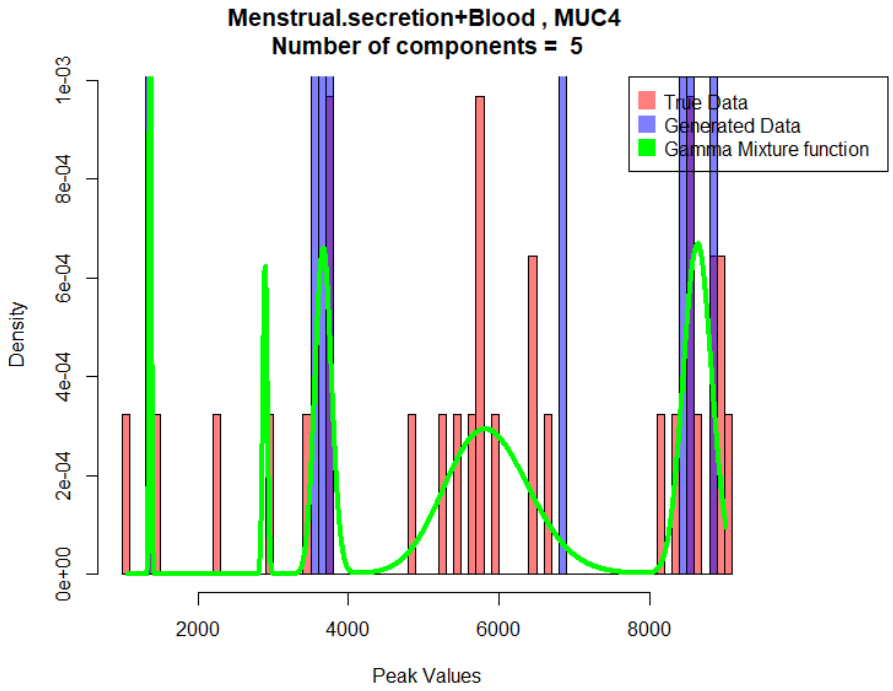

Figure: Gamma mixture: 5 components

Gamma mixtures - Discussion

- For $p-$values, same considerations as in the Gaussian.
- Methods based on gamma distributions used in the past for DNA mixture analysis [2].
- Drawback 1: the EM algorithm may fail to converge.
- Drawback 2: prone to overfitting but less than Gaussian.

