Vignette to *gammadnamix*: An open source R-package to handle continuous model for evaluating STR DNA mixtures evidence with artefacts.

Øyvind Bleka

September 12, 2014

Abstract

In this vignette we are introducing a forensic software, the package *gammadnamix*, which utilises the peak height information in STR DNA data and takes care of artefacts such as allele drop-out,drop-in and stutters. Our model assumes that the resulting peak heights from the PCR-process are gamma-distributed as presented by Cowell et al. [4] . We will incorporate this model into a Bayesian framework to make it available for users to incorporate their own prior information about allele drop-in and stutters. *gammadnamix* is a open-source free R-package which does model inference using maximum likelihood estimation or marginalized likelihood for any number of known and unknown contributors and for multiple replicates replicates.

1 Introduction

In the forensic field the main problem is to determine if a suspect could be a possible contributor to a trace. Todays technology makes it possible to apply DNA analysis to evaluate this determination which is often done by comparing two competing hypothesis: "The suspect contributes to the evidence" versus "An unknown individual from the population contributes to the evidence". A weight-of-evidence evaluation is based on a quantitative calculation based on some assumed model. The analysis may be limited (for STR DNA data) in that if a trace is contributed by several contributors, we will expect to see more amplified alleles in the DNA evidence. Also, the DNA extraction process in the laboratory may produce artefacts which changes the orignal DNA information to a slightly different data. Such artefacts could be allele drop-out which happens because parts of the DNA is not amplified enough to exceed a certain detection threshold. Another is the allele drop-in event which happens as a contamination inside the lab. The third effect is the stutter effect which causes alleles positioned one base-pair in front of contributing alleles to be amplified. Such influence of artefacts makes the trace less evidencial and uncertain in that it's not longer so rare that a random person from the population may "fit in" the evidence anymore (compared to ideal stains).

Weigh-of-evidence evaluation where peak heights or peak areas are incorporated requires a continuous mixture model which contains noisance parameters. Such parameters in the models are often apriori unknown and are either estimated using maksimum likelihood estimator or integrated out by turning into a Bayesian framework [6]. At the moment there exists several commercial softwares that are based on continuous models which takes peak heights into account for weighting of evidence [12],[9],[4]. A full bayesian evaluation of such models are very computational intensive, even for 2-person mixtures. Taylor et al. [12] and Perlin et al. [9] suggests models that are so complex (high dimensional) that numerical integral methods would be intractable to use. As an approximation, the integral is estimated by Markov Chain Monte Carlo (MCMC) simulations. However, the models considered in [2],[3], [4] and [11] are more simplified and could be calculated with a required accuracy using numerical integration [8].

We consider the model proposed by Cowell et al. [4] which handles different mixture proportion, peak height variation and stutter ratio as noisance parameters. They evaluated the weight of evidence frequentisticly based on the maximum log-likelihood properties. In our software we extend to Bayesian inference by integrating out the noisance parameters in the model such that the weight of evidence formula becomes the Bayes Factor[1]. Both the frequentistic inference and the Bayesian inference for any given hypotheses are implemented in the open-source R-package *gammadnamix*.

2 The continuous model

For a particular locus we observe alleles $\mathbf{A} = (A_1, ..., A_J) \subseteq \mathbb{A}$ with corresponding peak heights $\mathbf{Y} = (Y_1, ..., Y_J)$ where J is number of observed alleles. We let the outcome of possible alleles be given as $\mathbb{A}$. We introduce an individual $k \in \{1, .., C\}$ to have genotype $g_k = (a_k/b_k)$. By combining the genotypes for the C contributors we obtain the genotype vector $\mathbf{g} = (g_1, ..., g_C)$.

2.1 The allele peak height distribution

Cowell et al. [4] assumed that the peak height at allele A_j indiced at $j \in \{1, ..., J\}$ for contributor k is distributed as

$$Y_{jk} \sim_{iid.} \text{gamma}\left(\rho n_{jk} m_k, \tau\right)$$

where ρ is proportional to the total amount of DNA in the mixture before amplification and τ is the scale parameter. For a given genotype g_k, $n_{jk} = \mathbb{I}(A_j = a_k) + \mathbb{I}(A_j = b_k)$ gives contribution amount for contributor k to allele A_j. Further, *iid.* means that the observed peak heights are drawn independently for each contributor and allele. The mixture proportion parameter $\mathbf{m} = (m_1, ..., m_C)$ is defined on $simplex(\mathbf{1}_C)$ (i.e. $\mathbf{m} \in [0, 1]^C$ with $m_C = \sum_{k=1}^{C-1} m_k = 1$.

By summing the independent peak heights over each contributors the observed

[1] The Bayes inference analogy to Likelihood Ratio is Bayes Factor

peak heights is given to have the following distribution

$$Y_j \sim_{iid.} \text{gamma}\left(\rho \mathbf{m}^T \mathbf{n}_j, \tau\right)$$

where $\mathbf{n}_j = (n_{j1}, ..., n_{jC})$ is the vector giving the contribution amount for each contributor.

