The multivariate Dirichlet-multinomial distribution and its application in forensic genetics to adjust for sub-population effects using the $\theta$-correction

T. TVÆDEBRINK∗ P. S. ERIKSEN

Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark
tvede@math.aau.dk svante@math.aau.dk

N. MORLING

Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V’s Vej 11, DK-2100 Copenhagen, Denmark
niels.morling@sund.ku.dk

September 9, 2014

Abstract

We present a multivariate generalisation of the Dirichlet-multinomial distribution. An example of forensic genetic statistical analysis of DNA mixtures motivated the study.

In forensic genetics, adjustment of the DNA profile match probabilities due to remote ancestry in the population is often done using the $\theta$-correction, that corrects for the increased probability of observing multiple copies of rare alleles due to inbreeding in the population and thereby reduces the weight of the evidence of rare genotypes.

By numerical examples, we show how the $\theta$-correction incorporated by the use of the multivariate Dirichlet-multinomial distribution affects the weight of evidence. Furthermore, we demonstrate how the $\theta$-correction can be incorporated in Bayesian networks facilitating efficient computations.

Keywords: Multivariate Dirichlet-multinomial distribution; STR DNA mixture; Forensic genetics; $\theta$-correction

1 Introduction

When biological material from two or more contributors is observed at a scene of crime, the resulting DNA profile is a so-called DNA mixture. Recently, Cowell et al. (2015) published a statistical model for DNA mixtures, where the efficiency of their computational approach relied on a Markov structure for representing the influence of the individual genotypes (see Fig. 1). Here, we show how this Markov structure can be modified in order to incorporate positive correlations between alleles within and between the genotypes involved. A consequence of positive correlation is increased

∗Corresponding author
probability of homozygosity, which may be induced by sub-population structures in the population. In forensic genetics, this correlation is often modelled using the so-called $\theta$-correction (Balding and Nichols, 1994).

In this paper, we discuss a multivariate extension of the Dirichlet-multinomial distribution. The Dirichlet-multinomial distribution was first discussed by Mosimann (1962), who derived the distribution as a compound distribution, where the probability vector of a multinomial distribution is assumed to follow a Dirichlet distribution (Mosimann, 1962). After marginalisation over this distribution, the cell counts follow a Dirichlet-multinomial distribution (Mosimann, 1962; Johnson et al., 1997).

Modelling overdispersion of count data continues to be an area of research. Recently, Bouguila (2008); Samanta et al. (2009); Tvedebrink (2010); Avetisyan and Fox (2012); de Valpine and Harmon-Threatt (2013); Yu and Shaw (2014) discussed the modelling of overdispersed count data using a compound approach involving the multinomial distribution.

The paper is structured as follows: In Section 2, we discuss how the $\theta$-correction is implemented for a single DNA profile, and in Section 3, this is generalised for more contributors. This latter extension of the genotype model leads to the introduction of the multivariate Dirichlet-multinomial distribution. In Section 4, we derive the structures of the marginal and conditional distributions of the multivariate Dirichlet-multinomial distribution. Furthermore, the expression of the generalised factorial moments are derived, which is used to obtain the mean and covariance matrix of the distribution. In Section 5, we show by numerical examples how the $\theta$-correction affects the weight of the evidence.

2 Dirichlet-multinomial distribution

In order to adjust for genetic sub-population structures when computing the weight of evidence in forensic genetics, it is common to use the $\theta$-correction (Balding and Nichols, 1994). Several authors have discussed the interpretation of $\theta$; Curran et al. (1999, 2002) derived likelihood ratio expressions with $\theta$ being the probability that a pair of alleles is identical-by-descent (IBD). Tvedebrink (2010) defined $\theta$ as an overdispersion parameter in a multinomial sampling scheme, and Green and Mortera (2009) discussed $\theta$ in relation to assumptions made about founding genes in populations.

In forensic genetics, the prevailing genotyping technology is based on short tandem repeat (STR) loci. The genotype at a given STR locus is represented by a pair of alleles, where each of the alleles is inherited from the individual’s parents. Let $A$ denote the possible number of alleles, typically in the range of five to 20, at a given STR locus. The genotype of individual $i$ can be represented as a vector of alleles counts, $n_i$, where $n_{ia}$ is the number of $a$ alleles in the genotype. By forming a cumulative sum, $S_{ia} = \sum_{b=1}^{a} n_{ia}$, of alleles counts, $n_{ia}$, for alleles $a \in \{1, \ldots, A\}$, Graversen and Lauritzen (2014) showed that the multinomial distribution over allele counts for unknown contributors may be evaluated by the product of a sequence of binomial distributions (see Fig. 1), such that $n_{ia} \mid S_{i,a-1} \sim \text{bin}(2 - S_{i,a-1}, Q_a)$, where $Q_a = q_a / \sum_{b=1}^{A} q_b$.

