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

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

April 14, 2025

(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 $(\hat{q}, \hat{p})$. 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.

Keywords: admixture model, STRUCTURE, ADMIXTURE, non-unique estimator, label switching

1 Introduction

Inferring the ancestry of a sample of individuals from genetic data is a formidable challenge, given its importance in various domains, such as exploration of human history (Rosenberg et al., 2002), identification of missing persons (Phillips, 2015), corrections for population stratification (Pritchard and Donnelly, 2001), forensic genetics (Tvedebrink, 2022) or conservation genetics (Wasser et al., 2007). In these applications, individual genomes are assumed to originate from different ancestral populations with distinct allele frequencies. Both, model-based methods and model-free techniques were developed for estimating individual admixture (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., 2023; Falush et al., 2003, 2007; Hubisz et al., 2009; Raj et al., 2014; Ko et al., 2023) treat the number of ancestral populations (usually denoted by K), allele frequencies within populations and individual admixture of all sample individuals as parameters of a statistical model, which we will call the ADMIXTURE MODEL. 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 individual ancestry (see Wollstein and Lao (2015) for an overview) and links between these two approaches have been made (Engelhardt and Stephens, 2010). In recent years, neural networks have become yet another tool for inference of individual admixture, which follow the admixture model (Dominguez Mantes et al., 2023), or similar research questions (Battey et al., 2020).

There are various cases where the output of STRUCTURE and ADMIXTURE must be carefully interpreted. Different historic scenarios might lead to similar genetic data, and researchers have been warned not to over interpret estimates of the admixture model (Lawson et al., 2018). In addition, population structure is not correctly recovered in all cases (Puechmaille, 2016) and it needs to be evaluated whether the assumptions of the admixture model are met, Garcia-Erill and Albrechtsen (2020). Another discussion in the admixture model comes from the choice of K. Since the K populations are ancestral, it is per se unclear what the right choice of K is, given an unstructured sample of genetic material (Evanno et al., 2005; Pritchard et al., 2000). As the admixture model does not account for the potential relationships between ancestral populations, it can result in contradictory predictions when using different values for K (Burger et al., 2024). Frequently, inferring K from the data leads to the choice $K = 2$, see e.g. (Janes et al., 2017; Friedlaender et al., 2008; Rosenberg et al., 2005; Lawson et al., 2018; Hamlin and Arnold, 2014; Phillips et al., 2009), while it is questioned if K can reasonably inferred at all (Verity and Nichols, 2016; Wang, 2017). Due to the increasing number of parameters, STRUCTURE is in practice not applied for many different populations, i.e. $K \geq 10$ (Lawson et al., 2012).

In the present study, we focus on principal limitations of the admixture model from a statistical perspective. Usually, STRUCTURE and ADMIXTURE result in maximum likelihood estimators (or maximum aposteriori estimators) of both, allele frequencies and admixture proportions. In a sample of N individuals, M bi-allelic markers and K ancestral

populations, we have to estimate both, the $N \times K$-matrix of individual admixtures, and the $K \times M$-matrix of allele frequencies.

One of the well-known limitation of the admixture model is usually called label-switching, which means that relabeling ancestral populations leads to equivalent results. This corresponds to applying the same permutation to the population indices in both matrices. Label-switching is of practical importance, since it is one reason why different runs of STRUCTURE or ADMIXTURE 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). As a possible solution, a series of investigations tries to resolve label-switching with high computational effort in order to be able to compare different outputs of STRUCTURE (Behr et al., 2016; Rosenberg, 2004; Jakobsson and Rosenberg, 2007; Earl and VonHoldt, 2012; Kopelman et al., 2015), even for different choices of K (Margaryan et al., 2020; Larena et al., 2021; Jia et al., 2020; Liu et al., 2024).

