Revealing the range of maximum likelihood estimates in the admixture model.

Carola Sophia Heinzel[1,*], Franz Baumdicker[2,3], Peter Pfaffelhuber[1]

October 17, 2024

(1) Department of Mathematical Stochastics, Albert-Ludwigs-University Freiburg, 79104 Freiburg im Breisgau, Germany
(2) Cluster of Excellence "Controlling Microbes to Fight Infections", Mathematical and Computational Population Genetics, University of Tübingen, Sand 14, 72076 Tübingen, Germany
(3) Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Germany
(∗) Corresponding author: E-Mail: carola.heinzel@stochastik.uni-freiburg.de
E-Mail-Adresses: carola.heinzel@stochastik.uni-freiburg.de (CSH),
peter.pfaffelhuber@stochastik.uni-freiburg.de (PP)

Abstract

Many ancestry inference tools, including STRUCTURE and ADMIXTURE, rely on the admixture model to infer both, allele frequencies p and individual admixture proportions q for a collection of individuals relative to a set of hypothetical ancestral populations. We show that under realistic conditions the likelihood in the admixture model is typically flat in some direction around a maximum likelihood estimate (MLE) $(\hat{q}, \hat{p})$. In particular, the maximum likelihood estimator is non-unique and there is a complete spectrum of possible estimates. Common inference tools typically identify only a few points within this spectrum.

We provide an algorithm which computes the set of equally likely $(\tilde{q}, \tilde{p})$, when starting from $(\hat{q}, \hat{p})$. It is analytic for $K = 2$ ancestral populations and numeric for $K > 2$. We apply our algorithm to data from the 1000 genomes project, and show that inter-European estimators of q can come with a large set of equally likely possibilities. In general, markers with large allele frequency differences between populations in combination with individuals with concentrated admixture proportions lead to small areas with a flat likelihood.

Our findings imply that care must be taken when interpreting results from STRUCTURE and ADMIXTURE if populations are not separated well enough.

Key Words: Admixture Model, STRUCTURE, ADMIXTURE, Non-Unique Estimator, Label Switching

1 Introduction

Inferring the individuals' ancestry (IA) from genetic data is a formidable challenge, given its importance in various domains, such as the identification of missing persons (Phillips, 2015), the exploration of human history (Rosenberg et al., 2002), corrections for population stratification (Pritchard and Donnelly, 2001), forensic genetics (Tvedebrink, 2022) or conservation genetics (Wasser et al., 2007). Two categories of approaches exist for estimating the IA: model-based methods and model-free techniques (Wollstein and Lao, 2015). Model-based methods, such as STRUCTURE (Pritchard et al., 2000), ADMIXTURE (Alexander et al., 2009) and improvements of them (Alexander and Lange, 2011; Shringarpure et al., 2016; Dominguez Mantes et al., 2023a,b; Falush et al., 2003, 2007; Hubisz et al., 2009; Raj et al., 2014; Ko et al., 2023) treat the coefficients as parameters of a statistical model. Researchers have to be careful to not over interpret the output of STRUCTURE (Lawson et al., 2018) and STRUCTURE does not recover the population structure in every case (Puechmaille, 2016). Model-free methods use for example Principal Component Analysis or Spectral Graph Theory (Klei et al., 2011; Lee et al., 2010; Jombart et al., 2010) for inferring the IA. An overview of the existing methods can be found in (Wollstein and Lao, 2015). As an alternative to the admixture model assumption that discrete ancestral populations existed, neural network approaches can be used to infer the geographic sampling location (Battey et al., 2020). However, the boundaries between model-based and model-free approaches are not always clearly defined. For example, a neural network autoencoder can be constrained to follow the assumptions of the admixture model (Dominguez Mantes et al., 2023), which allows to quickly infer IAs in the admixture model even for large datasets. In the present study, we focus on the underlying admixture model that is the basis of the most often used model-based methods, as well as some model-free approaches. See e.g. (Lovell et al., 2021; Badouin et al., 2017; Mogensen et al., 2020) for application examples. To evaluate whether the assumptions of the admixture model are met, Garcia-Erill and Albrechtsen (2020) can be used.

For the model-based methods, the choice of the markers is important. Numerous papers have already proposed methods to choose the best marker set to estimate the IAs, see e.g. (Kidd et al., 2014; Phillips et al., 2014; Pfaffelhuber et al., 2020; Ko et al., 2023). Additionally, it is already known that there are markers, i.e. Ancestry Informative Markers (AIMs), that are more important for the estimation than other markers. Hence, researchers often use AIM sets with only 50 - 200 markers rather than thousands of markers to reduce sequencing effort and run time (Ko et al., 2023).

There are two different scenarios in which we are interested in the IA: the allele frequencies are known (supervised setting) or they need to be estimated (unsupervised setting). In the present work, we address exclusively the latter scenario. Hence, we aim to estimate both, allele frequencies and IA. The estimation of the allele frequencies is particularly crucial for enhancing familial research strategies (Rohlfs et al., 2012). Specifically, we consider maximum likelihood estimators (MLEs) in the admixture model to estimate the IAs and the allele frequencies at the same time. Therefore, we have to fix the number of ancestral populations K.

Determining the number of ancestral populations is a common problem (Evanno et al., 2005;

Pritchard et al., 2000). In particular, (Janes et al., 2017) mentioned that the two most common used methods (Evanno et al., 2005; Pritchard et al., 2000) to infer K often lead to the choice $K = 2$, see e.g. (Friedlaender et al., 2008; Rosenberg et al., 2005; Lawson et al., 2018; Hamlin and Arnold, 2014; Phillips et al., 2009). As (Lawson et al., 2012), STRUCTURE is in practice not applied for many different populations, i.e. $K \geq 10$. In this study, we provide further evidence that researchers should choose K as small as possible. (Wang, 2017) even said that it is difficult to use STRUCTURE because of its large amount of possible parameters like K and (Verity and Nichols, 2016) mentioned that using STRUCTURE to estimate K often leads to wrong results.

It is already known that different runs of STRUCTURE can result in different estimates (Jakobsson and Rosenberg, 2007; Alexander et al., 2009; Pritchard et al., 2010; Verdu et al., 2014; Kopelman et al., 2015; Behr et al., 2016; Cabreros and Storey, 2019; van Waaij, 2022; Liu et al., 2024). This is mostly attributed to label-switching, although (Liu et al., 2024) acknowledge the fact that different patterns of co-clustering of individuals may as well lead to different results. This is in line with general theory, since MLEs are known to be non-unique in general (Dharmadhikari and Joag-Dev, 1985). (Behr et al., 2016) developed among others (Rosenberg, 2004; Jakobsson and Rosenberg, 2007; Earl and VonHoldt, 2012; Kopelman et al., 2015; Liu et al., 2024) algorithms to analyse the different outputs of STRUCTURE in the unsupervised setting and resolve label-switching. This can also mean that the IAs were estimated for a different number of ancestral populations see e.g. (Margaryan et al., 2020; Larena et al., 2021; Jia et al., 2020; Liu et al., 2024). Usually, these approaches have to run STRUCTURE many times to find some different estimators. This computational expensive method consequences that there might be other possible outcomes with the same likelihood as the output of STRUCTURE.

We show that the local maximal points and even the global optima are generally not unique in the unsupervised setting. This non-uniqueness might result in a large range of the IA and the allele frequencies with the same likelihood as the estimator identified e.g. by STRUCTURE. The differences between equally likely estimates are not just nuances, they can be substantial. In this case, there is no reliable claim concerning the IA and the allele with MLEs possible. Hence, it is important to know every possibility of the IA and the allele frequencies with the same maximum likelihood. Additionally, we provide information how to get estimators that are meaningful, i.e. where the range of MLEs is small. Our results highlight that reliable ancestry estimates depend on at least a few SNPs with large allele frequency differences between the populations.

Our objective here is to find every MLE given one MLE and to compute the differences between them depending on the estimated allele frequencies and the estimated IAs. Therefore, we first explain the admixture model and the general idea. Then, we introduce an algorithm, EMALAM (Every MAximum Likelihood estimator in the Admixture Model) available at our GitHub repository or as online version that explores the range of MLEs. To show the consequences of the non-unique MLEs, we apply our theoretical results to genetic data from the 1000 Genomes Project (Consortium et al., 2015). Furthermore, we compare our approach to the method proposed by (Behr et al., 2016), i.e. to running STRUCTURE several times. In

contrast to the identification of the different modes discovered by STRUCTURE/ADMIXTURE, here we explore the complete space of equally likely MLEs.

2 Methods

2.1 The admixture Model

Let us reintroduce the admixture model in the unsupervised setting. For M markers, where marker m has J_m different alleles, $m = 1, \dots, M$, the genetic data of N diploid individuals is given by $x = (x_{ijm})_{(i=1,\dots,N, j=1,\dots,J_m, m=1,\dots,M)}$. Additionally, $x_{ijm} \in \{0, 1, 2\}$ determines the number of copies of allele j in individual i at locus m. We assume that we are dealing with diploid individuals, i.e., $\sum_{j=1}^{J_m} x_{ijm} = 2$. We fix the number of distinct ancestral populations K and aim to infer the IAs $q = (q_{ik})_{i=1,\dots,N, k=1,\dots,K}$ and the allele frequencies $p = (p_{kjm})_{k=1,\dots,K, j=1,\dots,J_m, m=1,\dots,M}$ from the genetic data. Here, q_{ik} represents the proportion of the genome of individual i that originates from population k (i.e., the IA of individual i in population k), and p_{kjm} represents the frequencies of allele j in the assumed ancestral population k at marker m.