2.1.1 Model properties

The peak height mean and variance of the model is given as $\mu_j = E[Y_j] = \rho\tau\mathbf{m}^T\mathbf{n}_j$ and $Var[Y_j] = \rho\tau^2\mathbf{m}^T\mathbf{n}_j$. Hence the coeffecient of variation is given as $CV[Y_j] = \rho^{-\frac{1}{2}}$ which is a strongly decaying function down to 200rfu and hence it will model the low-template DNA effect in that smaller peak heights has greater variation than larger peak heights (NEED REF).

2.1.2 Reparameterization

As in [4] the parameters of the model are reparameterizated such that that the parameters orthogonally spanned and easier to interpret directly. By requiring $\mu = \rho\tau \geq 0$ to be the expected mean peak height (for a single heterozygote contributor) and $\sigma = \rho^{-\frac{1}{2}} \geq 0$ to be the coefficent of variance, we have that $\rho = \sigma^{-2}$ and $\tau = \mu\sigma^2$. However, when presenting the gamma model we will still use ρ and τ for simplificity.

2.2 Genotype probabilities

The continuous model requires that we make a suggestion of a combined genotype profiles $\mathbf{g}$ as a possible contribution to the evidence. Often a hypothesis assumes that one or more profile componononents in $\mathbf{g}$ are unknown. When such genotype profiles are unknown, it is reasonable to assign these components to a discrete random variable set, such that a given combined genotype profile has probability $p(\mathbf{g})$. In our software we will use the point estimates of the population allele-frequencies to determine $p(\mathbf{g})$. However we will extend the model of $p(\mathbf{g})$ to include the scalar f_{st} to include the possibility that the components of $\mathbf{g}$ are a sub-population of the population which the point estimates are based on [5].

With the Hardy-Weinberg assumption ($f_{st} = 0$), we have that

$$p(g_k) = \begin{cases} 2^{\mathbb{I}(a_k \neq b_k)} P(a_k)P(b_k) & \text{if k is unknown contributor} \\ 1 & \text{if k is known contributor} \end{cases}$$

The correction formula when $f_{st} > 0$ is given as:

$$p(g_k) = \begin{cases} 2^{\mathbb{I}(a_k \neq b_k)} \prod_{j \in \{a_k, b_k\}} \left(\frac{u_j f_{st} + (1-f_{st})P(j)}{1+(v-1)f_{st}}\right) & \text{if k is unknown contributor} \\ 1 & \text{if k is known contributor} \end{cases}$$

where $P(j)$ is allele frequency of allele $j \in \{a_k, b_k\}$, u_j denotes previously number of sampled alleles of allele j and v denotes previously number of total sampled alleles.

By assuming that each contributors are independent of each others (given the f_{st} relation) we have that $p(\mathbf{g}) = \prod_{k=1}^{C} p(g_k)$.

2.3 Modeling artefacts

In this subsection we follow the models provided by Cowell et al. [4] to handle artefacts as stutter and drop-out. To handle drop-in events, we provide our own method similar to Puch-Solis [10]. We also remind the reader that the expected model peak height of allele j is denoted as $\mu_j = E[Y_j]$ as given in earlier sub-section.

2.3.1 Incorporate stutter

The stutter ratio parameter ξ is modelled to be the ratio of contributed peak heights from the contribution of the $j+1$ alleles. The non-stuttered and stuttered peak heights are independently modelled as gamma $\big((1-\xi)\rho\mathbf{m}^T\mathbf{n}_j, \tau\big)$ and gamma $\big(\xi\rho\mathbf{m}^T\mathbf{n}_{j+1}, \tau\big)$ respectively such that the accumulated peak heights on allele j is indepedent distributed as

$$Y_j \sim \text{gamma}\left(\rho\mathbf{m}^T\big((1-\xi)\mathbf{n}_j + \xi\mathbf{n}_{j+1}\big), \tau\right) = f_j(y_j|\mathbf{g}, \boldsymbol{\theta})$$

where f_j becomes the density function of the peak height given the set of genotype profiles $\mathbf{g}$ and the reparameterized model parameters is given as $\boldsymbol{\theta} = (\mathbf{m}, \mu, \sigma, \xi)$.

2.3.2 Handling allele drop-out

Sometimes, the peak heights of a contributing allele (i.e. the allele of a true donor) is not seen in the evidence because the amplification produces a rfu peak height below a certain detection threshold T. We define this as an allele drop-out. For a given genotype $\mathbf{g}$ in our model, we have the set of non-observed alleles given in $\mathbf{g}$ as

$$\Omega = \{A_j \in \mathbf{g} : \mu_j > 0 \cap Y_j < T\}$$

If Ω is empty (i.e. we have no allele drop-out), $p(\Omega|\mathbf{g}, \boldsymbol{\theta}) = 1$. Otherwise we incorporate the allele drop-out event in the model by including the probability factor

$$p(\Omega|\mathbf{g}, \boldsymbol{\theta}) = \prod_{A_j \in \Omega} Pr(y_j < T|\mathbf{g}, \boldsymbol{\theta}) = \prod_{A_j \in \Omega} \int_0^T f_j(x|\boldsymbol{\theta}, \mathbf{g})dx$$