We show that the non-uniqueness of point estimates substantially extends beyond the phenomenon of mere label-switching and can be more severe. The differences between equally likely estimates are not just nuances, they can be substantial. However, the set of equally likely estimators is small if markers are able to distinguish clearly between populations. Therefore, our results highlight that reliable ancestry estimates depend on at least a few SNPs with large allele frequency differences between the populations. If such SNPs are unavailable, estimates within the admixture model must be taken with caution.

The structure of the paper is as follows: After introducing the admixture model, we start our investigation with a single estimator of individual admixtures and allele frequencies, and look for the range of estimators which are equally likely. From this we can compute the differences between all the identified estimators with respect to the estimated allele frequencies and the estimated individual admixtures. This results in 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 equally likely estimators. To show the consequences of the flatness of the likelihood curve, we apply our theoretical results to genetic data from the 1000 Genomes Project (Consortium et al., 2015). Furthermore, we compare our approach to Pong, a method proposed by Behr et al. (2016), i.e. to running STRUCTURE several times while accounting for label-switching to identify different modes. In contrast to the identification of the different modes discovered by STRUCTURE/ADMIXTURE, here we explore the complete space of equally likely estimators.

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, $m = 1, \ldots, M$, the genetic data of N diploid individuals is given by $x = (x_{ijm})_{(i=1,\ldots,N, j=1,\ldots,J_m, m=1,\ldots,M)}$. 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 individual admixtures $q = (q_{ik})_{i=1,\dots,N,k=1,\dots,K}$ 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 individual ancestry 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)

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

i.e., the individual admixtures q of each individual sum to one and allele frequencies p sum to one at each locus. Given x_{ijm}, the number of alleles j of individual i at marker m, there is a constant C_x which depends only on $x = (x_{ijm})$, such that 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 $(q_{i\cdot} p_{\cdot jm}^\top)_{i=1,\dots,N,j=1,\dots,J_m,m=1,\dots,M} = qp$. Maximum likelihood estimators are then given by

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

2.2 Equally likely estimators

We start with an explanation why there are different estimators in the admixture model with the same likelihood. 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, $\tilde{q}_{i\cdot} \tilde{p}_{\cdot jm}^\top = \hat{q}_{i\cdot} S (\hat{p}_{\cdot jm} S^{-1})^\top = \hat{q}_{i\cdot} S S^{-1} \hat{p}_{\cdot jm}^\top = \hat{q}_{i\cdot} \hat{p}_{\cdot jm}^\top$, 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. $S 1_K^\top = 1_K^\top$. (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

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

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

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

side conditions (i) and (iii) are satisfied for $\tilde{q}, \tilde{p}$ as well. So, we have shown that the likelihood curve is flat in some directions of the parameter space, and hence likelihood-based estimators are 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 individual admixtures 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 individual admixtures 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 the estimators $Sq, S^{-1}p$ and q, p are equally likely. However, label switching is only one example, and the goal here is to identify the complete range of equally likely estimators beyond label switching.

2.3 EMALAM

Our algorithm for exploring the set of equally likely estimators starts with a single estimator $\hat{q}, \hat{p}$, 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 individual ancestry in k for all individuals in $\mathcal{I}$. In (II), we rather minimize (favoring less admixed individuals) or maximize (favoring admixed individuals) the average entropy for individuals in $\mathcal{I}$. See also Figure 4 for an illustration how an equally likely estimator looks like for the four different options in an example data set.

2.4 Non-identifiability and allele frequency differences

Let us consider a toy example for the claimed non-identifiability. We simulate 1000 markers with different allele frequencies among 100 admixed individuals from two populations. The true individual admixtures are depicted as the black line in Figure 1, and the range of equally likely estimators as given by (3.1) are the colored area. Depending on whether we only have

1000 markers with intermediate allele frequency differences, or additionally two markers with a large frequency differential, the range of equally likely estimators varies; see Figure 1. Recall that markers which are (almost) fixed in one, and polymorphic in all other populations, are called anchor markers. Equivalently, individuals with individual ancestry almost exclusively in one population, are called anchor individuals (Cabreros and Storey, 2019). Figure 1 thus also illustrates how the addition of anchor markers and individuals influences the range of equally likely estimators. Without such anchors, the range of equally likely admixture proportions can span a substantial area such that the contribution of the two populations is basically impossible to disentangle (Figure 1 top left). In the case with two populations A and B, if for an equally likely estimator the admixture proportion for population A of a certain individual is increased, so is the admixture proportion for population A in all other individuals, and the allele frequencies change accordingly. However, when $K > 2$ the dependencies will be more complex.