Comparison: Semen.fertile+Vaginal.mucosa - MUC4

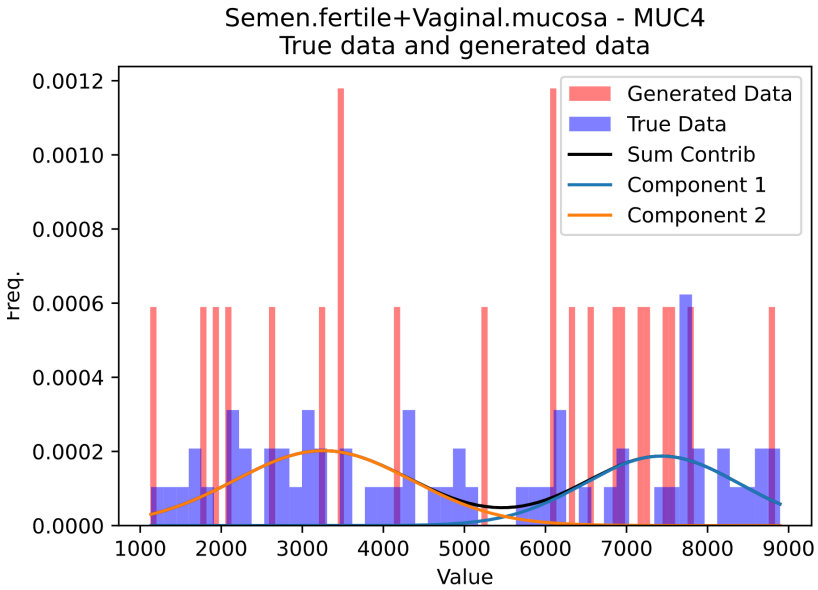

Figure: Gaussian mixture

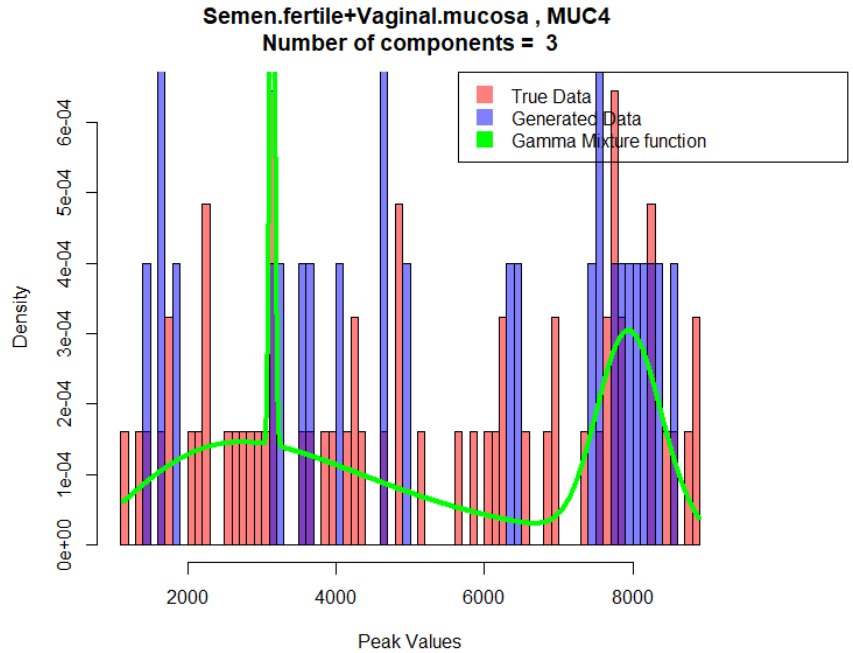

Figure: Gamma mixture

	#Components	BIC	p-value
Gaussian	2	737	0.872
Gamma	3	680	0.710

On the independence assumption

Many markers correlate!

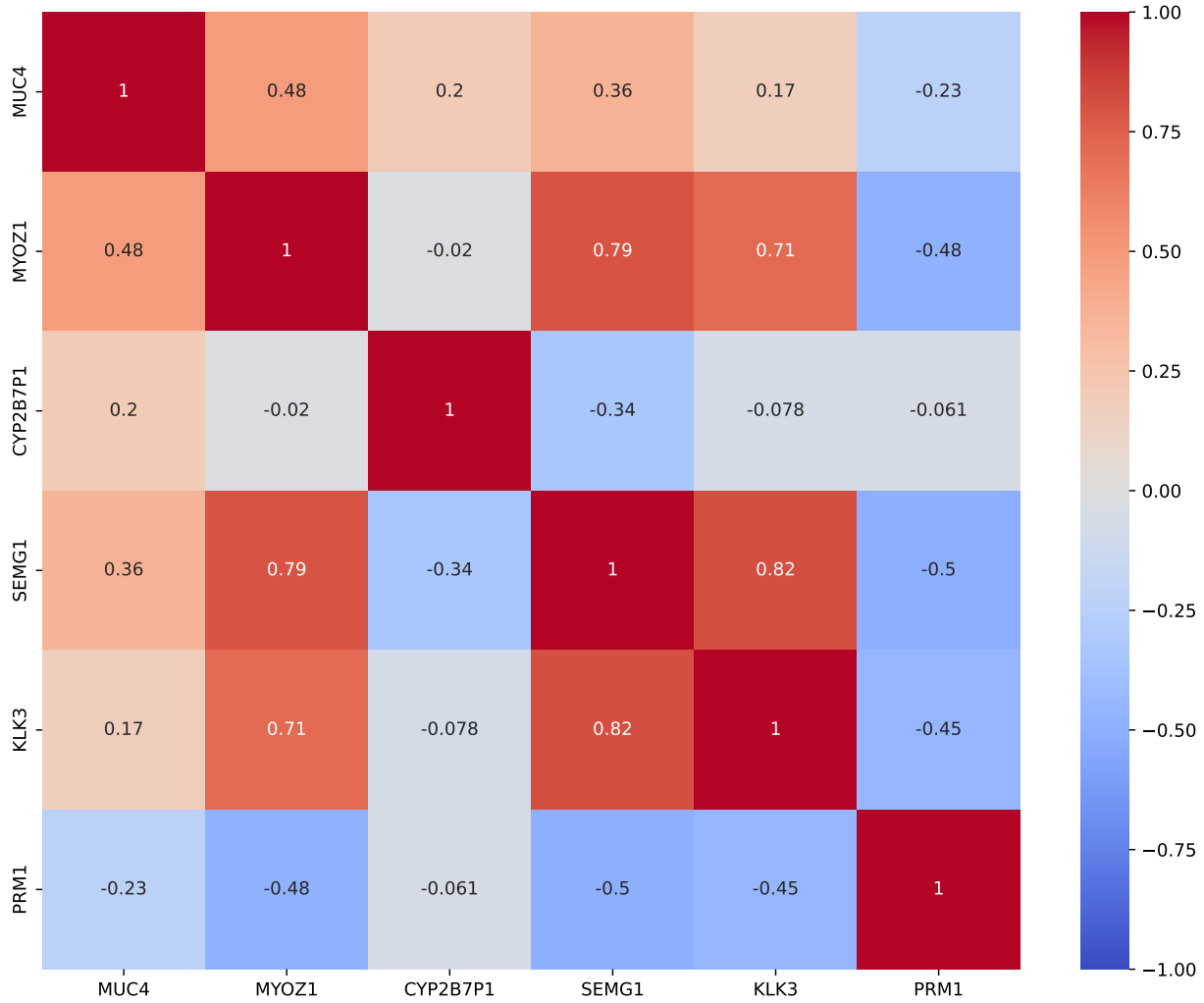

Figure: Correlation matrix of markers conditioned on semen (fertile) and vaginal mucosa.