2.3.3 Handling allele drop-in

The event of an allele drop-in is defined to be a non-explained peak height which is not contributed by any contributors nor as a stuttered peak height. The dropped in allele is assumed to originate from a contaminating cell from the population with probability p_C. Such contaminations are often identified by running "negative controls" which aims to detect contaminating alleles. By summarising the negative controls for a lab it is possible to both estimate the probability of drop-in, p_C, and infer a drop-in density function as a function of allele peak height Y given as $Pr(Y = y|\text{Drop-in}) = h(y)$. Puch-Solis [10] found that h fitted a gamma distribution well for SGM with threshold given as 25. However we assume that the detection threshold is given above $T = 50$. We found that a shifted exponential density function with a rate parameter λ

starting from threshold T was a reasonable choice for h. We will further assume λ to be prior knowledge to the model such that $h(y) = h(y|\lambda)$.

For a given genotype combination $\mathbf{g}$, we define the allele drop-in set as

$$\Psi = \{A_j \in \mathbf{A} : (\mu_j = 0 \cap Y_j \geq T)\}$$

which is the set of non-explained alleles (a peak height is observed for an allele which the model assumes no contribution to).

We incorporate the drop-in peak height information to the likelihood with the factor

$$p(\Psi|\mathbf{g}) = \begin{cases} \prod_{A_j \in \Psi} (p_C P(A_j) h(Y_j|\lambda)) & \text{if } \Psi \text{ is non-empty} \\ (1 - p_C) & \text{if } \Psi \text{ is empty} \end{cases}$$

where

$$h(y|\lambda) = \lambda \exp^{-\lambda(y-T)} \mathbb{I}(y \in [T, \infty))$$

is the shifted exponential density function with parameter λ.

2.3.4 Handling no-contributed markers

If the stain is low-template, there may be some markers which have no observed peak heights. The explanation for this will be that at least one peak height is below the detection threshold. This is a part of the observation and should be taken into account to the model to downjust the parameter(s) connected to the amount of dna.

2.4 Apriori distribution for model parameters

A Bayesian framework requires prior distribution to the unknown parameters in the model to be incorporated. Prior distributions for the parameters of $\boldsymbol{\theta}$ are given as

$$p(\boldsymbol{\theta}) = p(\mathbf{m})p(\tau)p(\xi)$$

where we specify non-informative prior on each parameters

$$\begin{aligned} p(\mathbf{m}) &= Dirichlet(\mathbf{1}_C) \\ p(\mu) &= \mathbb{I}(\mu \in [0, \mu_1]) \\ p(\sigma) &= \mathbb{I}(\sigma \in [0, \sigma_1]) \\ p(\xi) &= \mathbb{I}(\xi \in [0, min(\xi_1, 1)]) \end{aligned}$$

The posterior distribution of μ and σ should be investigated closer in order to find a suitable values of μ_1 and σ_1. The prior density of the stutter ratio can be choosen flexible in the routine such that it may be infered by stutter information from the laboratory (i.e. choosing $p(\xi) \approx p(\xi|\text{stutter-data})$).

3 Inference of the model

3.1 One replicate

A specified hypothesis H defines a set $\mathbb{Q}$ which consists of all possible combined genotypes $\mathbf{g}$ which satisfies H. Let the mixture evidence $E = (\mathbf{A}, \mathbf{Y})$ be the observed data. When threating I loci simultaneously we let the vectorized alleles, peak heights and the combined genotype outcome be denoted as $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_I)$, $\mathbf{A} = (\mathbf{A}_1, ..., \mathbf{A}_I)$ and $\mathbb{Q} = (\mathbb{Q}_1, ..., \mathbb{Q}_I)$. We are marginalizing out all combined genotypes (latent variables) with respect to the possible outcome, such that simultaneous over all loci, the likelihood function of the model parameters ($\boldsymbol{\theta} = (\mathbf{m}, \mu, \sigma, \xi)$) becomes

$$L(\boldsymbol{\theta}|E,H) = \prod_{i=1}^{I} L_i(\boldsymbol{\theta}|E_i,H) = \prod_{i=1}^{I} \left(\sum_{\mathbf{g}_i \in \mathbb{Q}_i} p(E_i|\mathbf{g}_i,\boldsymbol{\theta}) p(\mathbf{g}_i|f_{st},H) \right)$$

For a given locus i we have that the probability of the observations is given by the expression

$$p(E|\mathbf{g},\boldsymbol{\theta}) = p(\Psi|\mathbf{g}) p(\Omega|\mathbf{g},\boldsymbol{\theta}) \prod_{A_j \in \chi} f_j(y_j|\mathbf{g},\boldsymbol{\theta})$$

where $\chi = \mathbf{A} \setminus \Psi = \{A_j \in \mathbf{A} : (\mu_j > 0 \cap Y_j \geq T)\}$ is the observed set of alleles where we expect a contributing peak height. The peak height density function f_j, drop-in density function $p(\Psi|\mathbf{g})$ and drop-out density function $p(\Omega|\mathbf{g},\boldsymbol{\theta}))$ were all defined in the previous section.