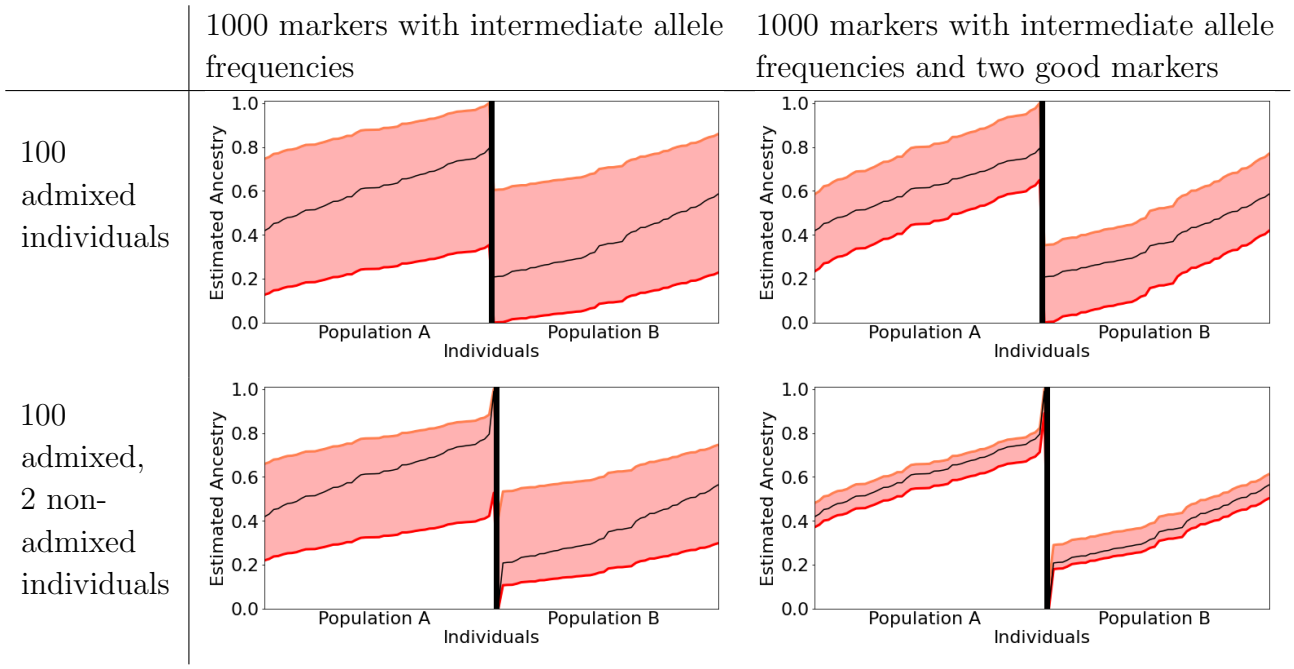

Figure 1: Range of equally likely estimators for two simulated populations. The x-axis represents the individuals and the y-axis indicates the individual ancestry for population A. The black lines show the true (simulated) individual ancestries. The area between the red lines symbolizes equally likely estimators. 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 $p_{1,1,m} + \epsilon$, where ϵ is uniformly distributed in [0, 0.3] if $p_{2,1,m} \in [0.2, 0.8]$ and $p_{2,1,m} = 0.5$ else. Additionally, we added in the right column two additional markers with frequencies $(0.01, 0.1), (0.9, 0.99)$ in populations A and B, respectively. In the second row, we included two non-admixed individuals with ancestry $(0.01, 0.99)$ and $(0.99, 0.01)$.