![Fig. 1: Markov structure for allele counts of contributor $i$ for a marker with six possible alleles.](image)

If the distribution of allele probabilities is assumed to follow a Dirichlet distribution, then the marginal distribution of allele counts under a multinomial sampling scheme follows a Dirichlet-
multinomial distribution \cite{Tvedebrink2010}. Using similar derivations as in \cite{GraversenLauritzen2014}, we show that the \( \theta \)-correction can be incorporated by evaluating the Dirichlet-multinomial distribution by a sequence of beta-binomial distributions.

Let \( n = \sum_{b=1}^{A} n_b \) and suppress the subscript \( i \), then the Dirichlet-multinomial distribution can be specified by

\[
P(n_1, \ldots, n_A) = \frac{n! \Gamma(\alpha_\star) \prod_{b=1}^{A} \left\{ \frac{\Gamma(n_b + \alpha_b)}{n_b! \Gamma(\alpha_b)} \right\}}{\Gamma(n + \alpha_\star)}, \quad \text{where} \quad \alpha_\star = \sum_{b=1}^{A} \alpha_b,
\]

with \( \alpha = (\alpha_1, \ldots, \alpha_A) \) being positive real valued parameters \cite[pp. 81]{Johnson1997}. The joint distribution over sums of disjoint subsets of cell counts is also Dirichlet-multinomial \cite[pp. 81]{Johnson1997}. In particular, when collapsing the last \( A - a \) cells into one cell, it will yield a parameter-vector of \( (\alpha_1, \ldots, \alpha_a, \alpha_{a+1}) \) with \( \alpha_{a+1} = \sum_{b=a+1}^{A} \alpha_b \). In the case where \( n \) denotes the alleles counts, \( n = 2 \) and the distribution of allele counts \( (n_1, \ldots, n_a) \), \( a \in \{1, \ldots, A-1\} \) is given by:

\[
P(n_1, \ldots, n_a) = \frac{2! \Gamma(\alpha_\star) \Gamma(2 - S_a + \alpha_{a+1}) \prod_{b=1}^{a} \left\{ \frac{\Gamma(n_b + \alpha_b)}{n_b! \Gamma(\alpha_b)} \right\}}{\Gamma(2 + \alpha_\star) \Gamma(2 - S_a)! \Gamma(\alpha_{a+1})}.
\]

Using this result, we obtain the conditional distribution of \( n_a \) given \( n_1, \ldots, n_{a-1} \) as

\[
P(n_a \mid n_{a-1}, \ldots, n_1) = \frac{P(n_a, n_{a-1}, \ldots, n_1)}{P(n_{a-1}, \ldots, n_1)} = \frac{2! \Gamma(\alpha_\star) \Gamma(2 - S_a + \alpha_{a+1}) \prod_{b=1}^{a} \left\{ \frac{\Gamma(n_b + \alpha_b)}{n_b! \Gamma(\alpha_b)} \right\}}{\Gamma(2 + \alpha_\star) \Gamma(2 - S_a)! \Gamma(\alpha_{a+1})} \cdot \frac{\Gamma(\alpha_{a+1} + \alpha_a) \Gamma(n_a + \alpha_a) \Gamma(2 - S_{a-1} - n_a + \alpha_{a+1})}{\Gamma(\alpha_a) \Gamma(\alpha_{a+1}) \Gamma(2 - S_{a-1} + \alpha_{a+1} + \alpha_a)}
\]

where we from the second to the third line used that \( S_a = S_{a-1} + n_a \) and \( \alpha_{a+1} = \alpha_a - \alpha_a \). This is a beta-binomial distribution \cite[pp. 81]{Johnson1997} with parameters \( (2 - S_{a-1}, \alpha_a, \alpha_{a+1}) \) that are similar to those of the binomial distribution, \( (2 - S_{a-1}, Q_a) \). Similarly to \cite{GraversenLauritzen2014}, we observe directly from the expression that \( n_a \perp \{n_1, \ldots, n_{a-1}, S_{a-1}, \ldots, S_2 \} \mid S_{a-1} \). Finally, we note that the allele probabilities \( q = (q_1, \ldots, q_A) \) and \( \theta \) are related to \( \alpha \) through \( q_a = \alpha_a / \alpha_\star \) and \( \theta = (1 + \alpha_\star)^{-1} \).