3.2 Multiple replicates

Assume that a sample has been replicated such that each repliacted sample is assumed to contain the same contributors and satisfy the same model assumptions. Consider the number of replicates as S, and let the observed replicated evidences to be given as $\mathbf{E} = (E^{(1)}, ..., E^{(S)})$. THen the formula above is extended to

$$L(\boldsymbol{\theta}|\mathbf{E},H) = \prod_{i=1}^{I} \left(\sum_{\mathbf{g}_i \in \mathbb{Q}_i} p(\mathbf{g}_i|f_{st},H) \prod_{s=1}^{S} p(E_i^{(s)}|\mathbf{g}_i,\boldsymbol{\theta}) \right)$$

by assuming the conditional replicated observations are independent of each others.

3.3 Reducing the mixture proportion space

Note that $\mathbb{Q} = \{\mathbb{Q}_1, ..., \mathbb{Q}_I\}$ must contain all possible genotype combinations, also symmetrical possibilities for unknown genotypes. The reason for this is that when we consider multiple loci simultanously, the order of the combined genotype profiles is relevant when scaling with the global mixture proportion parameter. However, we will now show how the symmetry may support to reduce the space of the mixture proportions.

Denote the parameter space of the mixture proportions $\mathbf{m}$ as M where $M =$

simplex$(\mathbf{1}_C)$ (i.e. $\mathbf{m} \in [0,1]^C$ with restriction$m_C = 1 - \sum_{k=1}^{C-1} m_k \geq 0$). When there is at least 2 unknown contributors in a hypothesis, there is possible to reduce the space M. Let $\mathbf{m} = \{\mathbf{m}_K, \mathbf{m}_U\} \in M$ such that $\mathbf{m}_K$ is the vector for known contributors, while $\mathbf{m}_U$ is the vector for unknown contributors. By requiring the elements in $\mathbf{m}_U$ to be in decreasing order, changing M to $\widetilde{M}$, this avoids a multimodal posterior density in the space of $\mathbf{m}$ when the hypothesis includes at least 2 unknowns.

3.4 Estimating mode of the posterior distribution (Frequentistic inference)

Instead of working in the restricted space of θ we transform[2] to the unrestricted space of ϕ and make unrestricted optimization of the posterior distribution

$$p(\boldsymbol{\theta}(\boldsymbol{\phi})|E,H) \propto L(\boldsymbol{\theta}(\boldsymbol{\phi})|E,H)p(\boldsymbol{\theta}(\boldsymbol{\phi}))$$

to find the restricted maximum argument $\boldsymbol{\theta}^*$. The reporting likelihood for the hypothesis is then given as $L(\boldsymbol{\theta}^*|E,H)p(\boldsymbol{\theta}^*)$. Because the uncertainty of $\boldsymbol{\theta}$, the routine also returns the posterior covariance structure Σ on $\boldsymbol{\theta}^*$ using the delta method approximation[3]. Σ can also further be used as the covariance of the proposal distribution when running Random Walk Metropolis Hastings in the next section to draw samples from $p(\boldsymbol{\theta}|E,H)$. In next section we also see how it can also be used together with the maximum likelihood value to make an Laplace approximation to the marginalized likelihood of the evidence.

3.5 The marginalized likelihood of evidence (Bayesian inference)

The reporting likelihood in previous sub-section is depending on the point estimation given as the mode of the posterior distribution, and hence it does not take into account the uncertainty of the model parameters $\boldsymbol{\theta}$. We let $\boldsymbol{\theta} = (\mathbf{m}, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_2 = (\mu, \sigma, \xi)$ excludes the mixture proportion parameters. Since we are working in a Baysian framework, the marginalized likelihood of the given hypothesis H can be calculated by integrating out the parameters as

$$L(H|E) = \int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|E,H)p(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\mathbf{m}\in\mathcal{M}} p(\mathbf{m}|E,H)d\mathbf{m} \tag{1}$$

where

$$p(\mathbf{m}|E,H) = \int_{\boldsymbol{\theta}_2} L(\mathbf{m}, \boldsymbol{\theta}_2|E,H)p(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2 p(\mathbf{m})$$

Note here that $\mathcal{M} = $ *simplex*$(\mathbf{1}_C)$ (i.e. $\mathbf{m} \in [0,1]^C$ with restriction $m_C = 1 - \sum_{k=1}^{C-1} m_k \geq 0$).

[2] See Reparametrization Appendix A

[3] To get the full covariance structure of θ we used the derivings in Appendix A

3.5.1 Integral over a simplex

To integrate with respect to $\mathcal{M}$ we need to take the simplex restrictions into account. We have that since $m_C = 1 - \sum_{k=1}^{C-1} m_k$, we only need to consider the integral over the variables $(m_1, ..., m_{C-1})$. With the restriction $\sum_{k=1}^{C-1} m_k \leq 1$ we get the marginalized likelihood as

$$L(H|E) = \int_{m_1 \in [0,1]} \int_{m_2 \in [0,1-m_1]} \cdots \int_{m_{C-1} \in [0,1-\sum_{k=1}^{C-2} m_k]} p(\mathbf{m}|E,H) dm_{C-1} \cdots dm_1$$

3.5.2 Reduction of mixture proportion space

Recall that we may reduce the space of $\mathcal{M}$ to $\tilde{\mathcal{M}}$ if the hypothesis includes at least 2 unknowns. The marginalized likelihood can then be written as

$$L(H|E) = U! \int_{\mathbf{m} \in \tilde{\mathcal{M}}} p(\mathbf{m}|E,H) d\mathbf{m}$$

where $U = |\mathbf{m}_U|$, the number of unknown contributors. Reducing the space from M to $\tilde{\mathcal{M}}$ is efficient since we reduce the space of $\mathbf{m}_U$ by require the mixture proportions of the unknown contributors to be decreasing ordered.