2.5 Data applications

Since our main focus is on a proof of concept of non-identifiability in the admixture model, we use genetic data from the 1000 genomes project (Consortium et al., 2015). In our applications, we choose every 10.000^{th} bi-allelic SNP among all SNPs with a minor allele frequency of 5%, resulting in about 10^5 markers for intercontinental and about $7 \cdot 10^4$ for intra-continental applications. Moreover, we considered the Kidd AIM set (Kidd et al., 2014) with 55 bi-allelic markers.

3 Results

3.1 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 an estimator $\hat{q}, \hat{p}$, we give for each individual i the range of possible individual admixtures with the same likelihood.

We show in Theorem 1 in Appendix A that equally likely estimators for individual ancestry of individual i in population 1, i.e. q_{i1}, are in

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

with

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

Most notably, the range of possible admixture proportions of individual i among the equally likely estimators 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-identifiability 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 $v_* \ll 1 \ll v^*$. This gives a more quantitative view on the fact that anchor individuals and anchor markers are necessary for uniqueness of the equally likely estimator; see the anchor conditions from Cabreros and Storey (2019).

The case $K > 2$

In this case, we did not obtain an analytical solution for the possible individual admixtures. 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 individual admixtures. 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.

3.2 Non-identifiability 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, and then study the range of equally likely estimators using the STRUCTURE estimator and EMALAM. We use data from northern Europeans from Utah (CEU, 99 individuals), Great Britain (GBR, 91 individuals), Iberian Populations in Spain (IBS, 107 individuals) and Tuscans from Italy (TSI, 107 individuals) in different combinations in the following. For the (small) marker set from Kidd et al. (2014), we make 20 independent runs of STRUCTURE, whereas we make only one run for the large marker set.

First, we consider two ancestral populations, $K = 2$. Using only data from GBR and IBS, Figures 2 and 3 show the range of equally likely individual admixtures (the colored area). In order to put our results into context, we (i) run STRUCTURE 20 times and apply pong (Behr et al., 2016) to the outputs to avoid label switching for the small marker set and (ii) compare the area of equally likely estimators when starting with the outputs of STRUCTURE and ADMIXTURE for the larger marker set. For (i), consider the blue and green lines in Figure 2, which uses the marker set from Kidd et al. (2014). Here, the differences between the 20 runs of STRUCTURE are much smaller than the differences between the equally likely estimators found by EMALAM.

For (ii) consider the same individuals in Figure 3, but using the large marker set of 71185 SNPs. Here we see that the range of estimators with the same likelihood as the output of STRUCTURE is still large, comparable with the range when using the marker set from Kidd et al. (2014) (Figure 2). We also see that the difference between the outputs of STRUCTURE and ADMIXTURE is small. Interestingly, we see that the difference between the maximal and the minimal estimator with the same likelihood as the output of ADMIXTURE is much smaller, i.e. so small that we cannot see any differences in Figure 2.

Second, we consider another combination of populations, namely CEU, TSI and IBS where we apply EMALAM to the data with $K = 3$. In Figure 4, the data was created by running STRUCTURE once and by applying option (I) (a) and (b) (recall from (2.4)) to the blue population and option (II) (a) and (b) to the output of STRUCTURE for $K = 3$. The different options of EMALAM aim to highlight different aspects where the interpretation of STRUCTURE and ADMIXTURE results can be particularly misleading when the complete range of estimators