On the independence assumption

But how do marker pairs look?

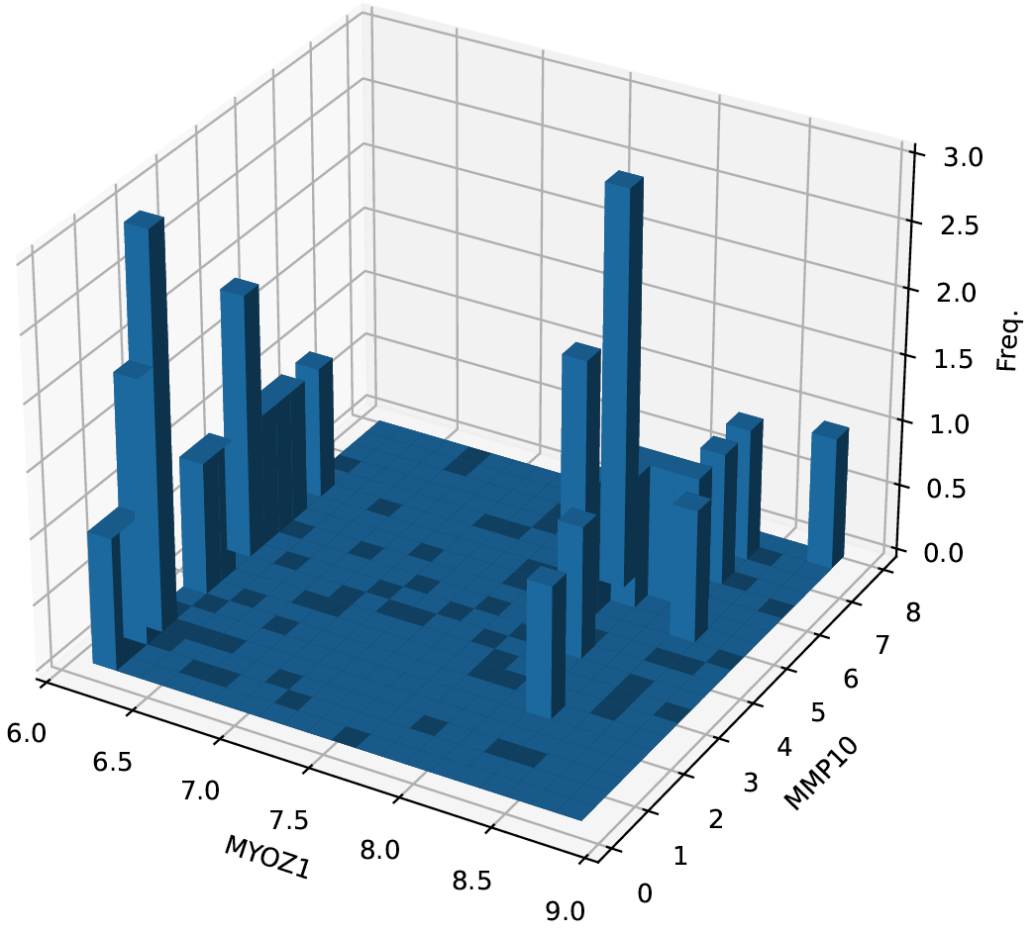

Figure: Histogram of MYOZ1 and MMP10 markers conditioned on blood and menstrual secretion.

Unclear how to approach. Stick to independence assumption.

Conclusions

Model summary:

- Mixtures can effectively model marker values, but not always.
- Implementation/adaptation straightforward.
- Allow generation of new data.

Drawbacks:

- No clear biological interpretation.
- Few data results in non-convergence of EM algorithm.
- Many singletons or outliers not well modelled.

Adaptations:

- Alternative mixture models could be more appropriate.
- Incorporate correlations.
- Acquire more data.

References

[1] Ø. Bleka, G. Storvik, and P. 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.

[2] R. Cowell, S. Lauritzen, and J. Mortera. A gamma model for dna mixture analyses. *Bayesian Analysis*, 2:333–348, 06 2007.

[3] R. Ypma, P. 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.