Estimation of these integrals are done by numerical multidimensional integration.

3.5.3 Laplace approximation of the marginalized likelihood

Consider $\boldsymbol{\theta}^*$ to be the estimated mode and $\Sigma(\boldsymbol{\theta}^*)$ the inverse negative hessian from previous section. When there is sufficient large enough amount of data, the likelihood will be approximately normal distributed (CLT). A second order taylor expansion (see Appendix A.4 for technical details) gives a rough approximation to the marginalized likelihood as

$$\hat{L}(H|E) = U!(2\pi)^{\frac{p}{2}} |\Sigma(\boldsymbol{\theta}^*)|^{\frac{1}{2}} L(\boldsymbol{\theta}^*|E,H) p(\boldsymbol{\theta}^*)$$

where U is number of unknowns in hypothesis U and p is number of paramaters in the model ($p = C + 2$). If at least 2 unknowns are considered in H, the posterior of the mixture proportions will be multimodal and hence we must scale with number of possible ordered outcomes for the unknown contributors.

3.5.4 Numerical integration methods

The marginalized likelihood in equation 1 requires to calculate an integral over the space of the noisance parameters $\boldsymbol{\theta}$. Steven G. Johnson implemented the R-package *cubature* to do adaptive multidimensional integration based on algorithms in [7] and [1]. A relative error requirement is given as an input for the accuracy of the integration. The method also estimates the relative error which can again be used to provide an error interval for the marginalized likelihood (see Appendix A.3).

3.6 Mixture deconvolution

The specified model can also be used to propose the most probable genotype combinations for the given evidence(s) E. Let a suggested combined genotypes over all loci be given as $G = (\mathbf{g}_1, ..., \mathbf{g}_I) \in \mathbb{Q}(H)$ where H is a specified hypothesis.

3.6.1 Conditioned on maximum likelihood estimates

Assume that $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimates obtained from the optimizing routine. Then by using Bayes theorem we have that

$$p(G|E, \hat{\boldsymbol{\theta}}, H) = \frac{p(E|G, \hat{\boldsymbol{\theta}})p(G|H)}{p(E|H, \hat{\boldsymbol{\theta}})}$$

where $p(E|H, \hat{\boldsymbol{\theta}})$ is the maximized likelihood value found in the optimizing routine.

We require to return a ranked list of genotypes $\mathbb{G}$ which ensures that $\sum_{G \in \mathbb{G}} p(G|E, \hat{\boldsymbol{\theta}}, H) > \alpha$ for a given selection of α. Because of this, the number of elements in $\mathbb{G}$ will vary.

3.6.2 Integrating out parameter uncertainty

We may integrate the uncertainty of $\boldsymbol{\theta}$ as following:

$$p(G|E, H) = \frac{p(G|H)}{p(E|H)} \int_{\boldsymbol{\theta}} \prod_{i=1}^{I} p(E_i|g_i, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

where $p(E|H)$ is the marginalised likelihood value found in the integration routine. This is not yet implemented...

3.6.3 Discrete optimalization search algorithm

To find the δ most probable genotype combinations we must apply a local search algorithm which is described in details in Appendix. The purpose of the algorithm is to find the global most probable genotypes by searching locally at each loci. Tvedebrink et al. [13] introduced a greedy search which keeped the most probable combined genotype combination at each steps. We extend this idea to keep the δ most probable to insure that we find the global most probable.

3.7 Validation of the model

For a given locus i and model parameter $\boldsymbol{\theta}$ we have that the peak heights $\mathbf{y}$ has density

$$p(\mathbf{y}|\boldsymbol{\theta}, E, H) = \sum_{\mathbf{g} \in \mathbb{Q}} p(\mathbf{y}|\mathbf{g}, \boldsymbol{\theta})p(\mathbf{g}|f_{st}, H)$$

If the model fits the data, we would expect that the cumulative probabilities up to the observed peak heights are uniformly distributed. We define this cumulative probability as

$$q(\mathbf{y}|\boldsymbol{\theta}, E, H) = \int_T^{y_1} \cdots \int_T^{y_n} p(x_1, ..., x_n|\boldsymbol{\theta}, E, H) dx_1 \cdots dx_n$$

4 Examples

In these examples we will calculate the Weight-of-evidence (WoE) by imputing the maximum likelihood estimates $\boldsymbol{\theta}^*$ under each of the hypothesis to the likelihood function. When comparing two rivaling hypothesis H_p and H_d the weight-of-evidence becomes

$$WoE(H_p, H_d) = \log_{10} L(E|H_p, \theta = \theta^*_{H_p}) - \log_{10} L(E|H_d, \theta = \theta^*_{H_d})$$

where $\theta^*_{H_p}$ and $\theta^*_{H_d}$ are the maximum likelihood estimates under each of the hypothesis.