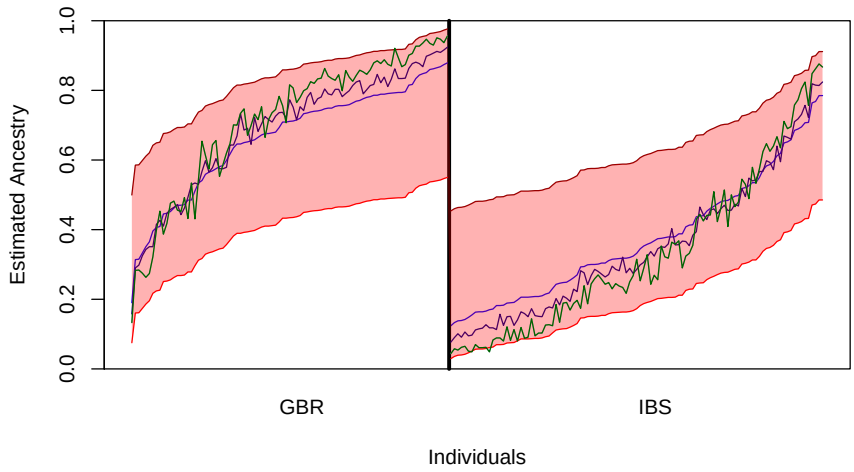

Figure 2: Range of equally likely estimators for GBR and IBS in the admixture model with $K = 2$ with the marker set from Kidd et al. (2014). The colored area is between the minimal and maximal equally likely estimator for the individual ancestry, if we apply the formula (3.1). The three colored lines 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 dark blue line represent minor modes, while the blue line represents the major mode.

with equal likelihood is not considered. Option (I) maximizes and minimizes the ancestry proportions of a particular population, which can be useful to assess the confidence with which an individual is attributed to a certain bio-geographic background. Option (II) can highlight a different aspect as it maximizes or minimizes the entropy of admixtures in all individuals. This option thus helps to assess the identifiability of separate populations and whether the individuals can be separated into multiple populations. This option helps determine if the estimators clearly suggest that individuals can be categorized into distinct populations or whether individuals are more admixed.

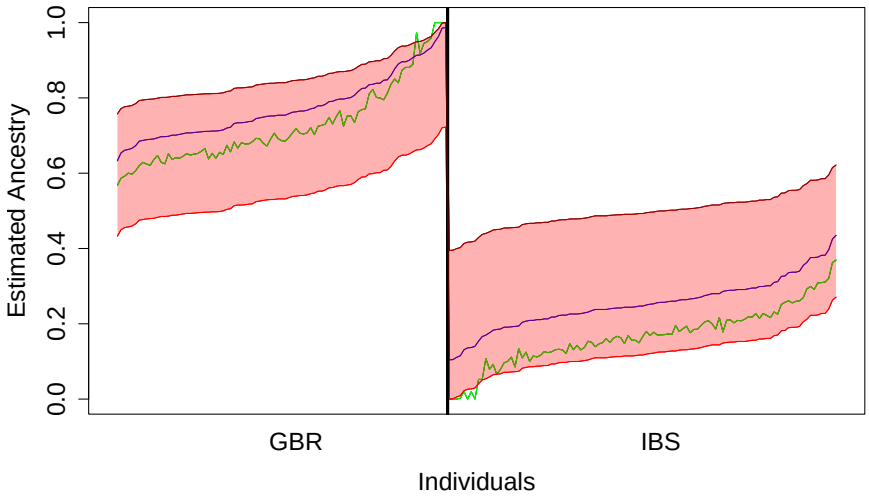

Figure 3: Range of equally likely estimators for GBR and IBS populations in the admixture model with $K = 2$. Ancestry proportions of the estimators are depicted as in Figure 2 but here we use 71185 SNPs. We only consider bi-allelic SNPs, where the allele frequency of both alleles is higher than 0.05. Additionally, we used only every 10000th SNP. The green line corresponds to the estimated ancestry of ADMIXTURE. The blue line is the output of STRUCTURE.