3 More contributors

In order to incorporate the \( \theta \)-correction for more contributors, it is necessary to modify the Markov structure in Fig. 1. This is caused by the need to model the joint distribution of \( n_{ia} \) and \( n_{ja} \) in order to incorporate the positive correlation from remote ancestry. Thus, the Markov structure depicted in \cite[Fig. 4]{Cowell2015} should be replaced by the Markov structure in Fig. 2.

In Fig. 2 the distribution of the allele probabilities, \( q_a \), is modelled by a Dirichlet distribution. This distribution can be specified sequentially by the following relation: \( Q_a \sim \text{beta}(\alpha_a, \alpha_{a+1}) \), where \( Q_a = q_a / \sum_{b=a}^{A} q_b \) for \( a = 1, \ldots, A-1 \). Furthermore, these beta-distributions are independent \cite{Johnson1997}, which implies that the Dirichlet distribution can be written as a product of beta distributions.
First, we observe that, conditioned on \( Q_a \) and cumulative sums, the allele counts from the two individuals are independent:

\[
P(n_{ia}, n_{ja} \mid S_{i,a-1}, S_{j,a-1}, Q_a) = P(n_{ia} \mid S_{i,a-1}, Q_a) P(n_{ja} \mid S_{j,a-1}, Q_a)
\]

\[
= \binom{2-S_{i,a-1}}{n_{ia}} \binom{2-S_{j,a-1}}{n_{ja}} Q^\alpha_a (1-Q_a)^{4-S_{a-1}+n_{ia}}
\]

where \( n_{ia} = n_{ia} + n_{ja} \) and \( S_{a-1} = S_{i,a-1} + S_{j,a-1} \). Second, we marginalise over \( Q_a \), which is beta distributed with parameters \( (\alpha_d, \alpha_{a+1}) \):

\[
\int_0^1 P(n_{ia} \mid S_{i,a-1}, Q_a) P(n_{ja} \mid S_{j,a-1}, Q_a) f(Q_a) \, dQ_a = \frac{\Gamma(\alpha_a) \Gamma(\alpha_d) \Gamma(n_{ia} + \alpha_d) \Gamma(n_{ja} + \alpha_d) \Gamma(4-S_{a-1}+n_{ia}+\alpha_d+1)}{\Gamma(\alpha_d) \Gamma(\alpha_{a+1}) \Gamma(4-S_{a-1}+\alpha_d+1)}
\]

which is the integral of a non-normalised beta-distribution. By letting \( S_{i0} = S_{j0} = 0 \), we have that \( P(n_{ia}, n_{ja} \mid S_{i,a-1}, S_{j,a-1}) \), for \( 1 \leq a < A \), is given by

\[
\binom{2-S_{i,a-1}}{n_{ia}} \binom{2-S_{j,a-1}}{n_{ja}} \frac{\Gamma(\alpha^d_a) \Gamma(n_{ia} + \alpha_d) \Gamma(4-S_{a-1}-n_{ia}+\alpha_{d+1})}{\Gamma(\alpha_d) \Gamma(\alpha_{a+1}) \Gamma(4-S_{a-1}+\alpha_d+1)}.
\]

A consequence of marginalising over \( Q_a \) is that the clique size in the network decreases. For the two profile situation in Fig. 2, this marginalisation implies that the relevant clique size decreases from eight to six nodes as \( Q_{a-1} \) and \( Q_a \) are removed, while the imposed correlation connects the nodes \( n_{ia} \) and \( n_{ja} \) (graph not shown).

In the general setting, where we consider a DNA mixture of \( I \) contributors, we denote \( n = (n_1, \ldots, n_I) \), where each \( n_i = (n_{i1}, \ldots, n_{iA}) \) denotes the allele counts for profile \( i \) and, similarly, for the cumulative sums, \( S_{it} = \sum_{b=1}^t n_{ib} \). Hence, \( P(n_{1a}, \ldots, n_{IA} \mid S_{1,a-1}, \ldots, S_{I,a-1}) \) is given by

\[
\frac{\prod_{i=1}^I \binom{2-S_{i,a-1}}{n_{ia}} \Gamma(\alpha^d_a) \Gamma(n_{ia} + \alpha_d) \Gamma(2I-S_{a-1}-n_{ia}+\alpha_{d+1})}{\Gamma(\alpha_d) \Gamma(\alpha_{a+1}) \Gamma(2I-S_{a-1}+\alpha_d+1)}.
\]
Table 1: Sufficient statistics of a table when modelled using the multiplicative Dirichlet-multinomial (MDM) distribution. From the construction of the MDM distribution, the row sums, \( n_{i*} \), and the total sum, \( n_{**} \), are known and fixed.