4.1 MC15

As in [4] we consider the rivaling hypothesis H_p : "$K_1 + K_2 + K_3 + U$" versus H_d : "$K_1 + K_2 + U_1 + U_2$". Our software used 3 and 202 seconds to analyse the MLE under H_p and H_d respectively, with resulting $WoE(H_p, H_d) = 13.336$. The estimates with corresponding standard errors is given in the table below.

	MLE (H_p)	SE (H_p)	MLE (H_d)	SE (H_d)
mx1	0.819	0.223	0.797	0.060
mx2	0.047	0.065	0.039	0.021
mx3	0.125	0.018	0.082	0.031
mx4	0.009	0.269	0.082	0.041
mu	914.028	35.010	915.293	40.613
sigma	0.170	0.018	0.197	0.023
xi	0.074	0.014	0.072	0.019
log10lik	-117.973		-129.309	

Applying the Laplace approximation we attain $WoE(H_p, H_d) = 10.055$.

References

[1] J. Berntsen, T. O. Espelid, and A. C. Genz. An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Transactions on Mathematical Software*, 17(4):437?451, 1991.

[2] R. G. Cowell, S. L. Lauritzen, and J. Mortera. A gamma model for dna mixture analysis. *Bayesian Analysis*, 2(2):333–348, 2007.

[3] R. G. Cowell, S. L. Lauritzen, and J. Mortera. Identification and seperation of dna mixtures using peak area information. *Forensic Science International: Genetics*, 166, 2007.

[4] R. G. Cowell, T. Graversen, S. L. Lauritzen, and J. Mortera. Analysis of forensic dna mixtures with artefacts. *Appl. Statist.*, 64(1):1–32, 2015.

[5] J. Curran, P. Gill, and R. Bill, M. Interpretation of repeat measurement dna evidence allowing for multiple contributors and population substructure. *Forensic Science International*, 148:47–53, 2005.

[6] I. W. Evett, P. Gill, and J. A. Lambert. Taking account of peak areas when interpreting mixed DNA profiles. *J Forensic Sci.*, 43:62–69, 1998.

[7] A. C. Genz and A. A. Malik. An adaptive algorithm for numeric integration over an n-dimensional rectangular region. *Journal of Computational and Applied Mathematics*, 6(4):295–302, 1980.

[8] T. Hahn. CUBA: A Library for multidimensional numerical integration. *Comput.Phys.Commun.*, 168:78–95, 2005. doi: 10.1016/j.cpc.2005.01.010.

[9] M. W. Perlin, M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose, and B. W. Duceman. Validating trueallele?? dna mixture interpretation. *Journal of Forensic Sciences*, 56:1430???1447, 2011.

[10] R. Puch-Solis. A dropin peak height model. *Forensic Sci. Int. Genet.*, (11):80–84, 2014.

[11] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. W. Evett, J. Curran, and D. Balding. Evaluating forensic dna profiles using peak heights, allowing for multiple donors, allelelic dropout and stutters. *Forensic Sci. Int. Genet.*, 2013.

[12] D. Taylor, J. A. Bright, and J. Buckleton. The interpretation of single source and mixed dna profiles. *Forensic Science International: Genetics*, 7, 2013.

[13] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling. Identifying contributors of DNA mixtures by means of quanitative information of str typing. *Journal of Computational Biology*, 18, 2011.

A Appendix I: Derivings

A.1 Expanded covariance structure of restricted mixture proportion

To get the full covariance structure of $\boldsymbol{\theta} = (\mathbf{m}, \mu, \sigma, \xi)$ we need to calculate $Cov[m_C, \boldsymbol{\theta}]$. The reason is that our implemented likelihood function does not depend on m_C directly. The formula for the covariance with the other parameters is given as

$$Cov[m_C, \boldsymbol{\theta}] = Cov[-\sum_{k=1}^{C-1} m_k, \boldsymbol{\theta}] = -\sum_{k=1}^{C-1} Cov[m_k, \boldsymbol{\theta}]$$

with the variance given as

$$Var[m_C] = Var[\sum_{k=1}^{C-1} m_k] = \sum_{k=1}^{C-1}\sum_{l=1}^{C-1} Cov[m_k, m_l]$$

A.2 Reparameterization

There are advantageous in transforming the restriced space to unrestricted space. For instance it could be a good idea to transform the simplex space $\mathcal{M} = simplex(\mathbf{1}_C)$ (i.e. $\mathbf{m} \in [0,1]^C$ with restriction $m_C = 1 - \sum_{k=1}^{C-1} m_k \geq 0$) into to domain $\mathbb{R}^{C-1}$. Optimization, Laplace approximation and MCMC performance will perhaps be better in an unrestricted space. We now present the transformation from the restricted parameter space of $\boldsymbol{\theta} = (\mathbf{m}, \mu, \sigma, \xi) \in \left(\mathcal{M}, \mathbb{R}_+^2, [0,1]\right)$ to the unrestricted space $\boldsymbol{\phi} = \left(\boldsymbol{\nu}, \tilde{\mu}, \tilde{\sigma}, \tilde{\xi}\right), \in \mathbb{R}^{(C+2)}$.