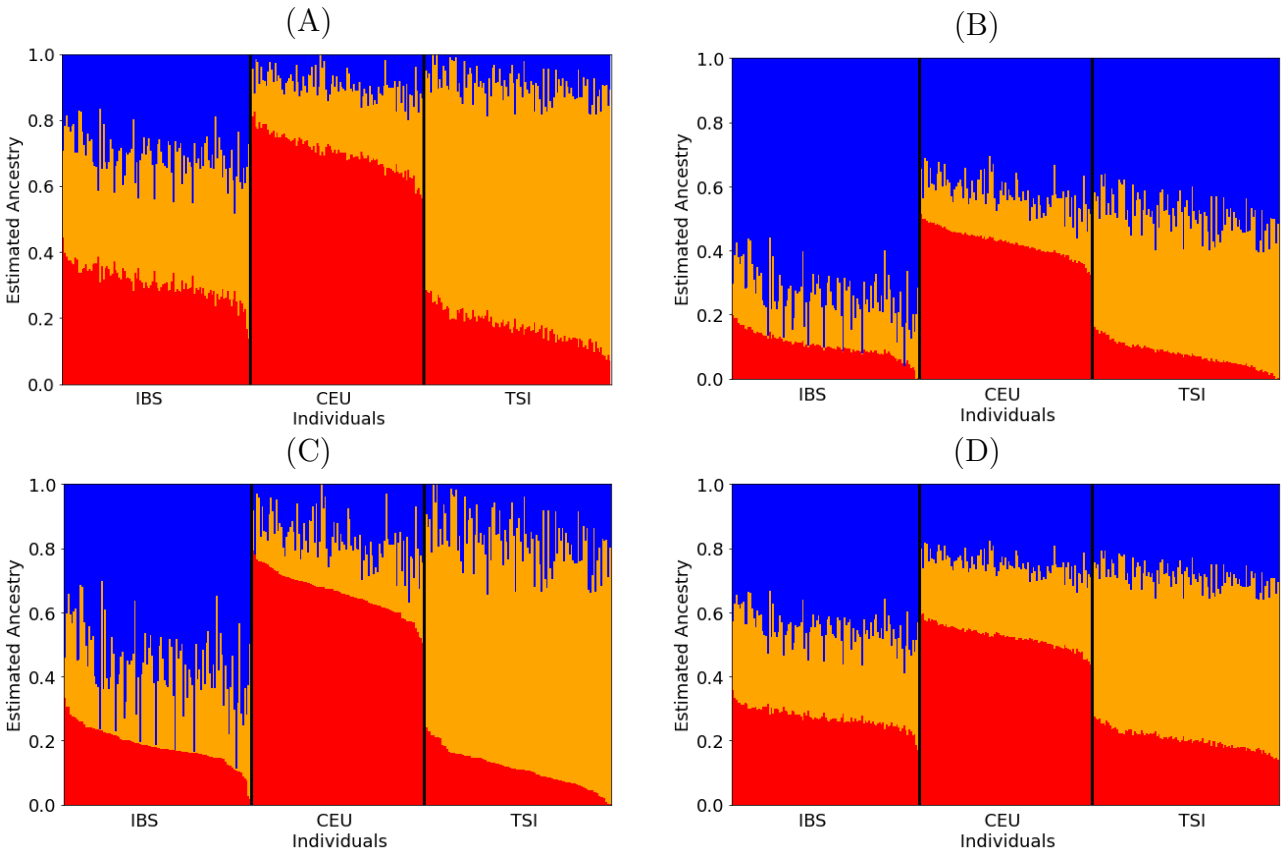

Figure 4: Estimated individual admixtures for the TSI, IBS, and CEU populations with 125350 markers. All four figures show estimators with the same likelihood. We used EMALAM with $K = 3$ in order to optimize the following target in the four figures: (A) Minimize the ancestry from the blue population (option (I)(a)); (B) Maximize the ancestry from the blue population (option (I)(b)); (C) Minimize the entropy (favoring non-admixed individuals), i.e. option (II)(a); (D) Maximize the entropy (favoring admixed individuals), i.e. option (II)(b). Orders of individuals on the x-axis is the same for (A)–(D).

