A Dirichlet model of alignment cost in mixed-membership unsupervised clustering

Xiran Liu
Institute for Computational and Mathematical Engineering, Stanford University
and
Naama Kopelman
Faculty of Sciences, Holon Institute of Technology
and
Noah A. Rosenberg
Department of Biology, Stanford University

September 14, 2022

Abstract

Mixed-membership unsupervised clustering is widely used to extract informative patterns from data in many application areas. For a shared data set, the stochasticity and unsupervised nature of clustering algorithms can cause difficulties in comparing clustering results produced by different algorithms, or even multiple runs of the same algorithm, as outcomes can differ owing to permutation of the cluster labels or genuine differences in clustering results. Here, with a focus on inference of individual genetic ancestry in population-genetic studies, we study the cost of misalignment of mixed-membership unsupervised clustering replicates under a theoretical model of cluster memberships. Using Dirichlet distributions to model membership coefficient vectors, we provide theoretical results quantifying the alignment cost as a function of the Dirichlet parameters and the Hamming permutation difference between replicates. For fixed Dirichlet parameters, the alignment cost is seen to increase with the Hamming distance between permutations. Data sets with low variance across individuals of membership coefficients for specific clusters generally produce high misalignment costs—so that a single optimal permutation has far lower cost than suboptimal permutations. Higher variability in data, as represented by greater variance of membership coefficients, generally results in alignment costs that are similar between the optimal permutation and suboptimal permutations. We demonstrate the application of the theoretical results to data simulated under the Dirichlet model, as well as to membership estimates from inference of human-genetic ancestry. The results can contribute to improving cluster alignment algorithms that seek to find optimal permutations of replicates.

Keywords: admixture, Dirichlet model, label-switching, multimodality

1 Introduction

In mixed-membership unsupervised clustering, statistical models of a set of clusters and a set of entities are considered, so that the total "membership" of an entity is distributed across the clusters (Airoldi et al., 2015a). Using the models, patterns inferred among the entities are interpreted by examining their co-clustering, and a cluster itself is interpreted by examining the entities that possess large membership fractions for the cluster.

Mixed-membership unsupervised clustering has found diverse applications in such areas as document and text classification, statistics of networks, and medical diagnostics (Airoldi et al., 2015b). In one of the most prominent areas of application—the field of population genetics—it has long been a central technique for recovering information about genetic relationships of individuals and populations. In typical population-genetic studies, researchers collect genotypes from individuals within a species, measure features of genetic variation among the individuals, and infer evolutionary processes that have generated those features. Mixed-membership unsupervised clustering techniques designed specifically for population-genetic data—Structure (Pritchard et al., 2000), Admixture (Alexander et al., 2009), and Baps (Corander et al., 2003), for example—use stochastic iterative clustering algorithms to infer membership fractions for individuals in clusters. The membership fraction for an individual in a cluster is interpreted, depending on the setting, in one of two ways. In some settings, it represents the proportion of the individual's genome originating in the cluster, or the probability that within the individual, an observation of a specific site in the genome originates from that cluster, with different sites having independent and identical probabilities. In others, it gives the probability that the individual's *entire* genome originates from the cluster, so that different sites are identically distributed but fully dependent.

In unsupervised clustering, two challenges to data analysis have long been recognized: label-switching and genuine multimodality (Stephens, 2000; Jasra et al., 2005; Jakobsson and Rosenberg, 2007; Airoldi et al., 2015a). *Label-switching* describes the fact that because the methods include stochastic steps, if K clusters are labeled $1, 2, \ldots, K$, then $K!$ distinct permutations of the cluster labels have equivalent meaning. For example, Figure 1A shows two permutations of the clusters for a single set of cluster memberships; the panels differ