To depict the likelihood, we write $q_{i,\cdot}$ for the row vector $(q_{i1}, \dots, q_{iK})$, and the allele frequencies as $p_{\cdot jm} = (p_{1jm}, \dots, p_{Kjm})$ and note that q, p satisfy the conditions (where 1_K is the 1-vector of length K)

$$\begin{aligned} &\text{(i) } 0 \le q_{ik} \le 1, && \text{(ii) } q_{i\cdot} 1_K = 1_K \\ &\text{(iii) } 0 \le p_{kjm} \le 1, && \text{(iv) } p_{k\cdot m} 1_{J_m} = 1 \end{aligned} \tag{2.1}$$

i.e., the IAs q of each individual sum to one and allele frequencies p sum to one at each locus. Assuming that x_{ijm} is the number of alleles j of individual i at marker m and C_x is a constant which depends only on $x = (x_{ijm})$, the log-likelihood for p, q given x is, (using $\top$ to indicate column vectors)

$$\ell(q, p|x) = C_x + \frac{1}{2MN} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{j=1}^{J_m} x_{ijm} \log(q_{i\cdot} p_{\cdot jm}^{\top}). \tag{2.2}$$

Here, $q_{i\cdot} p_{\cdot jm}^{\top}$ is a scalar product, i.e. $q_{i\cdot} p_{\cdot jm}^{\top} = \sum_{k=1}^{K} q_{ik} p_{kjm}$. Apparently, the likelihood (2.2) depends on q, p only via $(q_{i\cdot} p_{\cdot jm}^{\top})_{i=1,\dots,N, j=1,\dots,J_m, m=1,\dots,M} = qp$. MLEs, as e.g. computed by STRUCTURE and ADMIXTURE, are given by

$$(\hat{q}, \hat{p}) := (\hat{q}_x, \hat{p}_x) := \operatorname{argmax}\{(q, p) \mapsto \ell(q, p|x)\}.$$

STRUCTURE and ADMIXTURE (as well as their relative FRAPPE (Tang et al., 2005) and descendants fastSTRUCTURE Raj et al. (2014) and openADMIXTURE Ko et al. (2023) differ in the ways this argmax is computed.

2.2 Non-unique MLEs

We start with an explanation why the MLE in the admixture model is not unique. As stated above, the likelihood $\ell(q, p|x)$ depends on q, p only via the scalar products $q_{i\cdot} p_{\cdot jm}^{\top}$. As a

consequence, $\ell(\hat{q},\hat{p}|x) = \ell(\tilde{q},\tilde{p}|x)$ provided that

$$\hat{q}_{i\cdot}\hat{p}_{\cdot jm}^\top = \tilde{q}_{i\cdot}\tilde{p}_{\cdot jm}^\top \text{ for all } i,j,m. \tag{2.3}$$

In order to see that this is in fact possible, let S be an invertible $K \times K$ matrix as well as

$$\tilde{q}_{i\cdot} = \hat{q}_{i\cdot}S, \qquad \tilde{p}_{\cdot jm} = \hat{p}_{\cdot jm}(S^{-1})^\top.$$

Writing out the scalar product, $\tilde{q}_{i\cdot}\tilde{p}_{\cdot jm}^\top = \hat{q}_{i\cdot}S(\hat{p}_{\cdot jm}S^{-1})^\top = \hat{q}_{i\cdot}SS^{-1}\hat{p}_{\cdot jm}^\top = \hat{q}_{i\cdot}\hat{p}_{\cdot jm}^\top$, i.e. (2.3) holds. However, we have to be careful in order not to violate the side conditions (2.1). In order to achieve this, restrict S to a matrix with rows summing to 1, i.e. $S1_K^\top = 1_K^\top$. (If S has non-negative entries, it is called a stochastic matrix.) Multiplying this equation by S^{-1}, it is clear that $S^{-1}1_K^\top = 1_K^\top$ as well. Then, provided $\hat{q}$ and $\hat{p}$ satisfy (2.1), we have

$$\tilde{q}_{i\cdot}1_K^\top = \hat{q}_{i\cdot}S1^\top = \hat{q}_{i\cdot}1^\top = 1^\top, \qquad \tilde{p}_{k\cdot m}1_{J_m}^\top = \sum_{k'}(S^{-1})_{kk'}\hat{p}_{k'\cdot m}1_{J_m}^\top = \sum_{k'}(S^{-1})_{kk'} = 1,$$

i.e. $\tilde{q},\tilde{p}$ satisfy the side conditions (ii), (iv) as well. Moreover, if S is from

$$\mathbb{S}_{\hat{q},\hat{p}} := \Big\{S : S1_K = 1_K, \hat{q}_{i\cdot}S \geq 0 \text{ for all } i, \sum_{k'}\hat{p}_{k'jm}(S^{-1})_{kk'} \geq 0 \text{ for all } j,m\Big\},$$

conditions (i), (iii) are satisfied for $\tilde{q},\tilde{p}$ as well. So, we have shown that the MLE is not unique provided that $\mathbb{S}_{\hat{q},\hat{p}}$ is not trivial. Moreover, we can search $\mathbb{S}_{\hat{q},\hat{p}}$ for parameters which lead to equally likely IAs and allele frequencies.

A well-known but not exclusive example of this non-uniqueness deals with label-switching for populations. For this, consider a permutation matrix S. (Such a matrix has entries in $\{0,1\}$ with a single 1 in each row/column.) Here, $Sq^\top$ are IAs where we have switched labels $k \to l$ if $(S)_{kl} = 1$. Similarly, labels of allele frequencies p, switch label accordingly, i.e. using $S^{-1}p^\top$. For such matrices, conditions (i), (iii) are always met and it has been discussed frequently (see e.g. Behr et al., 2016; Ko et al., 2023) that $Sq, S^{-1}p$ is a MLE iff q,p is an MLE. However, the goal here is to identify the complete range of MLEs beyond label switching.

2.3 EMALAM

Our algorithm for exploring the set of MLEs starts with a single MLE $\hat{q},\hat{p}$, as e.g. provided by the output of STRUCTURE or ADMIXTURE. We then search for $S \in \mathbb{S}_{\hat{q},\hat{p}}$ as follows: Pick a subset $\mathcal{I} \subseteq \{1,...,N\}$ of individuals and (recalling that $H(x) = -\sum_k x_k \log x_k$ is the entropy of x which satisfies $x1^\top = 1$, which is maximal for the uniform distribution, and minimal for point measures) either

$$\begin{aligned} \text{(I): (a) minimize or (b) maximize } & \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(\hat{q}_{i\cdot}S)_k \text{ for some } k \\ \text{or (II): (a) minimize or (b) maximize } & \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}H(\hat{q}_{i\cdot}S). \end{aligned} \tag{2.4}$$

So, in (I), we look at a concrete population k and minimize / maximize the average IA in k for all individuals in $\mathcal{I}$. In (II), we rather minimize (favouring less admixed individuals) or maximize (favouring admixed individuals) the average entropy for individuals in $\mathcal{I}$.

2.4 Data applications

Since our main focus is on a proof of concept of non-identifyability in the admixture model, we use genetic data from the 1000 genomes project (Consortium et al., 2015). In our applications, we only use a small number of SNPs which were pre-selected in order to distinguish ancestry in forensic genetics, such that run times are hardly an issue. We consider the SNP-set by (Kidd et al., 2014) with 55 bi-allelic markers, and the VISAGE enhanced tool (Woźniak et al., 2021) with 104 markers, 73 of which are multi-allelic. We only used the 100 autosomal markers of this AIM set. Since ADMIXTURE can only handle bi-allelic markers, whereas STRUCTURE also allows for multiple alleles, we use STRUCTURE throughout all our analyses.

3 Results

3.1 Non-identifyability and allele frequency differences

Let us consider a toy example for the claimed non-identifyability. We simulate 1000 markers with different allele frequencies among 100 admixed individuals from two populations. The true IAs are depicted as the black line in Figure 1, and the range of MLEs as given by (3.1) are the shaded area. Depending on whether we use all five or just the three markers with small allele frequency differences, the range of MLEs varies. Only two, very good markers, influence the output of the estimation significantly. Figure S3 shows the connection between the difference of the allele frequencies in two populations and the range of MLEs. Recall that markers with are (almost) fixed in one, and polymorphic in all other populations, are called anchor markers. Equivalently, individuals with IA almost exclusively in one population, are called anchor individuals (Cabreros and Storey, 2019).

3.2 Some theoretical insights

The case $K = 2$

The case $K = 2$ is important from an application perspective (Friedlaender et al., 2008; Rosenberg et al., 2005; Lawson et al., 2018; Hamlin and Arnold, 2014; Phillips et al., 2009). In this case, starting with a MLE $\hat{q}, \hat{p}$, we give for each individual i the range of possible IAs with the same likelihood.