For given contributors C, $\mathbf{m} = (m_1, ..., m_C) \in \mathcal{M}$ are given as the mixture proportion with $m_C = 1 - (m_1 + ... + m_{C-1})$. Transformations for given $\mathbf{m} \in \mathcal{M}$ (constrained) to unconstrained domain $\mathbb{R}^{C-1}$ is given as

$$\begin{aligned}
\nu_1 &= \log\left(\frac{m_1}{1-m_1}\right) \\
\nu_2 &= \log\left(\frac{m_2/(1-m_1)}{1-m_2/(1-m_1)}\right) \\
&\vdots \\
\nu_{(C-1)} &= \log\left(\frac{m_{(C-1)}/(1-\sum_{k=1}^{C-2} m_k)}{1-m_{(C-1)}/(1-\sum_{k=1}^{C-2} m_k)}\right)
\end{aligned}$$

with $\boldsymbol{\nu} = (\nu_1, ..., \nu_{(C-1)}) \in \mathbb{R}^{(C-1)}$. This transformation also takes care of $\sum_{k=1}^{C-1} m_k \leq 1$

We also do transformations of (μ, σ, ξ) such that $\tilde{\sigma} = \log\sigma$, $\tilde{\mu} = \log\mu$ and $\tilde{\xi} = \log\left(\frac{\xi}{1-\xi}\right)$

A.2.1 Inverse transformations

The inverse transformation of the variables $\boldsymbol{\nu} \in \mathbb{R}^{(C-1)}$ back to $\mathcal{M}$ is given as

$$\begin{aligned}
m_1 &= (1+e^{-v_1})^{-1} \\
m_2|m_1 &= (1+e^{-v_2})^{-1}(1-m_1(v_1)) \\
&\vdots \\
m_{C-1}|(m_1,...,m_{(C-2)}) &= (1+e^{-v_{(C-1)}})^{-1}\left(1-\sum_{k=1}^{C-2} m_k(v_1,...,v_k)\right) \\
m_C|(m_1,...,m_{(C-1)}) &= 1-\sum_{k=1}^{C-1} m_k(v_1,...,v_k)
\end{aligned}$$

Also, the transformations back to (μ, σ, ξ) is given as

$$\begin{aligned}
\sigma &= e^{\tilde{\sigma}} \\
\mu &= e^{\tilde{\mu}} \\
\xi &= (1-e^{-\tilde{\xi}})^{-1}
\end{aligned}$$

A.2.2 Jacobian matrix of the inverse transformation

For a given number of contributors C, the elements (i,j) in the $(C+2)\times(C+2)$-Jacobian matrix $J(\boldsymbol{\phi})$ of the inverse transformation of the variables $\boldsymbol{\phi}$ back to the variables $\boldsymbol{\theta}$ is given as

$$\begin{aligned}
J_{ij} &= \frac{\partial m_i(v_1,...,v_i)}{\partial v_j} = -(1+e^{-v_i})^{-1}\sum_{l=j}^{i-1}\frac{\partial m_l(v_1,..,v_l)}{\partial v_j} && 0<j<i<(C-1) \\
J_{ij} &= \frac{\partial m_i(v_1,...,v_i)}{\partial v_i} = e^{-v_i}(1+e^{-v_i})^{-2}\left(1-\sum_{l=1}^{i-1} m_l(v_1,...,v_l)\right) && 0<i=j<C \\
J_{ij} &= \frac{\partial m_i(v_1,...,v_i)}{\partial v_j} = 0 && 0<i<j<C \\
J_{ij} &= \frac{\partial \mu(\tilde{\mu})}{\partial \tilde{\mu}} = e^{\tilde{\mu}} && i=j=C \\
J_{ij} &= \frac{\partial \sigma(\tilde{\sigma})}{\partial \tilde{\sigma}} = e^{\tilde{\sigma}} && i=j=C+1 \\
J_{ij} &= \frac{\partial \xi(\tilde{\xi})}{\partial \tilde{\xi}} = e^{-\tilde{\xi}}(1+e^{-\tilde{\xi}})^{-2} && i=j=C+2 \\
J_{ij} &= 0 && \text{else}
\end{aligned}$$

where else is the set $\{\{i,j\} : (C-1)<j<i<(C+3) \text{ or } (C-1)<i<j<(C+3)\}$.