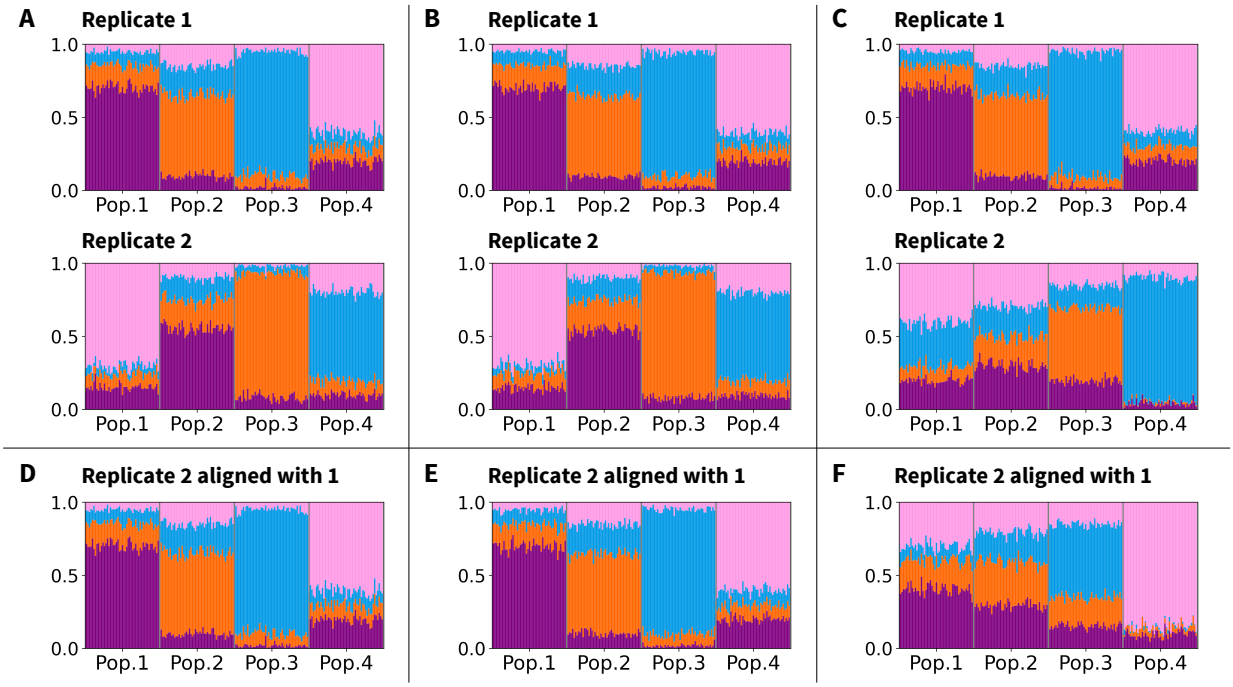

Figure 1: Label-switching and genuine multimodality. (A-C) Pairs of replicates. (D-F) Optimal permutations of replicate 2 to align with replicate 1. Replicates are simulated using the Dirichlet parameter values in Appendix A. (A) Label-switching only. (B) Label-switching with two independent replicates simulated from the same parameter values. (C) Genuine multimodality. (D) Optimal permutation for (A). (E) Optimal permutation for (B). (F) Optimal permutation for (C). The number of individuals per population is 50.

only in that the cluster labels, each of which is represented by a color, differ between the permutations. When label-switching is present, clustering replicates can be aligned by identifying the unique permutation that makes the replicates equivalent (Figure 1D). *Genuine multimodality* arises if no permutation exists that makes replicates equivalent, such as in Figure 1C. Replicates using the same data can fail to produce identical memberships, even after they are permuted to align in many features (Figure 1F). A common scenario is that in which replicates are not strictly equivalent (Figure 1B), but a permutation makes them extremely similar (Figure 1E); this situation is informally described as possessing label-switching rather than genuine multimodality.

To make use of cluster analyses from multiple data sets, algorithms, or settings, a method is needed for identifying the permutations that eliminate label-switching and reveal genuine multimodality. In the population-genetic context, early attempts at permutation proceeded informally, as the particular features of data sets often rendered the optimal permutations relatively easy to identify (e.g. Rosenberg et al., 2001; Rosenberg, 2004). To advance on this situation, several algorithms, including CLUMPP (Jakobsson and Rosenberg, 2007), CLUMPAK (Kopelman et al., 2015), and PONG (Behr et al., 2016), have been introduced for identifying optimal alignments, where an optimal alignment is one that minimizes a cost function or maximizes a similarity function. These algorithms are now widely used with unsupervised clustering methods to clarify the results that the methods produce.