\[
\begin{array}{cccc}
1 & \ldots & a & A \\
1 & n_{11} & \ldots & n_{1a} & n_{1A} & n_{1*} \\
& \vdots & \ddots & \vdots & \vdots & \vdots \\
i & n_{i1} & \ldots & n_{ia} & n_{iA} & n_{i*} \\
& \vdots & \ddots & \vdots & \vdots & \vdots \\
I & n_{I1} & \ldots & n_{Ia} & n_{IA} & n_{I*} \\
\hline
n_{*1} & \ldots & n_{*a} & \ldots & n_{*A} & n_{**}
\end{array}
\]

where \( n_{\alpha} = \sum_{i=1}^{I} n_{i\alpha} \) and \( S_{\alpha-1} = \sum_{i=1}^{I} S_{i\alpha-1} \).

In full generality, consider a set of vectors \( n = (n_1, \ldots, n_I) \), where \( n_i = (n_{i1}, \ldots, n_{iA}) \) and \( n_{i*} = \sum_{a=1}^{A} n_{ia} \) for \( n_{i*} \in \mathbb{Z}_0 \). Then, the probability mass function for \( n \) is given by

\[
P(n) = \left\{ \prod_{i=1}^{I} \left( \frac{n_{i}}{n_i} \right)^{n_{i|}} \right\} \frac{\Gamma(n_{**})}{\prod_{a=1}^{A} \Gamma(n_{a|} + n_{a})} \prod_{a=1}^{A} \frac{\Gamma(n_{*a} + n_{a})}{\Gamma(n_{a})},
\]

where \( n_{**} = \sum_{a=1}^{A} n_{*a} = \sum_{a=1}^{A} \sum_{i=1}^{I} n_{ia} \). For a single component \( n_i \), i.e. \( n = (n_1, \ldots, n_A) \), this distribution simplifies to the Dirichlet-multinomial distribution. Hence, we may call this distribution the multivariate Dirichlet-multinomial (MDM) distribution, which we denote \( MDM(n_{**}, \alpha) \), where \( n_{**} = (n_{1*}, \ldots, n_{*A}) \) is the vector of trails per experiment (row sums in Table 1) or e.g. the number of alleles per DNA profile. Furthermore, we observe from (2), that inference about the model parameters, \( \alpha \), only depends on \( n_{**} = (n_{1*}, \ldots, n_{*A}, \ldots, n_{*A}) \), i.e. the column sums shown in Table 1.

To emphasise the difference between row and column marginals, we let \( n_{*B} = (n_{1B}, \ldots, n_{*B}, \ldots, n_{*A}) \) denote the column sums, which we will use in the next section when discussing conditional and marginal distributions.

Furthermore, let \( B \subset \{1, \ldots, A\} \) be a subset of the cells, e.g. a subset of the alleles in a genetics context, and by \( C \) we denote the complement of \( B \). The counts associated with \( B, C \) and row sums over \( C \) are defined by

\[
n_{*B} = \{ n_{ia} \}_{a \in B}, \quad n_{*C} = \{ n_{ia} \}_{a \in C} \quad \text{and} \quad n_{i(C)} = \left\{ \sum_{a \in C} n_{ia} \right\} \quad \text{for} \quad i = 1, \ldots, I, \text{ respectively.}
\]

Similarly, we may consider subsetting over index \( i \), such that \( J \) and \( K \) specify two disjoint and exhaustive partitions of \( \{1, \ldots, i, \ldots, I\} \), where \( n_{j*} \) and \( n_{K *} \) denotes the counts, respectively. In the DNA mixture context this corresponds to partition the set of \( I \) contributors into two disjoint groups.

## 4 Properties of multivariate Dirichlet-multinomial distribution

### 4.1 Conditional and marginal distributions

The construction of the MDM distribution implies that it carries many similarities with the Dirichlet-multinomial distribution. For the MDM distribution, one can consider marginalisation and conditioning over both \( i \) and \( a \) in the \( n_{ia} \) notation. Furthermore, we may also condition on \( n_{*B} \) and \( n_{*C} \) to obtain a generalisation of the hypergeometric distribution.