We show in Theorem 1 in Appendix 8 that the MLEs for IA of individual i in population 1 are in

$$\left[\hat{q}_{i1}\frac{u^* - 1}{u^* + v_*} + \hat{q}_{i2}\frac{1 - u^*}{1 + u^*/v_*}; \hat{q}_{i1}\frac{1 + v^*}{u_* + v^*} + \hat{q}_{i2}\frac{1 + v^*}{1 + v^*/u_*}\right] \tag{3.1}$$

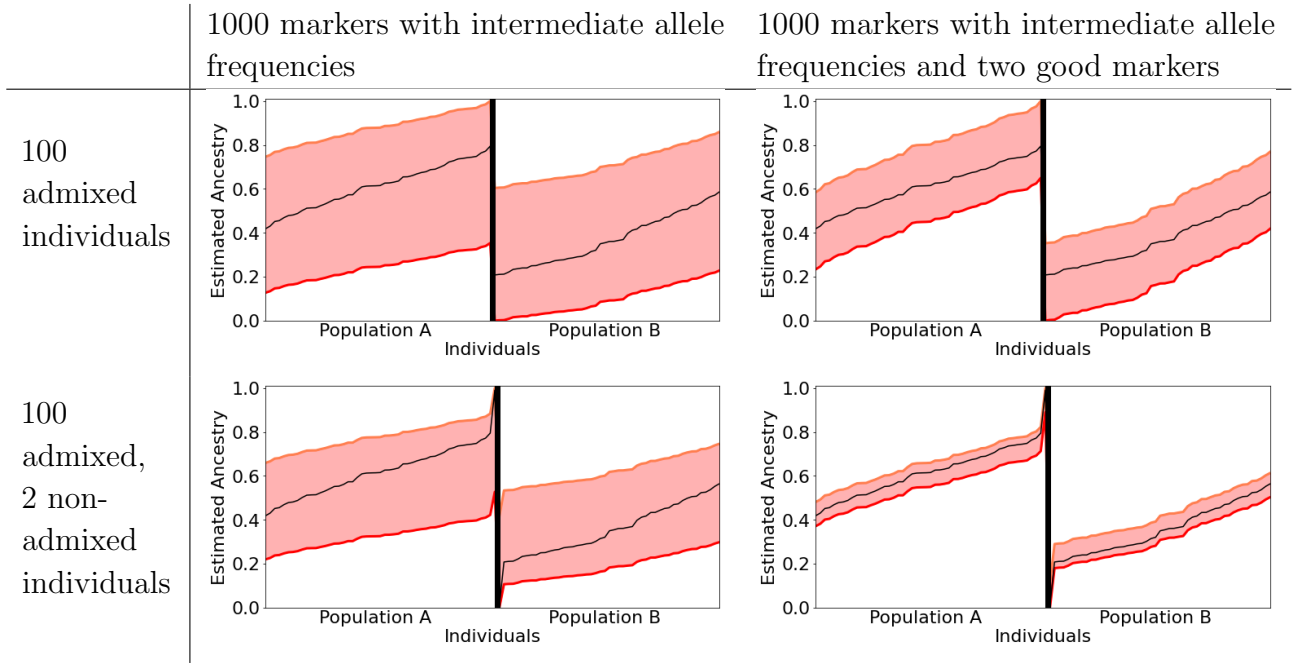

Figure 1: Range of equally likely MLEs for two simulated populations. The x-axis represents the individuals and the y-axis indicates the IA for population A. The black lines show the true simulated individual ancestries. The area between the red lines symbolizes the possible MLEs for this case. The dark (light) red line at the bottom (top) represents the MLE with the minimal (maximal) IA for Population A. The estimated allele frequencies in population A for 1000 bi-allelic markers were simulated according to a uniform distribution in [0.25, 0.75]. The allele frequencies in population B are $p_{m,2}+\epsilon, \epsilon \sim \mathcal{U}([0,0.3])$, if $p_{m,2}+\epsilon \in [0.2, 0.8]$ and $p_{m,2} = 0.5$ else. Additionally, we added in the right column two additional markers with frequencies $(0.01, 0.1), (0.9, 0.99)$. In the second row, we included two non-admixed individuals with ancestry $(0.01, 0.99)$ and $(0.99, 0.01)$.

with

$$\hat{u}_* := \min\left\{\frac{\hat{p}_{2jm}}{\hat{p}_{1jm}} : m = 1, ..., M, j = 1, ..., J_m\right\},$$
$$\hat{u}^* := \max\left\{\frac{\hat{p}_{2jm}}{\hat{p}_{1jm}} : m = 1, ..., M, j = 1, ..., J_m\right\},$$
$$\hat{v}_* := \min\left\{\frac{\hat{q}_{i1}}{\hat{q}_{i2}} : i = 1, ..., N\right\},$$
$$\hat{v}^* := \max\left\{\frac{\hat{q}_{i1}}{\hat{q}_{i2}} : i = 1, ..., N\right\}.$$

Most notably, the range of possible MLEs of individual i only depends on $\hat{q}_i, \hat{u}_*, \hat{u}^*, \hat{v}_*$ and $\hat{v}^*$. So, adding markers m' or individuals i' with less extreme quotients $\hat{p}_{2jm'}/\hat{p}_{1jm'}$ and $\hat{q}_{i'1}/\hat{q}_{i'2}$ than markers/individuals already present in the dataset does not solve the problem of non-identifyability per se. From the formula, the lower bound is near $\hat{q}_{i1}$ provided that both, $v_* \ll 1 \ll u^*$, and the upper bound is close to $\hat{q}_{i1}$ iff $u_* \ll 1 \ll v^*$. In other words, we need

both, markers which can distinguish well between both populations, leading to $u_* \ll 1 \ll u^*$, as well as individuals which are almost non-admixed for both populations, leading to $v_* \ll 1 \ll v^*$. In other words, (3.1) gives a more quantitative view on the fact that anchor individuals and anchor markers are necessary for uniqueness of the MLE; see the anchor conditions in (Cabreros and Storey, 2019).

The case $K > 2$

In this case, we did not obtain an analytical solution for the possible IAs. However, we note that only changing $\hat{p}$ and $\hat{q}$ in two populations could be treated in the same way as the case $K = 2$. However, this only leads to a lower bound for the set of possible IAs. Therefore, we rely on numerical minimizations and maximizations in cases (I) and (II) for $K > 2$ in order to find extreme, but equally likely estimators $\tilde{q}, \tilde{p}$. Some more information can be found in the Documentation of EMALAM.

3.3 Non-identifyability in the 1000 genomes data

We apply EMALAM to the data from the 1000 Genomes Project (Consortium et al., 2015). Additionally, we consider the difference between our method and running STRUCTURE many times as proposed by (Behr et al., 2016).

In our workflow, we run STRUCTURE once for the data, and then study the range of MLEs using the STRUCTURE estimator. We use data from northern Europeans from Utah (CEU, 99 individuals), Great Britain (GBR, 91 individuals), Iberian Populations in Spain (IBS, 107 individuals), Finland (FIN, 99 individuals) and Tuscans from Italy (TSI, 107 individuals). Furthermore, to compare our methods to previous approaches, we run STRUCTURE 20 times and apply pong (Behr et al., 2016) to the outputs to avoid label switching.

First, we consider two ancestral populations, $K = 2$, using markers from Kidd et al. (2014). Using only data from GBR and IBS, Figure 2 shows the maximal and minimal IAs (the red lines). We compared the results to running STRUCTURE many times (blue and green lines in Figure 2). However, the differences between these runs are much smaller than the differences between the MLEs in our approach. There are many values for the IA with the same likelihood as the output of STRUCTURE, e.g. for individual 1 every IA between 0.03 and 0.52 is possible.

Similarly, we consider FIN and IBS with $K = 2$. The output of STRUCTURE for these populations is shows in Figure 3 (blue line). According to pong, there is only one mode. Hence, we only have one blue line. Furthermore, we present the maximal and minimal IAs that is a MLE from Finland by the red lines. The difference between the possible estimated IAs are small. The reason for this is that the minimal estimated IA is very small and the maximal estimated IA is large and that there are markers with large estimated allele differences.