The alignment algorithms are generally seen to perform well in identifying permutations that visually align replicates (Jakobsson and Rosenberg, 2007; Kopelman et al., 2015; Behr et al., 2016). However, despite the widespread use of alignment algorithms in population genetics, formal evaluations of their success at identifying optimal permutations have not been performed. Further, relatively little understanding has been available regarding the alignment cost difference of suboptimal permutations in relation to minimal-cost permutations; thus, when permutations are suboptimal, the potential for reducing the alignment cost from that achieved by existing algorithms remains unclear.

In this study, we introduce a model for evaluating the cost difference of optimal and suboptimal permutations. We treat individual memberships as drawn from a Dirichlet

distribution with specified parameters. Under the Dirichlet model, we explore the cost of suboptimal permutations as a function of the number of misaligned clusters—the Hamming distance between permutations. We find that cost generally increases with the number of misaligned clusters. For examples in which Dirichlet parameters assign each individual primarily to a single cluster, the alignment cost for suboptimal permutations is generally substantially higher than for the optimum. For "noisy" data, as represented by Dirichlet parameters with similar mean values of membership components for different clusters, suboptimal permutations can possess cost similar to the optimum. The model can help in understanding challenges for algorithms that seek to produce minimal-cost permutations.

2 Model

2.1 Terminology

Model-based unsupervised clustering algorithms in population genetics produce a vector of membership coefficients for each individual. For N individuals and K clusters, the output of a clustering algorithm is an estimated $N \times K$ membership coefficient matrix $\hat{Q}$, where $\hat{q}_{ik}$ is the estimated coefficient for individual i in cluster k. In models in which each individual is treated as belonging to a single cluster, $\hat{q}_{ik}$ represents the estimated probability that individual i is a member of cluster k. In mixed-membership models, in which an individual possesses membership in multiple clusters, $\hat{q}_{ik}$ is the estimated fraction of the data from individual i that originates from cluster k.

For convenience, we use the language of mixed-membership unsupervised clustering in population genetics, but our analysis can also apply to cases of population-genetic clustering in which membership coefficients are interpreted as probabilities rather than ancestry fractions, as well as to related applications outside population genetics. Note that in the population-genetic context, we distinguish "populations," representing predetermined groups of individuals, from "clusters," the K groups for which membership is estimated.

Hence, each individual has an estimated membership vector $\hat{\mathbf{q}}_i = (\hat{q}_{i1}, \hat{q}_{i2}, \ldots, \hat{q}_{iK})$, for which the sum across clusters is $\sum_{k=1}^{K} \hat{q}_{ik} = 1$. The estimated membership matrix $\hat{Q}$ is

a right-stochastic matrix, and each column vector characterizes a cluster by the list of associated memberships of the N individuals.

In population genetics, unsupervised cluster analyses study the patterns of genetic variation of individuals from multiple populations. They infer a matrix that contains the estimated membership proportions of the individuals in the K clusters, where clusters correspond either to supervised ancestry groups or emergent groups appearing in specific data analyses. For instance, in a supervised analysis with three clusters 1, 2, and 3, representing ancestry in three distant populations, respectively, an individual with estimated membership vector $(0.6, 0.3, 0.1)$ has 60% of its genome estimated to originate from population 1, 30% from population 2, and 10% from population 3.

2.2 Dirichlet model

We consider membership coefficients drawn from a theoretical model. Each of a set of predetermined populations is assumed to have its own characteristic distribution of membership coefficients for a series of K clusters. For an analysis with K clusters, a natural choice to model the individual membership coefficients of a population is a Dirichlet distribution of order K (Kotz et al., 2004, chapter 49).