First, we consider the marginal and conditional distribution over index \( a \): The marginal distribution of \( n_{*B} \), can be thought of as the distribution when collapsing all elements of \( C \) into one
hypergeometric distribution (Johnson et al., 1997) with parameters
the results for contingency tables that

\[ n \]

have already observed counts \( n \)

conditional distribution of \( \alpha \)

results that the marginal and conditional distributions are MDM with parameters given by

\[ \alpha_B = \{a_a\}_{a \in B} \text{ and } \alpha_C = \sum_{a \in C} a_a. \]

Second, we handle the case of marginalising and conditioning over index \( i \): It follows directly from (2), that the distribution of \( n_{j*} \) is MDM with parameters \( n_{j*} \) and \( \alpha \), where \( n_{j*} = \{n_{i*}\}_{i \in J} \). The conditional distribution of \( n_{j*} \) given \( n_{K*} \) can be considered as a posterior distribution, such that we have already observed counts \( n_{K*} \), which are then factorised into the parameters. Thus, we have

\[ n_{j*} \sim \text{MDM}(n_{j*}, \alpha) \quad \text{and} \quad n_{j*} | n_{K*} \sim \text{MDM}(n_{j*}, \alpha + n_{K*}). \]

where \( n_{K*} = \{\sum_{i \in K} n_{i*}\} \) for \( a = 1, \ldots, A \), i.e. the number of \( a \) alleles observed for the profiles in \( K \).

Finally, when conditioning on the sufficient statistic, \( n_{s*} \), and the number trails, \( n_{e*} \), we recover the results for contingency tables that \( n_{**} | (n_{s*}, n_{e*}) \) follow a generalisation of the multivariate-hypergeometric distribution (Johnson et al., 1997) with parameters \( n_{e*} \) and \( n_{s*} \):

\[ P(n; n_{e*}, n_{s*}) = \prod_{i=1}^{l_i} \left( \frac{\binom{n_{i*}}{n_{r*}}}{\binom{n_{e*}}{n_{e*}}} \right) \prod_{i=1}^{l_i} \binom{n_{i*}}{n_{r*}}! \prod_{a=1}^{A} \binom{n_{e*}}{n_{e*}}! \prod_{i=1}^{l_i} \binom{n_{e*}}{n_{e*}}! \prod_{a=1}^{A} \binom{n_{e*}}{n_{e*}}!, \]

which is identical to Halton’s “exact contingency formula” (Halton, 1969) and utilised in Patefield’s algorithm (Patefield, 1981) to generate \( R \times C \) contingency tables.

4.2 Moments

In Appendix A we show that the moments of MDM can be computed using the generalised factorial moments, which are given by:

\[ \mathbb{E} \left( n^{(r)} \right) = \mathbb{E} \left( \prod_{i=1}^{l_i} \prod_{a=1}^{A} n_{i*}^{(r_{ia})} \right) = \left\{ \prod_{i=1}^{l_i} \binom{n_{i*}}{r_{i*}}! \right\} \prod_{a=1}^{A} \prod_{k=0}^{r_{a*}-1} (a_a + k) \prod_{i=1}^{l_i} \binom{n_{i*}}{r_{i*}}! \prod_{k=0}^{r_{a*}-1} (a_a + k), \tag{3} \]

where \( a^{(b)} = a(a-1) \cdots (a-b+1) = a!/(a-b)! \) is a rising factorial. Hence, in order to compute the mean of \( n_{i*} \), we set \( r_{ia} = 1 \) and \( r_{i*} = 1 \) (implying that \( r_{i} = 1 \) and \( r_{e} = 1 \)). Plugging this into (3), we obtain \( \mathbb{E}(n_{i*}) = n_{i*} \alpha_i / n_{e*} = n_{i*} q_a \) as expected. Furthermore, the covariance matrix can be computed for the different levels of correlation (left: within individual \( i \), and right: between individuals \( i \) and \( i' \)):

\[ \text{Cov}(n_{i*}, n_{i*}) = n_{i*} q_{a*} (1 - q_a)[1 + (n_{i*} - 1)\theta] \quad \text{and} \quad \text{Cov}(n_{i*}, n_{e*}) = n_{i*} n_{i*} q_{a*} (1 - q_a) \theta \]