4 Discussion

We demonstrated how the non-uniqueness of the estimators in the admixture model impacts the range of equally likely estimators.

Here, we introduced EMALAM, a method to explore the complete range of equally likely estimators. In addition, EMALAM is able to merge estimators that occur due to label switching; cf. pong (Behr et al., 2016). We stress that in general, the non-uniqueness of equally likely estimators 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 estimators with respect to specific aspects, i.e. the individual admixtures 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 identifiability of the admixture model. If such anchors are missing, the range of estimators which EMALAM computes can be too large for a useful interpretation of the results provided by STRUCTURE or ADMIXTURE concerning the inference of the population structure, such that for some MLEs the ancestry fractions of one populations can vanish. 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 equally likely estimators can be small (see Figure 1).

EMALAM helps to investigate the range of equally likely estimators and how meaningful the inferred ancestries are. As a rule of thumb, small differences between the estimated individual ancestries and the estimated allele frequencies lead to a large range of possible estimated individual admixtures 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 individual ancestry 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 individual ancestry 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 equally likely estimators in the admixture model can be larger than considering the different modes of multiple STRUCTURE runs suggests.

One caveat of applying EMALAM is that it only identifies equally likely estimators starting from a given initial estimator. Estimators with similar or even lower likelihoods are not detected, and the resulting range of equally likely estimators depends thus on the starting point. For example, if an estimator inferred by ADMIXTURE creates a spurious anchor individual or allele, the inferred range may appear narrow, which could potentially give a false impression of high

confidence. While this is an extreme case, it is advisable to run STRUCTURE/ADMIXTURE multiple times to capture the variability in inferred modes, and apply EMALAM to each mode to better assess the range of equally likely estimators for each of them.

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 et al., 2018; Burger et al., 2024) can circumvent the non-identifiability of the admixture model. However, while more complex models can lead to more precise individual admixtures 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 equally likely estimators 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.

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.

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 (2018). Inferring continuous and discrete population genetic structure across space. Genetics 210(1), 33–52.

Burger, K. E., S. Klepper, U. Von Luxburg, and F. Baumdicker (2024, December). Inferring ancestry with the hierarchical soft clustering approach tangleGen. Genome Research 34(12), 2244–2255.

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.

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.

Engelhardt, B. E. and M. Stephens (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS genetics 6(9), e1001117.

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.

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.

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

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.

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

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.

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.

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.

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 individual admixtures 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:

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

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

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

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)

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

Since all conditions are linear, they are satisfied 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.