In this model, for a given predefined population, the expected membership proportion of an individual in cluster k is the mean of the kth random variable in a Dirichlet-distributed random vector. Suppose a random vector $\mathbf{q}$ is drawn from the Dirichlet distribution of order K with parameters $\mathbf{a} = (a_1, a_2 \ldots, a_K)$, where $a_k > 0$ for all k. Writing $a_0 = \sum_{k=1}^{K} a_k$, this $\mathbf{q} \sim \mathrm{Dir}(\mathbf{a})$ has probability density function

$$f(\mathbf{q}; \mathbf{a}) = \frac{\Gamma(a_0)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} q_k^{a_k - 1}, \tag{1}$$

where $\Gamma(\cdot)$ is the gamma function (Kotz et al., 2004).

It is convenient to convert Dirichlet parameters $\mathbf{a} = (a_1, a_2, \ldots, a_K)$ to expected memberships, $(\mathbb{E}[q_1], \mathbb{E}[q_2], \ldots, \mathbb{E}[q_K])$, as the expected memberships are often more easily understood than the Dirichlet parameters. We model a set of populations using Dirich-

let distributions, with the Dirichlet parameter vector $\mathbf{a}$ chosen so that each mean value $\mathbb{E}[q_k] = a_k/a_0$ corresponds to the assumed mean proportion of cluster k in a population; the sum a_0 controls the variance. Memberships of different individuals from the same population are modeled as independent random vectors sampled from the Dirichlet model with a shared set of parameter values. Memberships of individuals from different populations follow distributions with different Dirichlet parameters.

We use superscript $\cdot^{(\ell)}$ on the Dirichlet parameters $\mathbf{a}$ to distinguish the parameter vector for population ℓ. Thus, for example, for a set I_1 of individuals from population 1 and a set I_2 of individuals from population 2, $\mathbf{q}_i \sim \mathrm{Dir}(\mathbf{a}^{(1)})$ for all $i \in I_1$ and $\mathbf{q}_i \sim \mathrm{Dir}(\mathbf{a}^{(2)})$ for all $i \in I_2$, where $\mathbf{a}^{(1)}/a_0^{(1)}$ denotes the population-wise parametric mean membership proportions for population 1 and $\mathbf{a}^{(2)}/a_0^{(2)}$ denotes those proportions for population 2.

2.3 Distance function

To quantify the alignment cost for a pair of replicate analyses, we will need a dissimilarity measure for pairs of membership coefficients on the same samples. Consider two replicate $N \times K$ membership coefficient matrices, P and Q. We follow Jakobsson and Rosenberg (2007) in relying on the Frobenius norm of their difference (see also Rosenberg et al. (2002)).

In particular, using $\|\cdot\|_F$ to denote the Frobenius norm, the distance between the membership matrices P and Q can be calculated as

$$D_{1,2} = \sum_{i=1}^{N}\sum_{k=1}^{K}(p_{ik} - q_{ik})^2 = \|P - Q\|_F \,. \tag{2}$$

Each sum $\sum_{k=1}^{K}(p_{ik} - q_{ik})^2$ lies in $[0, 2]$; considering the unit vectors $\mathbf{p}$ and $\mathbf{q}$, we have $\|\mathbf{p} - \mathbf{q}\|_2 \leq \|\mathbf{p}\|_2 + \|\mathbf{q}\|_2 = 2$ by Minkowski's inequality. The choice of the Frobenius norm to measure the difference between membership matrices not only accords with past studies, it is also mathematically convenient in our framework, as the sum of squares that it entails facilitates the computation of integrals with respect to the Dirichlet distribution.

3 Alignment cost for a single individual

3.1 Overview

With our Dirichlet model and distance function established, we now describe the computation of the alignment cost associated with a pair of replicate clusterings and a single individual. Under the model, the membership coefficient vector of an individual is a random vector drawn from the Dirichlet distribution with specified parameters.