Second, we use $K = 3$ and markers from the VISAGE enhanced tool (Woźniak et al., 2021). We apply EMALAM to the data from CEU, TSI and IBS. For every picture we ordered the individual according to the estimated part of the genomes from CEU for $K = 3$ in the major mode. Especially, the individuals at position $i = 1, \ldots, 313$ are for every figure the same. In

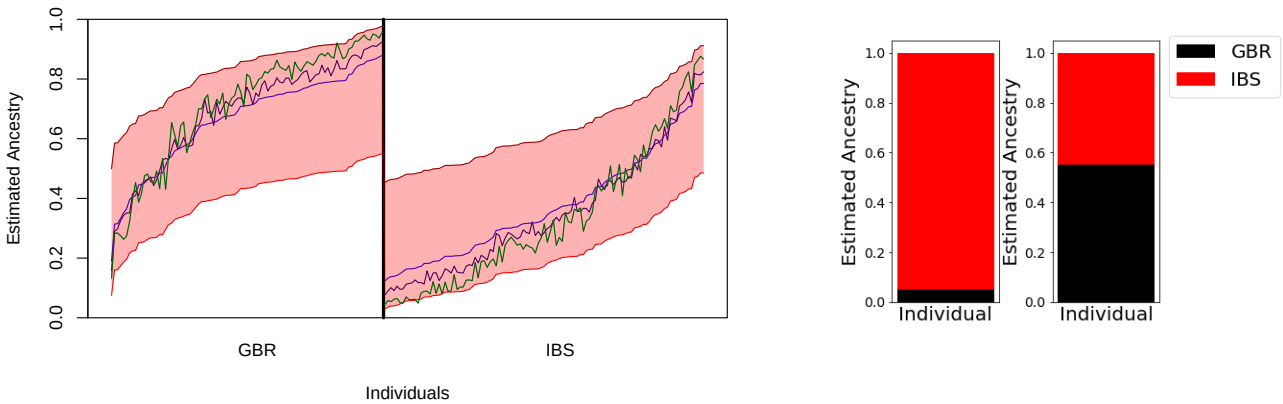

Figure 2: Range of MLEs for GBR and IBS populations in the admixture model with $K = 2$. Left: Estimated IAs for populations GBR and IBS among the MLEs. The red lines represent the minimal and maximal MLE for the IA from GBR, if we apply the formula (3.1). The three colored lines depict the output of 20 runs of STRUCTURE. The software pong was used to only depict the three result modes that are not similar. Here, the green and purple line represent minor modes, while the blue line represents the major mode. Right: Estimated IAs for the populations IBS and GBR with $K = 2$ for individual HG00096 from GBR, i.e. the first individual on the left side. Both barplots show the same individual with the largest MLE from GBR (dark red line in left panel) and the smallest MLE from GBR (light red line in left panel).

Figure 4, the data was created by running STRUCTURE once and by using (I) (a) and (b) (recall from (2.4)) to the blue population and (II) (a) and (b) to the output of STRUCTURE for $K = 3$.

We compare the results to running STRUCTURE 20 times and apply pong to the outputs to avoid label switching (Figure 5). Specifically, every row is one mode, i.e. one MLE, according to pong. We see that this values and the difference between the runs are much less extreme than in the results that are generated by EMALAM.

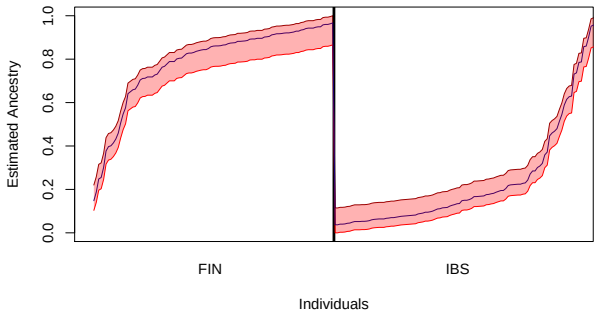

Figure 3: Range of MLEs for FIN and IBS populations in the admixture model with $K = 2$. Maximal and minimal possible estimated IAs for FIN and IBS with $K = 2$ if we apply (3.1) and use the markers from (Kidd et al., 2014). We also run STRUCTURE 20 time and use pong to only depict similar runs. This consequences just one mode (blue line).

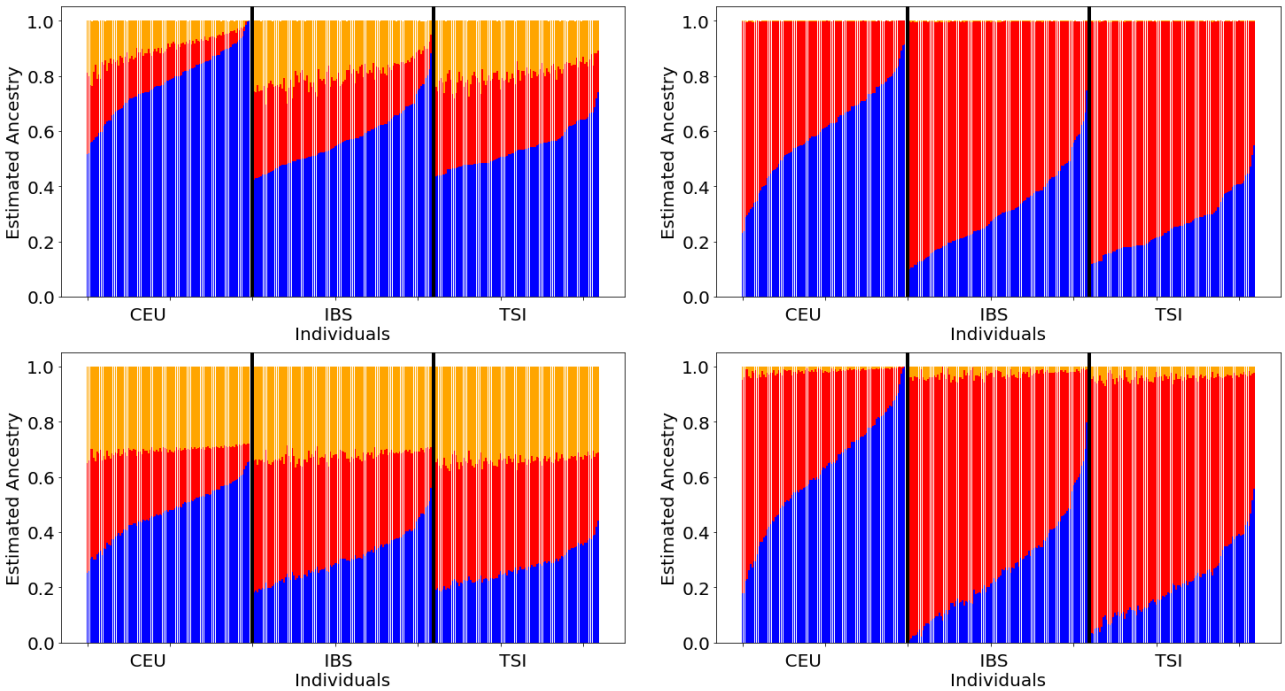

Figure 4: Estimated IAs for the TSI, IBS, and CEU populations. All four figures show estimators with the same likelihood. We used EMALAM with $K = 3$ in order to optimize the following target in the four figures: Top Left: Maximize the ancestry from the blue population (option (I)(b)). Bottom Left: Maximize the entropy (favouring admixd individuals), i.e. option (II)(b). Top Right: Minimize the ancestry from the blue population (option (I)(a)). Bottom Right: Minimize the entropy (favouring non-admixed individuals), i.e. option (II)(a).

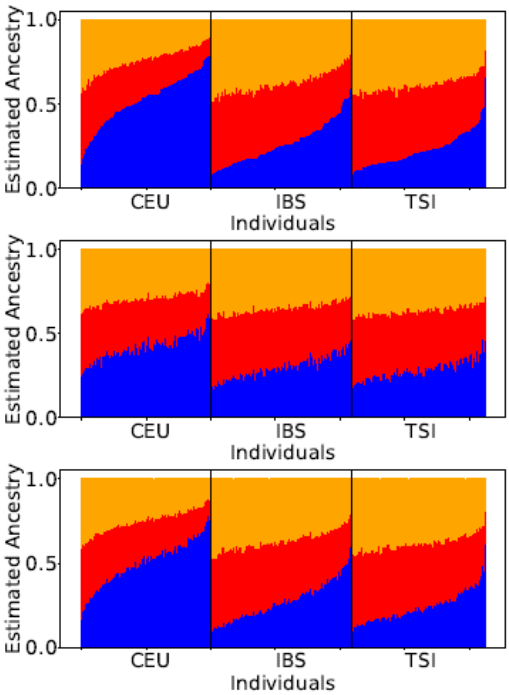

Figure 5: Estimated IAs with the populations CEU, IBS, TSI, K = 3. We ran STRUCTURE 20 times and used pong (Behr et al., 2016) to avoid label switching between the modes. The three resulting modes are shown.

4 Discussion

We demonstrated how the non-uniqueness of the maximum likelihood approach in the admixture model impacts the range of equally good MLEs. The typical approach to explore the flat likelihood of the admixture model by running STRUCTURE/ADMIXTURE several times (Behr et al., 2016) only partially reveals the range of MLEs.

Here, we introduced EMALAM, a method to explore the complete range of equally good MLEs. In addition, EMALAM is able to merge MLEs that occur due to label switching; compare with pong Behr et al. (2016). We stress that in general, the non-uniqueness of the MLE has way more possibilities than just switching labels. Since EMALAM relies on a numerical minimization of a $K \times K$-matrix, it is faster than running STRUCTURE or ADMIXTURE multiple times. Furthermore, it identifies the most extreme MLEs with respect to specific aspects, i.e. the IAs of a single individual or the admixture of a complete population.

In general, anchor markers, which have fixed alleles in only one population, and anchor individuals with non-admixed ancestry, play a vital role in identifyability of the admixture model. If such anchors are missing, the range of MLEs which EMALAM computes can be too large for a useful interpretation of the results provided by STRUCTURE or ADMIXTURE. Frequently, the lack of anchors goes hand in hand with choosing a too large number of ancestral populations. It is easy to see that the number of optima increases if K increases, i.e. in the admixture model K should be chosen as small as possible to prevent non-uniqueness. However, if suitable markers and individuals and a small enough K are considered the difference between

the MLEs can be small (see Figure 1).