A.2.3 Delta method

Given the covariance matrix $\tilde{\Sigma}(\hat{\boldsymbol{\phi}})$ of the maximum likelihood estimates of $\boldsymbol{\phi}$ (optimized in the space $\mathbb{R}^{(C+2)}$), the covariance matrix of the maximum likelihood estimates of $\boldsymbol{\theta}$ is approximately $\Sigma(\hat{\boldsymbol{\theta}}) \approx J(\hat{\boldsymbol{\phi}})^T\tilde{\Sigma}(\hat{\boldsymbol{\phi}})J(\hat{\boldsymbol{\phi}})$

A.3 Error interval from relative error

Let y be a measure for the real x. With Relative Error given as $\delta = \frac{|y-x|}{x}$, the error interval for x becomes $[\frac{y}{1-\delta}, \frac{y}{1+\delta}]$. With the absolute error $\Delta = |y - x|$ given, the error interval for x becomes $[y - \Delta, y + \Delta]$.

A.4 Laplace approximation

We now show how to derive the laplace approximation based on a second order taylor expansion and use this in weight-of-evidence. Consider we have a likelihood function $L(\theta)$ with $l(\theta) = log(L(\theta))$ to be the log-likelihood where we consider θ to be a p large parameter vector. The integral we want to calculate can be written as

$$L(E) = \int L(\theta|E)d\theta = \int e^{l(\theta|E)}d\theta = e^{l(\theta_0|E)} \int e^{l(\theta|E)-l(\theta_0|E)}d\theta$$

With a second order taylor approximation around θ_0 we have that

$$l(\theta|E) \approx l(\theta_0|E) + \Delta(\theta_0)(\theta - \theta_0) + 0.5(\theta - \theta_0)^T H(\theta_0)(\theta - \theta_0)$$

where $\Delta(\theta_0)$ is the partial derivative vector and H is the hessian matrix for $l(\theta|E)$, both evaluated at $\theta = \theta_0$.

Note that when $\theta_0 = \hat{\theta}_{ML}$, $\Delta l(\theta_0|E) = (0, ..., 0)$ (i.e. a singular point). And hence it follows that

$$L(E) \approx e^{l(\hat{\theta}_{ML}|E)} \int e^{0.5(\theta-\hat{\theta}_{ML})^T H(\hat{\theta}_{ML})(\theta-\hat{\theta}_{ML})}d\theta$$

Further note that if $H(\hat{\theta}_{ML})$ is positive definite (i.e. $\hat{\theta}_{ML}$ is a positive maximum),

$$p(\theta) = (2\pi)^{-\frac{p}{2}}|H(\hat{\theta}_{ML})|^{\frac{1}{2}}e^{-0.5(\theta-\hat{\theta}_{ML})^T(-H(\hat{\theta}_{ML}))(\theta-\hat{\theta}_{ML})}$$

is a multivariate normal density with mean $\hat{\theta}_{ML}$ and $-H(\hat{\theta}_{ML})^{-1}$, and the negative inverted hessian as the covariance matrix. Hence

$$L(E) \approx e^{l(\hat{\theta}_{ML}|E)}(2\pi)^{\frac{p}{2}}|H(\hat{\theta}_{ML})|^{-\frac{1}{2}} = L(\hat{\theta}_{ML}|E)(2\pi)^{\frac{p}{2}}|H(\hat{\theta}_{ML})|^{-\frac{1}{2}}$$

Note that this approximation is only good if the likelihood function is symmetrical and number of samples to the paramters are large (large sample theory convergence).

Further assume that we want to evaluate LR between the hypothesis H_p and H_d, with same number of parameters in both models. Then we have that

$$LR(H_p, H_d|E) \approx \frac{L(\hat{\theta}_{ML}|E, H_p)}{L(\hat{\theta}_{ML}|E, H_d)}\sqrt{\frac{|H(\hat{\theta}_{ML}|H_d)|}{|H(\hat{\theta}_{ML}|H_p)|}} \approx \frac{L(\hat{\theta}_{ML}|E, H_p)}{L(\hat{\theta}_{ML}|E, H_d)}$$

where the last identity is true if $H(\hat{\theta}_{ML}|H_d) \approx H(\hat{\theta}_{ML}|H_p)$ (i.e. the hessian matrix doesn't change very much for the two comparing hypotheses), Note that if any hypothesis contains at least 2 unknowns, the likelihood function will be multimodal, but symmetrical. Because of this, the number of multimodalities needs to be taken into account in the approximated integral. Let U_p be number of unknown contributors under H_p and U_d be number of unknown contributors under H_d. Then

$$LR(H_p, H_d|E) \approx \frac{U_p! L(\hat{\theta}_{ML}|E, H_p)}{U_d! L(\hat{\theta}_{ML}|E, H_d)}$$

When considering all the loci individualy, we expect there will be a group of combinations that have high D relative to other combinations. These are clustered out and keeped further in the algorithm. The algorithm proceeds as following:

1. Put the next ranked locus (outside the set $\mathbb{B}$) into the set $\mathbb{B}$.
2. Local optimization step: In this step, every genotype combinations combination G_j (for the loci in the set $\mathbb{B}$) is giving a weight D_j.
3. If there are still loci outside the set $\mathbb{B}$ ($|\mathbb{B}| < I$), the top ϵ combined genotype combinations of D_j are keeped. Stop otherwise.
4. Go to step (1).

B Model extensions

B.1 Handling degeneration

B.2 Handling locus variability