Consider two replicate draws from the Dirichlet model, representing outcomes of two cluster analyses. Both replicates have K clusters. However, owing to label-switching, multimodality, or both, the clusters are not necessarily aligned. In the most general case, in one replicate, an individual has membership vector $\mathbf{p}$ drawn from a Dirichlet distribution $\mathrm{Dir}(\mathbf{a})$, and the same individual has membership vector $\mathbf{q}$ drawn from $\mathrm{Dir}(\mathbf{b})$ in the second replicate. Both $\mathbf{p}$ and $\mathbf{q}$ are random vectors; when two replicates are aligned, $\mathbf{p}$ and $\mathbf{q}$ are drawn from the same distribution, and when the replicates are not aligned, the random vectors are drawn from different distributions.

The contribution of the individual to the distance between replicates 1 and 2 is the random variable $\sum_{i=1}^{K}(p_i-q_i)^2 = \|\mathbf{p}-\mathbf{q}\|_2^2$. Denote the mean value of this random variable $\|\mathbf{p}-\mathbf{q}\|_2^2$ by $A_{\mathbf{a},\mathbf{b}}$, where $\mathbf{p}\sim\mathrm{Dir}(\mathbf{a})$ and $\mathbf{q}\sim\mathrm{Dir}(\mathbf{b})$. Using the probability density function of the Dirichlet distribution in Eq. 1, this value can be computed as

$$
\begin{aligned}
A_{\mathbf{a},\mathbf{b}} \;=\; & \int_{p_1=0}^{1}\int_{p_2=0}^{1-p_1}\cdots\int_{p_{K-1}=0}^{1-\sum_{i=1}^{K-2}p_i}\int_{q_1=0}^{1}\int_{q_2=0}^{1-q_1}\cdots\int_{q_{K-1}=0}^{1-\sum_{i=1}^{K-2}q_i}\Bigg[\sum_{i=1}^{K-1}(p_i-q_i)^2 \\
& +\left((1-\sum_{i=1}^{K-1}p_i)-(1-\sum_{i=1}^{K-1}q_i)\right)^2\Bigg]\frac{(\prod_{i=1}^{K-1}p_i^{a_i-1})(1-\sum_{i=1}^{K-1}p_i)^{a_K-1}}{[\prod_{i=1}^{K}\Gamma(a_i)]/\Gamma(\sum_{i=1}^{K}a_i)} \qquad (3)\\
& \times\frac{(\prod_{i=1}^{K-1}q_i^{b_i-1})(1-\sum_{i=1}^{K-1}q_i)^{b_K-1}}{[\prod_{i=1}^{K}\Gamma(b_i)]/\Gamma(\sum_{i=1}^{K}b_i)}\,dq_{K-1}\cdots dq_2\,dq_1\,dp_{K-1}\cdots dp_2\,dp_1.
\end{aligned}
$$

Here, we have made use of the fact that $p_K = 1-\sum_{i=1}^{K-1}p_i$ and $q_K = 1-\sum_{i=1}^{K-1}q_i$. It is often convenient to consider the special case for label-switching, in which the entries of $\mathbf{b}$ represent a permutation of the entries of $\mathbf{a}$. In other words, denote the permutation between two replicates by ϕ, so that cluster $\phi(i)$ gives the number of the cluster in replicate

2 that corresponds to cluster i in replicate 1.

For the label-switching case, in replicate 1, the parameters associated with the K cluster memberships are $(a_1, a_2, \ldots, a_K)$. In replicate 2, corresponding parameters are $(b_1, b_2, \ldots, b_K) = (a_{\phi(1)}, a_{\phi(2)}, \ldots, a_{\phi(K)})$. For simplicity, we let $b_i = a_{\phi(i)}$. Only for the identity permutation ϕ_0, for which $\phi(i) = i$ for all $i = 1, 2, \ldots, K$, are the two replicates aligned. Because the a_i and b_i are the same set of items, permuted, $a_0 = \sum_{i=1}^{K} a_i = \sum_{i=1}^{K} b_i = b_0$.