EMALAM helps to investigate the range of MLEs and how meaningful the inferred ancestries are. As a rule of thumb, similar estimated IA or similar estimated allele frequencies lead to a large range of possible estimated IAs with the same likelihood as the output of STRUCTURE.

Our results highlight that a larger number of markers alone does not necessarily lead to a meaningful IA inference in the admixture model. It is crucial to include at least some markers with high allele differences between the populations of interest and to include for each population individuals with low admixture proportions. This implies that even when using thousands of markers (or individuals), the certainty of IA inference in the admixture model can rely on only a few ancestry informative markers (Figure 1). Consequently, for populations where admixture occurs frequently or no informative markers exist the range of MLEs in the admixture model can be larger than considering the different modes of multiple STRUCTURE runs suggests.

Using alternative models and methods that restrict the parameter space, e.g., by considering linkage disequilibrium and local ancestry (Falush et al., 2003; Corbett-Detig and Nielsen, 2017), or by assuming dependency structures between the ancestral populations (Bradburd, Coop, and Ralph, Bradburd et al.; Burger et al., 2024) can circumvent the non-identifiability of the admixture model. However, while more complex models can lead to more precise IAs and allele frequency estimates they typically rely on a larger number of markers or specific model assumptions. Instead, EMALAM can be used to assess the consequences of the non-identifiable MLEs in applications based on the admixture model. Especially for scenarios with small allele frequency differences between populations EMALAM will help to prevent misinterpretation of ADMIXTURE and STRUCTURE results.

5 Funding

CSH is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 499552394 – SFB Small Data. FB is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645, and EXC 2124 – Project number 390838134.

6 Data Availability Statement

The data is available at the 1000 Genomes Project website (Consortium et al., 2015). The implementation of EMALAM can be downloaded from the GitHub repository. An online version is also available.

7 Acknowledgments

We thank Jakob Stiefel for some early work on this project. We also thank Ferdinand Suchanek for extracting the markers of the VISAGE enhanced tool from the 1000 genomes data.

8 Conflict of Interest

The authors declare that there are no conflicts of interests.

References

Alexander, D. H. and K. Lange (2011). Enhancements to the admixture algorithm for individual ancestry estimation. *BMC bioinformatics 12*, 1–6.

Alexander, D. H., J. Novembre, and K. Lange (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research 19*(9), 1655–1664.

Badouin, H., J. Gouzy, C. J. Grassa, F. Murat, S. E. Staton, L. Cottret, C. Lelandais-Brière, G. L. Owens, S. Carrère, B. Mayjonade, et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. *Nature 546*(7656), 148–152.

Battey, CJ., P. L. Ralph, and A. D. Kern (2020). Predicting geographic location from genetic variation with deep neural networks. *eLife 9*, e54507.

Behr, A. A., K. Z. Liu, G. Liu-Fang, P. Nakka, and S. Ramachandran (2016). Pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics 32*(18), 2817–2823.

Bradburd, G. S., G. M. Coop, and P. L. Ralph. Inferring Continuous and Discrete Population Genetic Structure Across Space.

Burger, K. E., S. Klepper, U. von Luxburg, and F. Baumdicker (2024). Inferring Ancestry with the Hierarchical Soft Clustering Approach tangleGen.

Cabreros, I. and J. D. Storey (2019). A likelihood-free estimator of population structure bridging admixture models and principal components analysis. *Genetics 212*(4), 1009–1029.

Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature 526*(7571), 68.

Corbett-Detig, R. and R. Nielsen (2017). A hidden markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLoS genetics 13*(1), e1006529.

Dharmadhikari, S. and K. Joag-Dev (1985). Examples of nonunique maximum likelihood estimators. *The American Statistician 39*(3), 199–200.

Dominguez Mantes, A., D. Mas Montserrat, C. D. Bustamante, X. Giró-i Nieto, and A. G. Ioannidis (2023a). Neural admixture for rapid genomic clustering. *Nature computational science 3*(7), 621–629.

Dominguez Mantes, A., D. Mas Montserrat, C. D. Bustamante, X. Giró-i Nieto, and A. G. Ioannidis (2023b). Neural admixture for rapid genomic clustering. *Nature computational science 3*(7), 621–629.

Dominguez Mantes, A., D. Mas Montserrat, C. D. Bustamante, X. Giró-i-Nieto, and A. G. Ioannidis (2023). Neural ADMIXTURE for rapid genomic clustering. *Nature Computational Science 3*(7), 621–629.

Earl, D. A. and B. M. VonHoldt (2012). Structure harvester: a website and program for visualizing structure output and implementing the evanno method. *Conservation genetics resources 4*, 359–361.

Evanno, G., S. Regnaut, and J. Goudet (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular ecology 14*(8), 2611–2620.

Falush, D., M. Stephens, and J. K. Pritchard (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics 164*(4), 1567–1587.

Falush, D., M. Stephens, and J. K. Pritchard (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular ecology notes 7*(4), 574–578.

Friedlaender, J. S., F. R. Friedlaender, F. A. Reed, K. K. Kidd, J. R. Kidd, G. K. Chambers, R. A. Lea, J.-H. Loo, G. Koki, J. A. Hodgson, et al. (2008). The genetic structure of pacific islanders. *PLoS genetics 4*(1), e19.

Garcia-Erill, G. and A. Albrechtsen (2020). Evaluation of model fit of inferred admixture proportions. *Molecular ecology resources 20*(4), 936–949.

Hamlin, J. A. and M. L. Arnold (2014). Determining population structure and hybridization for two iris species. *Ecology and Evolution 4*(6), 743–755.

Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard (2009). Inferring weak population structure with the assistance of sample group information. *Molecular ecology resources 9*(5), 1322–1332.

Jakobsson, M. and N. A. Rosenberg (2007). Clumpp: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics 23*(14), 1801–1806.

Janes, J. K., J. M. Miller, J. R. Dupuis, R. M. Malenfant, J. C. Gorrell, C. I. Cullingham, and R. L. Andrew (2017). The k= 2 conundrum. *Molecular Ecology 26*(14), 3594–3602.

Jia, N., J. Wang, W. Shi, L. Du, Y. Sun, W. Zhan, J.-F. Jiang, Q. Wang, B. Zhang, P. Ji, et al. (2020). Large-scale comparative analyses of tick genomes elucidate their genetic diversity and vector capacities. *Cell 182*(5), 1328–1340.

Jombart, T., S. Devillard, and F. Balloux (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics 11*(1), 1–15.

Kidd, K. K., W. C. Speed, A. J. Pakstis, M. R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F. R. Friedlaender, and J. R. Kidd (2014). Progress toward an efficient panel of snps for ancestry inference. *Forensic Science International: Genetics 10*, 23–32.

Klei, L., B. P. Kent, N. Melhem, B. Devlin, and K. Roeder (2011). Gemtools: a fast and efficient approach to estimating genetic ancestry.

Ko, S., B. B. Chu, D. Peterson, C. Okenwa, J. C. Papp, D. H. Alexander, E. M. Sobel, H. Zhou, and K. L. Lange (2023). Unsupervised discovery of ancestry-informative markers and genetic admixture proportions in biobank-scale datasets. *The American Journal of Human Genetics 110*(2), 314–325.

Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across k. *Molecular ecology resources 15*(5), 1179–1191.

Larena, M., J. McKenna, F. Sanchez-Quinto, C. Bernhardsson, C. Ebeo, R. Reyes, O. Casel, J.-Y. Huang, K. P. Hagada, D. Guilay, et al. (2021). Philippine ayta possess the highest level of denisovan ancestry in the world. *Current Biology 31*(19), 4219–4230.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush (2012). Inference of population structure using dense haplotype data. *PLoS genetics 8*(1), e1002453.

Lawson, D. J., L. Van Dorp, and D. Falush (2018). A tutorial on how not to over-interpret structure and admixture bar plots. *Nature communications 9*(1), 3258.

Lee, A. B., D. Luca, L. Klei, B. Devlin, and K. Roeder (2010). Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 34*(1), 51–59.

Liu, X., N. M. Kopelman, and N. A. Rosenberg (2024). Clumppling: cluster matching and permutation program with integer linear programming. *Bioinformatics 40*(1), btad751.

Lovell, J. T., A. H. MacQueen, S. Mamidi, J. Bonnette, J. Jenkins, J. D. Napier, A. Sreedasyam, A. Healey, A. Session, S. Shu, et al. (2021). Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature 590*(7846), 438–444.

Margaryan, A., D. J. Lawson, M. Sikora, F. Racimo, S. Rasmussen, I. Moltke, L. M. Cassidy, E. Jørsboe, A. Ingason, M. W. Pedersen, et al. (2020). Population genomics of the viking world. *Nature 585*(7825), 390–396.

Mogensen, H. S., T. Tvedebrink, C. Børsting, V. Pereira, and N. Morling (2020). Ancestry prediction efficiency of the software genogeographer using a z-score method and the ancestry informative markers in the precision id ancestry panel. *Forensic Science International: Genetics 44*, 102154.