\[ \text{Cov}(n_{i*}, n_{i'}) = -n_{i*} q_{a*} q_{a'}[1 + (n_{i*} - 1)\theta] \quad \text{and} \quad \text{Cov}(n_{i*}, n_{e*}) = -n_{i*} n_{i*} q_{a*} q_{a'} \theta, \]

where \( \text{Cov}(n_{i*}, n_{i*}) = \text{Var}(n_{i*}) \). In the case where \( n \) represents a DNA profile, we have for all \( i \) that \( n_{i*} = 2 \). Thus, in this particular case, we obtain:

\[ \text{Cov}(n_{i*}, n_{i*}) = 2q_{a*}(1 - q_a)[1 + \theta] \quad \text{and} \quad \text{Cov}(n_{i*}, n_{e*}) = 4q_{a*}(1 - q_a)\theta \]

\[ \text{Cov}(n_{i*}, n_{i'}) = -2q_{a*}q_{a'}[1 + \theta] \quad \text{and} \quad \text{Cov}(n_{i*}, n_{e*}) = -4q_{a*}q_{a'}\theta. \]
5 Numerical results

In order to demonstrate how the \( \theta \)-correction affects \( P(n|H) \) in the evaluation of the \( L(H) \) expression of Equation 8 in Cowell et al. (2015), we evaluate

\[
\text{WoE}(n_{\bullet a}, S_{\bullet a-1}; Q, \theta) = \frac{P(n_{ia} | S_{ja-1}; Q_a)P(n_{ja} | S_{ja-1}; Q_a)}{P(n_{ia}, n_{ja} | S_{ja-1}; Q_a, \theta)}
= \frac{\Gamma(\alpha^a_i)\Gamma(n_{ja} + \alpha^a_j)\Gamma(4 - S_{ja-1} - n_{ja} + 4\alpha^a_{a+1})}{\Gamma(\alpha^a_j)\Gamma(\alpha^a_{a+1})\Gamma(4 - S_{ja-1} + \alpha^a_j)}
\]

where the last expression emphasises that this ratio only depends on the allele counts \((n_{ia}, S_{ja-1})\) and \((n_{ja}, S_{ja-1})\) through the margins \(n_{\bullet a} = n_{ia} + n_{ja}\) and \(S_{\bullet a-1} = S_{ja-1} + S_{ia-1}\), \(0 \leq n_{\bullet a} + S_{\bullet a-1} \leq 4\). Hence, the \( n_{\bullet a} = 2 \) situation covers both the combination of two heterozygous profiles and one homozygous profile together with a profile with no \( a \) allele. Similar symmetries can be identified for different values of \( n_{\bullet a} \) and \( S_{\bullet a-1} \).

For a two-person DNA mixture, only 15 non-symmetric combinations exist, although for \( S_{\bullet a-1} \geq 3 \), we have that \( n_{\bullet a} \leq 1 \), which implies that no correlation can be observed. Therefore, only 12 relevant combinations are shown in Fig. 3. The general picture in Fig. 3 is that \( \text{WoE}(n_{\bullet a}, S_{\bullet a-1}; Q, \theta) < 1 \), except for \( n_{\bullet a} = 1 \), where \( \text{WoE}(n_{\bullet a} = 1, S_{\bullet a-1}; Q, \theta) \geq 1 \). That is, the product of unrelated allele probabilities, \( P(n_{ia}|S_{ja-1})P(n_{ja}|S_{ja-1}) \), is smaller than the joint probability adjusting for shared ancestry/remote relatedness, \( P(n_{ia}, n_{ja}|S_{ja-1}, S_{ja-1}) \). Hence, in the case where two or more of the same alleles are observed simultaneously, the weight of evidence is decreased. Conversely, the increased probability of homozygosity for \( \theta > 0 \) implies that singletons, \( n_{\bullet a} = 1 \), are less frequent, which implies an increase in the weight of evidence (Buckleton et al. 2005).

We also analysed how the ratio between \( P(n_i)P(n_j) \) to \( P(n_i, n_j) \) behaves. We noted that due to the \( \theta \)-correction, there is an increased probability of shared alleles between and within DNA profiles. The behaviour is similar to that pictured in Fig. 3 since the evaluation is comprised by products of \( \text{WoE}(n_{\bullet a}, S_{\bullet a-1}; Q, \theta) \). In Fig. 4 we see that it is possible to identify the contributions from Fig. 3. For example, the probability of observing three alleles of one type together with an other allele, \((3, 1)\), is the product of \( \text{WoE}(n_{\bullet a}, S_{\bullet a-1}; Q, \theta) \) for \((n_{\bullet a}, S_{\bullet a-1}) \in \{(0, 0), (1, 0), (3, 0), (1, 3), (3, 1)\}\), which due to the positive correlation between alleles is dominated by \( \text{WoE}(n_{\bullet a} = 3, S_{\bullet a-1} = 0; Q, \theta) \) and \( \text{WoE}(n_{\bullet a} = 3, S_{\bullet a-1} = 1; Q, \theta) \).