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

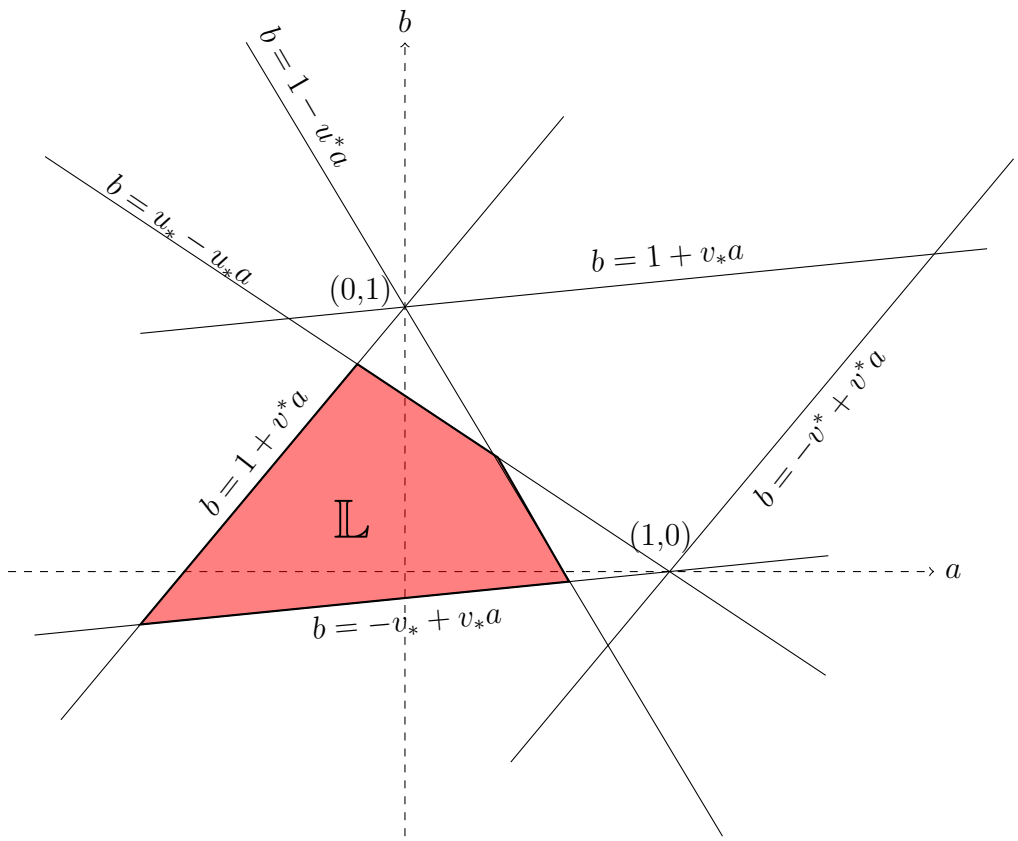

Figure S1: The set of (a, b), such that $\tilde{q} = \hat{q}S(a, b)$ and $\tilde{p} = \hat{p}(S(a, b)^{-1})^\top$ satisfies the conditions (2.1), is given by $\mathbb{L}$.

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

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

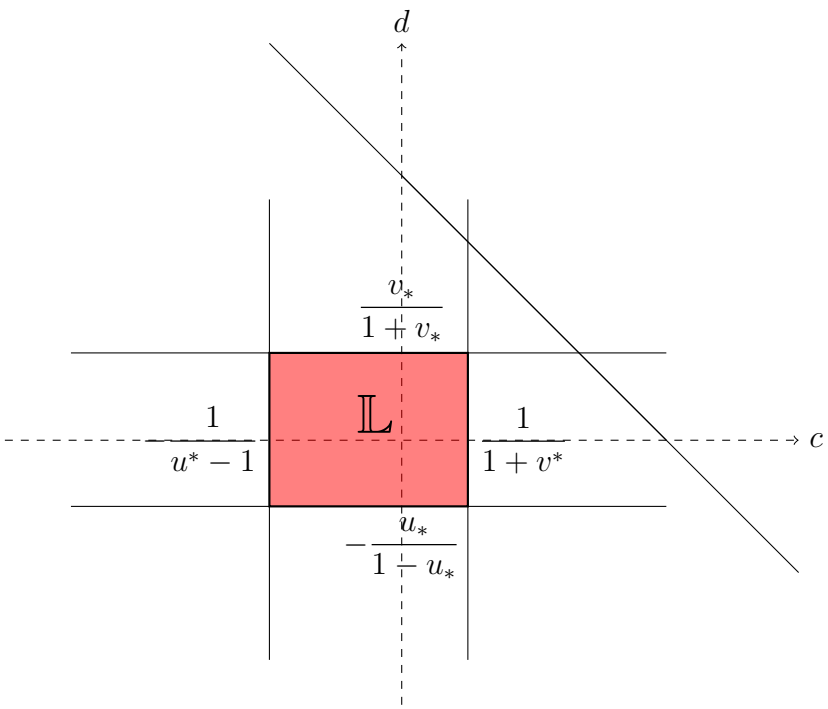

Figure S2: The set of (c, d), such that $\tilde{q} = \hat{q}S(a, b)$ and $\tilde{p} = \hat{p}(S(a, b)^{-1})^\top$ satisfies the conditions (2.1), is given by $\mathbb{L}$.

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.

□