Pfaffelhuber, P., F. Grundner-Culemann, V. Lipphardt, and F. Baumdicker (2020). How to choose sets of ancestry informative markers: A supervised feature selection approach. *Forensic Science International: Genetics 46*, 102259.

Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. *Forensic Science International: Genetics 18*, 49–65.

Phillips, C., W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, et al. (2014). Building a forensic ancestry panel from the ground up: The euroforgen global aim-snp set. *Forensic Science International: Genetics 11*, 13–25.

Phillips, C., L. Prieto, M. Fondevila, A. Salas, A. Gómez-Tato, J. Álvarez-Dios, A. Alonso, A. Blanco-Verea, M. Brión, M. Montesino, et al. (2009). Ancestry analysis in the 11-m madrid bomb attack investigation. *PloS one 4*(8), e6583.

Pritchard, J. K. and P. Donnelly (2001). Case–control studies of association in structured or admixed populations. *Theoretical population biology 60*(3), 227–237.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). Inference of population structure using multilocus genotype data. *Genetics 155*(2), 945–959.

Pritchard, J. K., X. Wen, and D. Falush (2010). Documentation for structure software: Version 2.3.

Puechmaille, S. J. (2016). The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Molecular ecology resources 16*(3), 608–627.

Raj, A., M. Stephens, and J. K. Pritchard (2014). faststructure: variational inference of population structure in large snp data sets. *Genetics 197*(2), 573–589.

Rohlfs, R. V., S. M. Fullerton, and B. S. Weir (2012). Familial identification: population structure and relationship distinguishability. *PLoS Genetics 8*(2), e1002469.

Rosenberg, N. A. (2004). Distruct: a program for the graphical display of population structure. *Molecular ecology notes 4*(1), 137–138.

Rosenberg, N. A., L. M. Li, R. Ward, and J. K. Pritchard (2003). Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics 73*(6), 1402–1422.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS genetics 1*(6), e70.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman (2002). Genetic structure of human populations. *science 298*(5602), 2381–2385.

Shringarpure, S. S., C. D. Bustamante, K. Lange, and D. H. Alexander (2016). Efficient analysis of large datasets and sex bias with admixture. *BMC bioinformatics 17*(1), 1–6.

Tang, H., J. Peng, P. Wang, and N. J. Risch (2005). Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 28*(4), 289–301.

Tvedebrink, T. (2022). Review of the forensic applicability of biostatistical methods for inferring ancestry from autosomal genetic markers. *Genes 13*(1), 141.

van Waaij, J. (2022). Necessary and sufficient conditions for identifiability in the admixture model. *arXiv preprint arXiv:2202.05540*.

Verdu, P., T. J. Pemberton, R. Laurent, B. M. Kemp, A. Gonzalez-Oliver, C. Gorodezky, C. E. Hughes, M. R. Shattuck, B. Petzelt, J. Mitchell, et al. (2014). Patterns of admixture and population structure in native populations of northwest north america. *PLoS genetics 10*(8), e1004530.

Verity, R. and R. A. Nichols (2016). Estimating the number of subpopulations (k) in structured populations. *Genetics 203*(4), 1827–1839.

Wang, J. (2017). The computer program structure for assigning individuals to populations: easy to use but easier to misuse. *Molecular ecology resources 17*(5), 981–990.

Wasser, S. K., C. Mailand, R. Booth, B. Mutayoba, E. Kisamo, B. Clark, and M. Stephens (2007). Using dna to track the origin of the largest ivory seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences 104*(10), 4228–4233.

Wollstein, A. and O. Lao (2015). Detecting individual ancestry in the human genome. *Investigative genetics 6*(1), 1–12.

Woźniak, A., A. Heidegger, D. Piniewska-Róg, E. Pośpiech, C. Xavier, A. Pisarek, E. Kartasińska, M. Boroń, A. Freire-Aradas, M. Wojtas, et al. (2021). Development of the visage enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones. *Aging (Albany NY) 13*(5), 6459.

Supplemental Material

A Theoretical results

A.1 Two ancestral populations

Of course, $K = 2$ in this section, so we suppress indices of 2 wherever possible. In addition, we assume we are given some $\hat{q} = (\hat{q}_i)_{i=1,\dots,I}$, i.e. some estimator for all IAs for all individuals, as well as $\hat{p} = (\hat{p}_{kjm})_{k=1,\dots K, j=1,\dots,J_m, m=1,\cdots,M}$, i.e. all estimated allele frequencies at all markers satisfying (2.1). We will use vector notation whenever possible, e.g. $p \geq 0$ for some vector p means that all entries in p are ≥ 0. We start by noting that for $S \in \mathbb{S}_{\hat{q},\hat{p}}$, we will require $\hat{q}S1^\top = 1^\top$, which we achieve by assuming that $S1^\top = 1^\top$, so we will work with S and the corresponding S^{-1} of the form

$$S = S(a,b) = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}, \qquad S^{-1} = S^{-1}(a,b) = \frac{1}{1-a-b}\begin{pmatrix} 1-b & -a \\ -b & 1-a \end{pmatrix}.$$

For such S, we set

$$\tilde{q} := \tilde{q}(a,b) = \hat{q}S(a,b), \qquad \tilde{p} = \hat{p}(S^{-1}(a,b))^\top \text{ or } \tilde{p}^\top = S^{-1}(a,b)\hat{p}^\top.$$

Since $S(a,b) \in \mathbb{S}_{\hat{q},\hat{p}}$ iff $S(1-a,1-b) = S(a,b) \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathbb{S}_{\hat{q},\hat{p}}$ (i.e. switching population labels leads to an additional solution), we prevent finding the solution with switched labels by using the condition $a + b \leq 1$, i.e. we do not consider the additional solution $(1-a, 1-b)$. So, our goal is to determine the set

$$\mathbb{L} := \{(a,b) : a + b \leq 1,\ \tilde{q}(a,b) \geq 0,\ \tilde{p}(a,b) \geq 0\}.$$

(Since summing q (over k) and p (over j) equals one as explained in Section 2, this suffices also for $\tilde{q}(a,b) \leq 1$ and $\tilde{p} \leq 1$.)

For the statement of our Theorem, we need the following notation:

$$\begin{aligned}
\hat{u}_* &:= \min\Big\{\frac{\hat{p}_{2jm}}{\hat{p}_{1jm}} : m = 1,\dots,M, j = 1,\dots,J_m\Big\}, \\
\hat{u}^* &:= \max\Big\{\frac{\hat{p}_{2jm}}{\hat{p}_{1jm}} : m = 1,\dots,M, j = 1,\dots,J_m\Big\}, \\
\hat{v}_* &:= \min\Big\{\frac{\hat{q}_{i1}}{\hat{q}_{i2}} : i = 1,\dots,I\Big\}, \\
\hat{v}^* &:= \max\Big\{\frac{\hat{q}_{i1}}{\hat{q}_{i2}} : i = 1,\dots,I\Big\}.
\end{aligned}$$

Theorem 1 (Range for $\tilde{q}$ and $\tilde{p}$)**.** *The set $\mathbb{L}$ is as given in Figure S1. Moreover, (a^*, b^*) and (a_*, b_*) maximizing and minimizing $\tilde{q}_1$ in all individuals are*

$$a^* = \frac{u_* - 1}{u_* + v^*}, \qquad b^* = \frac{1 + v^*}{1 + v^*/u_*}, \tag{A.1}$$

$$a_* = \frac{1 + v_*}{u^* + v_*}, \qquad b_* = \frac{1 - u^*}{1 + u^*/v_*}. \tag{A.2}$$

and for individual i, the resulting possible $\tilde{q}_{i1}$ are

$$\tilde{q}_{i1} \in \left[\hat{q}_{i1}\frac{u^* - 1}{u^* + v_*} + \hat{q}_{i2}\frac{1 - u^*}{1 + u^*/v_*}; \hat{q}_{i1}\frac{1 + v^*}{u_* + v^*} + \hat{q}_{i2}\frac{1 + v^*}{1 + v^*/u_*}\right]. \tag{A.3}$$

In addition, the resulting possible $\tilde{p}_{kjm}$'s are

$$\begin{aligned} &\textit{between } \hat{p}_{1jm}\frac{u^*}{u^* - 1} - \frac{1}{u^* - 1}\hat{p}_{2jm} \textit{ and } -\hat{p}_{1jm}\frac{v^*}{1 + v^*} + \frac{1}{1 + v^*}\hat{p}_{2jm} \textit{ for } \tilde{p}_{1jm}, \\ &\textit{between } -\frac{u_*}{1 - u_*}\hat{p}_{1jm} + \frac{1}{1 - u_*}\hat{p}_{2jm} \textit{ and } \frac{v_*}{1 + v_*}\hat{p}_{1jm} + \frac{1}{1 + v_*}\hat{p}_{2jm} \textit{ for } \tilde{p}_{2jm}. \end{aligned} \tag{A.4}$$

Proof.