Furthermore, the legend in Fig. 4 only specifies the alleles that are observed more than once. This is because the ratio of \( P(n_i)P(n_j) \) to \( P(n_i, n_j) \) for alleles observed only once cancel out. Therefore, the ratio is simplified into a polynomial in \( \theta \), which is independent of the allelic distribution. For example, in the upper left panel of Fig. 4 (dark grey curve), the ratio is the same for all vectors \{\( (2, 1, 1, 0, 0, 0), (2, 1, 0, 1, 0, 0), \ldots, (2, 0, 0, 0, 1, 1) \}\), i.e. all combinations with \( a_1 \) observed twice give the same ratio.

6 Conclusion

We have derived a multivariate generalisation of the Dirichlet-multinomial distribution for an application in forensic genetics. The conditional distribution over the cell counts of the multivariate Dirichlet-multinomial (MDM) distribution also follows a MDM distribution. Furthermore, the conditional distributions over vectors follow an extended hypergeometric distribution.

We have demonstrated how to incorporate the \( \theta \)-correction into the computational framework of the DNAmixtures package (Graversen 2014) and exemplified how the adjustment for positive correlation between alleles caused by population stratification affects the weight of evidence.
A Generalised factorial moments of MDM

In this section, we derive the generalised factorial moments of the MDM distribution. The generalised factorial moments are useful for count data, in that it allows relatively simple expressions
for most of the distribution’s moments. The generalised factorial moments can be considered a transformation, $f$, where we use that $\mathbb{E}\{f(n)\} = \sum_{n \in \mathcal{N}} f(n)P(n)$.

More specifically, $f(n) = \prod_{i=1}^{l} \prod_{a=1}^{A} n_{ia}^{r_{ia}}$, where $a^{(b)} = a!/(a-b)!$ and $r_{ia} \in \{0, \ldots, n_{ia}\}$ is a vector of constants. Hence, if we want to compute $\mathbb{E}(n_{ia})$, we set $r_{ia} = 1$ and all other to zero.

First, we use that conditioned on $q$, the distribution of $n$ is found by products of independent multinomial distributions.

$$\mathbb{E}\left\{ \prod_{i=1}^{l} \prod_{a=1}^{A} n_{ia}^{r_{ia}} \mid q \right\} = \sum_{n \in \mathcal{N}} \prod_{i=1}^{l} \prod_{a=1}^{A} \frac{n_{ia}^{r_{ia}}}{n_{ia}!} q_{ia}^{n_{ia}}$$

$$= \sum_{n \in \mathcal{N}} \prod_{i=1}^{l} \frac{n_{ia}!}{(n_{ia} - r_{ia})!} \left(\frac{n_{ia} - r_{ia}}{n_{ia} - r_{ia}}\right) \prod_{a=1}^{A} q_{ia}^{n_{ia} - r_{ia}} q_{ia}^{r_{ia}}$$

$$= \left\{ \prod_{i=1}^{l} \frac{n_{ia}!}{(n_{ia} - r_{ia})!} \right\} \prod_{a=1}^{A} q_{ia}^{r_{ia}}$$

where we moved terms constant over $\mathcal{N} = \{ n : \sum_{a} n_{ia} = n_{ia} \}$ outside the sum and identified the remaining terms as being the product of independent multinomial distributions for $n - r$, which by definition sum to unity.