We are interested in the mean contribution of an individual to the alignment cost, $C_{\mathbf{a},\mathbf{b}}$, which can be calculated as the difference between the contribution of a random individual to the distance between a pair of replicates aligned by permutation $\mathbf{b}$ (misaligned for $\mathbf{b} \neq \mathbf{a}$) and the contribution to the distance between correctly aligned replicates:

$$C_{\mathbf{a},\mathbf{b}} = \frac{A_{\mathbf{a},\mathbf{b}} - A_{\mathbf{a},\mathbf{a}}}{2}. \tag{4}$$

Here, we include a factor of $\frac{1}{2}$ to account for the fact that the Frobenius norm in Eq. 2 has maximum 2, so that $A_{\mathbf{a},\mathbf{b}} - A_{\mathbf{a},\mathbf{a}}$ has a maximum of 2; the cost $C_{\mathbf{a},\mathbf{b}}$ ranges from 0 to 1.

We now evaluate Eq. 3 to obtain the individual mean contribution to distance between replicates. First, we consider $K = 2$. We next examine $K = 3$, and we generalize to arbitrary K. We explore the effect of the Dirichlet parameters on these mean contributions.

3.2 $K = 2$

Consider two replicates, with Dirichlet parameters (a_1, a_2) and $(b_1, b_2) = (a_{\phi(1)}, a_{\phi(2)})$.

Theorem 3.1. *Consider a population of individuals with membership coefficients in $K = 2$ clusters. Suppose that in one replicate, the membership coefficients of the individuals follow a Dirichlet model with parameters $\mathbf{a} = (a_1, a_2)$, and in a second replicate, they follow a Dirichlet model with parameters $\mathbf{b} = (b_1, b_2)$. The mean contribution of a randomly chosen individual to the distance between replicates is*

$$A_{\mathbf{a},\mathbf{b}} = 2\left[\frac{(a_1+1)a_1}{(a_1+a_2+1)(a_1+a_2)} + \frac{(b_1+1)b_1}{(b_1+b_2+1)(b_1+b_2)} - \frac{2a_1b_1}{(a_1+a_2)(b_1+b_2)}\right]. \tag{5}$$

Proof of Theorem 3.1. When $K = 2$, Eq. 3, representing the mean contribution of a ran-

domly chosen individual to the distance between two replicates, becomes

$$\begin{aligned} A_{\mathbf{a},\mathbf{b}} &= \int_{p_1=0}^{1}\int_{q_1=0}^{1}\left((p_1-q_1)^2+\left[(1-p_1)-(1-q_1)\right]^2\right) \\ &\quad\times\frac{p_1^{a_1-1}(1-p_1)^{a_2-1}}{\Gamma(a_1)\,\Gamma(a_2)/\Gamma(a_1+a_2)}\frac{q_1^{b_1-1}(1-q_1)^{b_2-1}}{\Gamma(b_1)\,\Gamma(b_2)/\Gamma(b_1+b_2)}\,dq_1\,dp_1. \end{aligned}$$

We compute this integral in Appendix B to obtain the result. □

We can then apply Theorem 3.1 to obtain the contribution of an individual in the special case that the two replicates are aligned. In other words, we calculate $A_{\mathbf{a},\mathbf{a}}$:

$$A_{\mathbf{a},\mathbf{a}} = \frac{4a_1a_2}{(a_1+a_2)^2(a_1+a_2+1)}. \tag{6}$$

For misaligned replicates that differ by label-switching, consider a permutation ϕ with $(b_1, b_2) = (a_2, a_1)$. We have

$$A_{\mathbf{a},\phi(\mathbf{a})} = 2\frac{a_1^3+a_2^3+a_1^2+a_2^2-a_1^2a_2-a_1a_2^2}{(a_1+a_2)^2(a_1+a_2+1)}. \tag{7}$$