Step 0: Computing $\mathbb{L}$: As explained in Section 2, our task is to find S such that $\hat{q}S$ and $\hat{p}(S^{-1})^\top$ are non-negative. So, for $S = S(a, b)$, we find for $\hat{q} = (\hat{q}_1, \hat{q}_2)$ and $\hat{p} = (\hat{p}_1, \hat{p}_2)$ the conditions

$$\hat{q}S = \big(\hat{q}_1(1 - a) + b\hat{q}_2, \hat{q}_1 a + \hat{q}_2(1 - b)\big) \geq 0, \tag{A.5}$$

$$p(S^{-1})^\top = \frac{1}{1 - a - b}\big(\hat{p}_1(1 - b) - a\hat{p}_2, -b\hat{p}_1 + \hat{p}_2(1 - a)\big) \geq 0. \tag{A.6}$$

From these equations, we find the equivalent conditions (note that we are using $a + b \leq 1$ in the second line)

$$\begin{aligned} &-\frac{\hat{q}_1}{\hat{q}_2} + \frac{\hat{q}_1}{\hat{q}_2}a \leq b \leq 1 + \frac{\hat{q}_1}{\hat{q}_2}a \\ &b \leq \left(1 - a\frac{\hat{p}_2}{\hat{p}_1}\right) \wedge \left(\frac{\hat{p}_2}{\hat{p}_1} - a\frac{\hat{p}_2}{\hat{p}_1}\right) = \left(1 \wedge \frac{\hat{p}_2}{\hat{p}_1}\right) - a\frac{\hat{p}_2}{\hat{p}_1}. \end{aligned}$$

Since all conditions are linear, they are satsfied if and only if they hold for the maximal and minimal possible values for $\frac{\hat{q}_1}{\hat{q}_2}$ and $\frac{\hat{p}_2}{\hat{p}_1}$, i.e.

$$\begin{aligned} b &\geq -v_* + v_* a, \\ b &\geq -v^* + v^* a, \\ b &\leq 1 + v_* a, \\ b &\leq 1 + v^* a, \\ b &\leq u_* - a u_*, \\ b &\leq 1 - a u^*. \end{aligned}$$

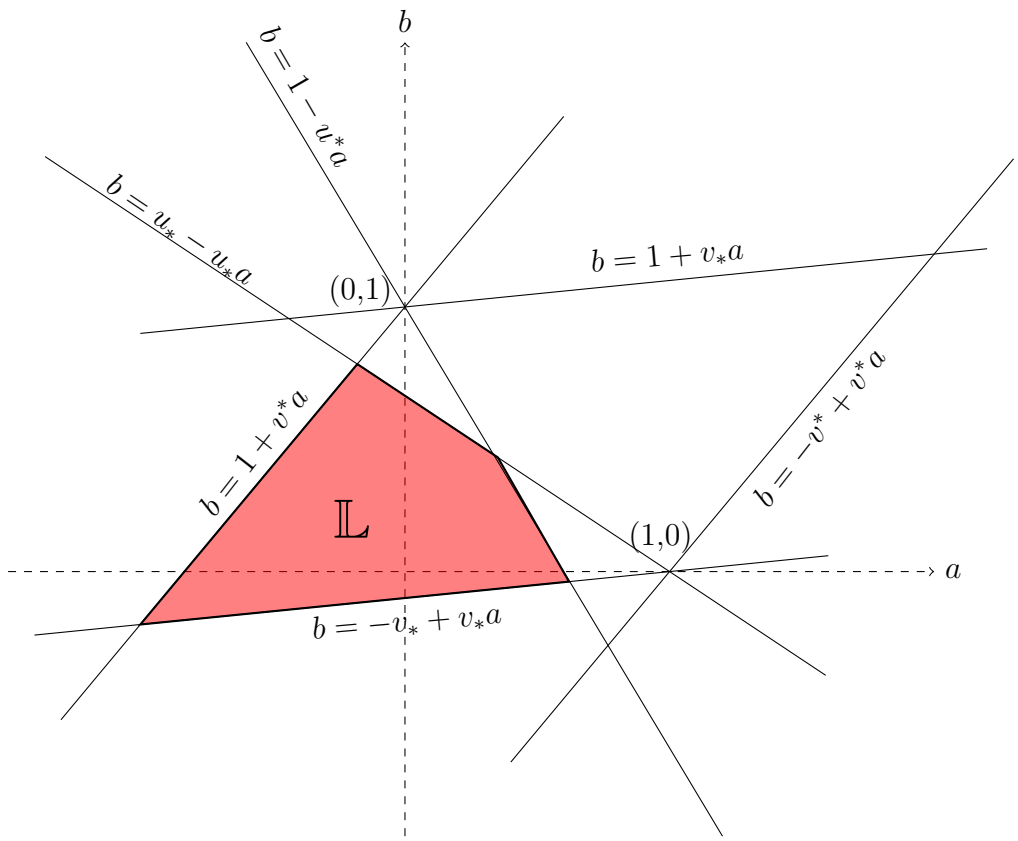

Figure S1: The set of (a, b), such that $\tilde{q} = \hat{q}S(a, b)$ and $\tilde{p} = \hat{p}(S(a, b)^{-1})^\top$ satisfies the conditions (2.1), is given by $\mathbb{L}$.

We have used that $\hat{u}_* \leq 1$ and $\hat{u}^* \geq 1$ by construction: if $\hat{p}_{2jm} \leq \hat{p}_{1jm}$ for some j, m, there must be j' with $\hat{p}_{2j'm} \geq \hat{p}_{1j'm}$ since summing over j is 1. We have illustrated all inequalities and the resulting $\mathbb{L}$ in Figure S1. Note that $a + b \leq 1$ is satisfied in $\mathbb{L}$.

Step 1: Maximizing and minimizing $\tilde{q}_{i1}$: We have to find combinations $(a, b) \in \mathbb{L}$ minimizing and maximizing $\tilde{q}_{i1} = \hat{q}_{i1}(1-a) + b\hat{q}_{i2}$.

In order to maximize $\tilde{q}_{i1} = \hat{q}_{i1}(1-a) + b\hat{q}_{i2}$, note that a maximizer $(a^*, b^*) \in \mathbb{L}$ must have the property that no $(a', b') \in \mathbb{L}$ with $a' < a^*, b' > b$ exists (since $\tilde{q}_{i1}(a', b') > \tilde{q}_{i1}(a^*, b*)$ in this case, contradicting the maximization property of (a^*, b^*)). This already implies that (a^*, b^*) is located on the line $b = 1 + v^* a$, since it is the line bounding $\mathbb{L}$ in the north-west. We find along this line

$$\tilde{q}_{i1}(a, 1 + v^* a) = \hat{q}_{i1}(1-a) + (1 + v^* a)\hat{q}_{i2} = \hat{q}_{i1} + \hat{q}_{i2} + \hat{q}_{i2}\big(v^* - \frac{\hat{q}_{i1}}{\hat{q}_{i2}}\big)a.$$

Since $v^* - \frac{\hat{q}_{i1}}{\hat{q}_{i2}} \geq 0$ by construction, $a \mapsto \tilde{q}_{i1}(a, 1 + v^* a)$ is increasing and the maximizer is at

$$u_*(1 - a^*) = 1 + v^* a^*,$$

i.e.

$$a^* = \frac{u_* - 1}{u_* + v^*}, \qquad b^* = 1 + v^* a^* = 1 + \frac{u_* v^* - v^*}{u_* + v^*} = \frac{u_*(1 + v^*)}{u_* + v^*} = \frac{1 + v^*}{1 + v^*/u_*},$$

which shows (A.1).

For the minimizer of $\tilde{q}_{i1} = \hat{q}_{i1}(1-a) + b\hat{q}_{i2}$, we analogously note that a minimizer $(a_*, b_*) \in \mathbb{L}$ must have the property that there is no $(a', b') \in \mathbb{L}$ with $a' > a^*, b' < b$. This implies that (a_*, b_*) is located on $b = -v_* + v_* a$, since this is the south-eastern boundary of $\mathbb{L}$. We find

$$\tilde{q}_{i1}(a, -v_* + v_* a) = \hat{q}_{i1}(1-a) - v_*(1-a)\hat{q}_{i2} = \hat{q}_{i2}\big(\frac{\hat{q}_{i1}}{\hat{q}_{i2}} - v_*\big)(1-a).$$

Since $\frac{\hat{q}_{i1}}{\hat{q}_{i2}} \geq v_*$ by construction, $a \mapsto \tilde{q}_{i1}(a, 1 + v^* a)$ is decreasing and the minimizer (a_*, b_*) is at

$$-v_*(1 - a_*) = 1 - u^* a_*,$$

i.e.

$$a_* = \frac{1 + v_*}{u^* + v_*}, \qquad b_* = 1 - u^* a_* = 1 - \frac{u^* + u^* v_*}{u^* + v_*} = \frac{v_*(1 - u^*)}{u^* + v_*} = \frac{1 - u^*}{1 + u^*/v_*},$$

giving (A.2).

Plugging (a^*, b^*) into $\tilde{q}_{i1} = \hat{q}_{i1}(1 - a^*) + b^* q_{i2}$, we obtain the maximal $\tilde{q}_{i1}$, which is given by

$$\tilde{q}_{i1} = \hat{q}_{i1}\frac{1 + v^*}{u_* + v^*} + \hat{q}_{i2}\frac{1 + v^*}{1 + v^*/u_*}.$$

Accordingly, we obtain the minimal $\tilde{q}_{i1}$ as

$$\tilde{q}_{i1} = \hat{q}_{i1}\frac{u^* - 1}{u^* + v_*} + \hat{q}_{i2}\frac{1 - u^*}{1 + u^*/v_*}$$

This gives (A.3) and we have shown all claims for the range of $\tilde{q}$.