Fig. 4: Plot of the different combinations for two contributors. The underlying allele distribution is $q = (0.025, 0.05, 0.1, 0.2, 0.4)$, where the remaining probability mass is assigned to a “rest class”.
Second, we marginalise over \( q \) in order to obtain the generalised factorial moments for the multivariate Dirichlet-multinomial distribution

\[
\mathbb{E}\{f(n)\} = \left\{ \prod_{i=1}^{I} \frac{n_{i\bullet}!}{(n_{i\bullet} - r_{i\bullet})!} \right\} \frac{\Gamma(\alpha_{\bullet})}{\prod_{a=1}^{A} \Gamma(\alpha_{a})} \int \prod_{a=1}^{A} q_{a}^{\alpha_{a} + r_{a\bullet} - 1} dq
= \left\{ \prod_{i=1}^{I} \frac{n_{i\bullet}!}{(n_{i\bullet} - r_{i\bullet})!} \right\} \frac{\Gamma(\alpha_{\bullet})}{\prod_{a=1}^{A} \Gamma(\alpha_{a} + r_{a\bullet})} \prod_{a=1}^{A} \Gamma(\alpha_{a} + r_{a\bullet}) \prod_{a=1}^{A} \Gamma(\alpha_{a} + r_{a\bullet}).
\]

For the remaining terms, we see that the ratios of gamma functions involve \( \Gamma(\beta + t) \) and \( \Gamma(\beta) \). For \( t > 0 \), the gamma function satisfies

\[
\frac{\Gamma(\beta + t)}{\Gamma(\beta)} = \prod_{k=0}^{t-1} (\beta + k).
\]

Hence, the expression for \( \mathbb{E}(n^{(r)}) \) may be simplified to

\[
\mathbb{E}(n^{(r)}) = \mathbb{E} \left\{ \prod_{i=1}^{I} \prod_{a=1}^{A} \frac{n_{i\bullet}!}{(n_{i\bullet} - r_{i\bullet})!} \right\} \frac{\prod_{a=1}^{A} (\alpha_{a} + k)}{\prod_{k=0}^{r_{a\bullet} - 1} (\alpha_{a} + k)}.\]

References

Avetisyan, M. and J.-P. Fox (2012). The Dirichlet-Multinomial Model for Multivariate Randomized Response Data and Small Samples. *PSICOLOGICA* 33(2), 362–390.

Balding, D. J. and R. A. Nichols (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci Int* 64, 125–140.

Bouguila, N. (2008). Clustering of count data using generalized Dirichlet multinomial distributions. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 20(4), 462–474.

Buckleton, J. S., J. M. Curran, and S. J. Walsh (2005). How reliable is the sub-population model in DNA testimony? *Forensic Sci Int* 157, 144–148.

Cowen, R. G., T. Graversen, S. L. Lauritzen, and J. Morthera (2015). Analysis of Forensic DNA Mixtures with Artefacts. *J R Stat Soc Ser C Appl Stat* 64(1), 1–32.

Curran, J. M., J. S. Buckleton, C. M. Triggs, and B. S. Weir (2002). Assessing uncertainty in DNA evidence caused by sampling effects. *Sci Justice* 42(1), 29–37.

Curran, J. M., C. M. Triggs, J. Buckleton, and B. S. Weir (1999). Interpreting DNA mixtures in structured populations. *J Forensic Sci* 44(5), 987–995.

de Valpine, P. and A. N. Harmon-Threatt (2013). General models for resource use or other compositional count data using the Dirichlet-multinomial distribution. *ECOLOGY* 94(12), 2678–2687.

Graversen, T. (2014). *DNAmixtures: Statistical Inference for Mixed Traces of DNA*. R package version 0.1-3.

Graversen, T. and S. Lauritzen (2014). Computational aspects of DNA mixture analysis. *Stat Comput*, 1–15. In Press.
Green, P. J. and J. Morera (2009). Sensitivity of inferences in forensic genetics to assumptions about founding genes. *Ann Appl Stat* 3(2), 731–763.

Halton, J. H. (1969). A rigorous derivation of the exact contingency formula. *Proc Camb Phil Soc* 65(2), 527–530.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1997). *Discrete Multivariate Distributions.* Wiley.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate $\beta$-distribution, and correlations among proportions. *Biometrika* 49(1-2), 65–82.

Patefield, W. M. (1981). Algorithm AS159. An efficient method of generating $R \times C$ tables with given row and column totals. *Applied Statistics* 30, 91–97.

Samanta, S., Y.-J. Li, and B. S. Weir (2009). Drawing inferences about the coancestry coefficient. *THEORETICAL POPULATION BIOLOGY* 75(4), 312–319.

Tvedebrink, T. (2010). Overdispersion in allelic counts and $\theta$-correction in forensic genetics. *Theor Popul Biol* 78(3), 200–210.

Yu, P. and C. A. Shaw (2014). An efficient algorithm for accurate computation of the Dirichlet-multinomial log-likelihood function. *BIOINFORMATICS* 30(11), 1547–1554.