Step 2: Maximizing and minimizing $\tilde{p}_{kjm}$: Here, we have to find combinations $(a,b) \in \mathbb{L}$ minimizing and maximizing $\tilde{p}_{1jm} = \frac{1-b}{1-a-b}\hat{p}_{1jm} - \frac{a}{1-a-b}\hat{p}_{2jm}$ and $\tilde{p}_{2jm} = \frac{-b}{1-a-b}\hat{p}_{1jm} + \frac{1-a}{1-a-b}\hat{p}_{2jm}$. We start by a reparametrization of S with

$$S^{-1} = \begin{pmatrix} 1-c & c \\ d & 1-d \end{pmatrix}, \qquad S = \frac{1}{1-c-d}\begin{pmatrix} 1-d & -c \\ -d & 1-c \end{pmatrix}.$$

Since we restrict to $a+b \leq 1$, we must assume

$$\frac{-c-d}{1-c-d} \leq 1 \text{ or } 1 - \frac{-c-d}{1-c-d} = \frac{1}{1-c-d} \geq 0,$$

which implies $c+d \leq 1$. So, we impose the restriction $c+d \leq 1$ in the sequel. In other words, we set

$$c = \frac{a}{1-a-b}, \qquad d = \frac{b}{1-a-b},$$

or

$$a = \frac{c}{1+c+d}, \qquad b = \frac{d}{1+c+d}.$$

The conditions (A.5) and (A.6) become

$$\begin{aligned} \hat{q}S &= \frac{1}{1-c-d}\big(\hat{q}_1(1-d) - d\hat{q}_2 = \hat{q}_1 - d, -c\hat{q}_1 + (1-c)\hat{q}_2 = \hat{q}_2 - c\big) \geq 0, \\ \hat{p}(S^{-1})^\top &= \big(\hat{p}_1(1-c) + c\hat{p}_2, d\hat{p}_1 + (1-d)\hat{p}_2\big) \geq 0. \end{aligned}$$

Since $c+d \leq 1$, we have

$$\begin{aligned} &d \leq \hat{q}_1, \\ &c \leq \hat{q}_2, \\ &c \leq \frac{\hat{p}_1}{\hat{p}_1 - \hat{p}_2} = \frac{1}{1-\hat{p}_2/\hat{p}_1} \text{ if } \hat{p}_1 > \hat{p}_2, \\ &c \geq -\frac{1}{\hat{p}_2/\hat{p}_1 - 1} \text{ if } \hat{p}_1 < \hat{p}_2, \\ &d \leq \frac{\hat{p}_2}{\hat{p}_2 - \hat{p}_1} = \frac{1}{1-\hat{p}_1/\hat{p}_2} \text{ if } \hat{p}_2 > \hat{p}_1, \\ &d \geq -\frac{1}{\hat{p}_1/\hat{p}_2 - 1} \text{ if } \hat{p}_2 < \hat{p}_1. \end{aligned}$$

The first three inequalities have to hold for all i, so we can write

$$\begin{aligned} d &\leq \min \hat{q}_{i1} = \frac{v_*}{1+v_*}, \\ c &\leq \min \hat{q}_{i2} = \frac{1}{1+v^*}. \end{aligned}$$

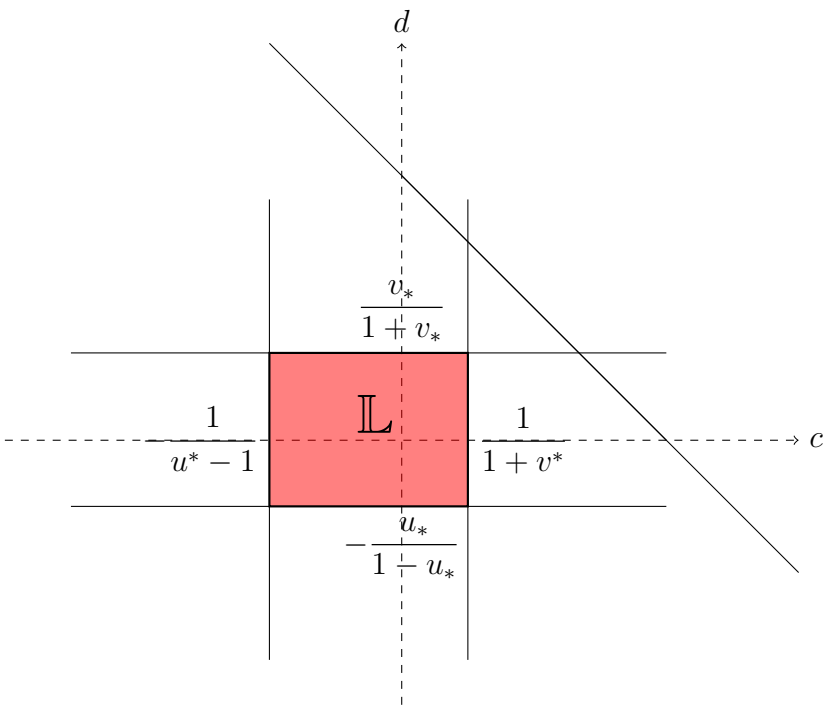

Figure S2: The set of (c, d), such that $\tilde{q} = \hat{q}S(a, b)$ and $\tilde{p} = \hat{p}(S(a, b)^{-1})^\top$ satisfies the conditions (2.1), is given by $\mathbb{L}$.

The latter four inequalities have to hold for all j, m, we can replace them by

$$\begin{aligned} c &\leq \frac{1}{1-u_*}, \\ c &\geq -\frac{1}{u^*-1}, \\ d &\leq \frac{1}{1-1/u^*}, \\ d &\geq -\frac{1}{1/u_*-1}. \end{aligned}$$

An illustration of all points (c, d) satisfying these inequalities is given in FigureS2.

In order to maximize (or minimize) $\tilde{p}_{1jm}$, we have to distinguish two cases. If $\hat{p}_{2jm} > \hat{p}_{1jm}$, we have to find the maximal (minimal) allowed c. If $\hat{p}_{2jm} < \hat{p}_{1jm}$, we must rather find the minimal (maximal) allowed c. In each case, $\tilde{p}_{1jm}$ is between

$$\hat{p}_{1jm}\frac{u^*}{u^*-1} - \frac{1}{u^*-1}\hat{p}_{2jm} \text{ and } -\hat{p}_{1jm}\frac{v^*}{1+v^*} + \frac{1}{1+v^*}\hat{p}_{2jm}.$$

Similarly, $\tilde{p}_{2jm}$ is between

$$-\frac{u_*}{1-u_*}\hat{p}_{1jm} + \frac{1}{1-u_*}\hat{p}_{2jm} \text{ and } \frac{v_*}{1+v_*}\hat{p}_{1jm} + \frac{1}{1+v_*}\hat{p}_{2jm}.$$

This gives (A.4) and we have shown all claims.

□

B A simple example

In order to see the dependence of the allele frequency differential on the range of possible MLEs, we consider the following example: For $J = K = 2, N = M = 1$, we write $\hat{p}_{1,1,1} = \hat{p}_1, \hat{p}_{2,1,1} = \hat{p}_2, \hat{q}_{1,1} = \hat{q}_1, \hat{q}_{1,2} = \hat{q}_2$. Figure S3 shows an overview of all conceivable combinations of $\hat{p}_1$, $\hat{p}_2$, $\hat{q}_1$, and $\hat{q}_2$ with respect to the quantities $\hat{q}_1\hat{p}_1 + \hat{q}_2\hat{p}_2 = 0.5$ and $\hat{p}_1 \leq \hat{p}_2$. This leads to the same likelihood. This visualization reveals that significant disparities between $\hat{p}_1$ and $\hat{p}_2$ correspond to a reduced number of feasible values for $\hat{q}_1$. Consequently, markers with large allele frequencies differences should be preferred, e.g. a marker with allele frequency 1 in one population and 0 in all other populations would be ideal. This is already known (Rosenberg et al., 2003). Nevertheless, Figure S3 underscores the significance of choosing a good marker set to decrease the range of MLEs and emphasizes that the difference between the MLEs might be large.

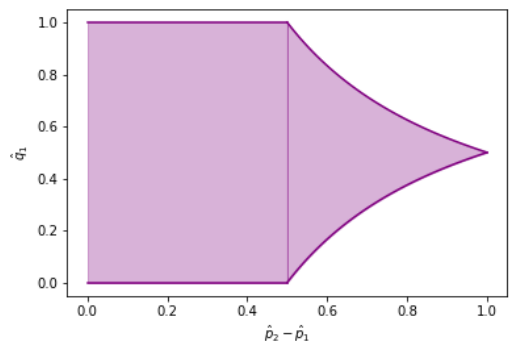

Figure S3: Possible MLEs for $\hat{q}_1$ depending on $\hat{p}_2 - \hat{p}_1$ with $\hat{q}_1\hat{p}_1 + \hat{q}_2\hat{p}_2 = 0.5, \hat{p}_1 \leq \hat{p}_2$.