Applying Eqs. 7, 6, and Eq. 4, the mean contribution of an individual to alignment cost is

$$C_{\mathbf{a},\phi(\mathbf{a})} = \frac{(a_1-a_2)^2}{(a_1+a_2)^2}. \tag{8}$$

Figure 2 plots Eq. 8 as a function of a_1 and a_2. For $a_2 = a_1$, the two replicates have the same parameters, and the mean contribution of an individual to the cost is 0. In each replicate, the mean membership coefficient for cluster 1 is $\frac{1}{2}$, and the mean for cluster 2 is also $\frac{1}{2}$. Starting from the $a_2 = a_1$ line in the a_1a_2-plane, as a_2 increases while holding a_1 constant, or as a_1 increases while holding a_2 constant, the cost increases. These parameter changes make the permuted replicate with parameters (a_2, a_1) quite different from the unpermuted replicate with parameters (a_1, a_2), so that a replicate is increasingly distinguishable from its label-switching permutation. The cost approaches 1 for a replicate with high a_1 and low a_2, or vice versa; in these cases, nearly all of the membership lies in one of the two clusters, so that the alignment cost of switching the two clusters is nearly 1.

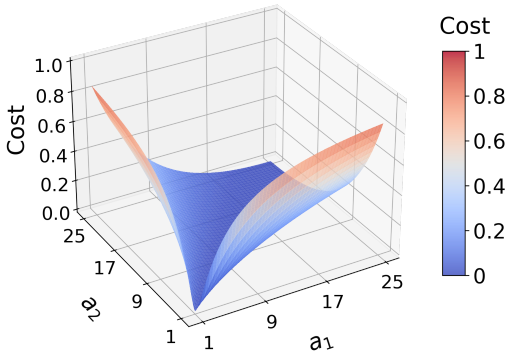

Figure 2: Alignment cost (Eq. 8) for permutation $\phi = (2, 1)$ as a function of the Dirichlet parameters a_1 and a_2 for a model with $K = 2$ clusters under label-switching. Parameters are varied in $[1, 25]$.

3.3 $K = 3$

Considering two replicates, the number of permutations possible for the clusters is $K! = 6$. As in the $K = 2$ case, for each permutation ϕ, we can obtain the mean individual contribution to the distance between two replicates. We compute $A_{\mathbf{a},\mathbf{b}}$ from Eq. 3.

$$\begin{aligned} A_{\mathbf{a},\phi} \quad = \quad & \int_{p_1=0}^{1} \int_{p_2=0}^{1-p_1} \int_{q_1=0}^{1} \int_{q_2=0}^{1-q_1} \Big((p_1 - q_1)^2 + (p_2 - q_2)^2 + \big[(1 - p_1 - p_2) \\ & - (1 - q_1 - q_2) \big]^2 \Big) \frac{p_1^{a_1-1} p_2^{a_2-1} (1 - p_1 - p_2)^{a_3-1}}{\Gamma(a_1)\,\Gamma(a_2)\,\Gamma(a_3)/\Gamma(a_1 + a_2 + a_3)} \\ & \times \frac{q_1^{b_1-1} q_2^{b_2-1} (1 - q_1 - q_2)^{b_3-1}}{\Gamma(b_1)\,\Gamma(b_2)\,\Gamma(b_3)/\Gamma(b_1 + b_2 + b_3)} \, dq_2 \, dq_1 \, dp_2 \, dp_1. \end{aligned} \tag{9}$$

For each ϕ, we compute the alignment cost for ϕ from Eq. 4.

The costs for the 6 permutations appear in Table 1. We omit their derivations, as each can be obtained from the general result that we present for arbitrary K (Section 3.4). In the table, the cost of 0 for the identity permutation appears in the first row. The next three rows show the costs associated with each of the permutations with Hamming distance 2 from the initial permutation $(1, 2, 3)$, where the Hamming distance tabulates the number of clusters that are misaligned between a pair of replicates. The last two rows show the costs for the two permutations with Hamming distance 3 from $(1, 2, 3)$.

Table 1: Alignment cost for each of the six label-switching cases with $K = 3$, as functions of the Dirichlet parameters a_1, a_2, a_3. Each cost is obtained by evaluating Eq. 9, and then applying Eq. 4.

Permutation number	Permutation $\phi(1,2,3)$	Alignment cost $C_{\mathbf{a},\phi(\mathbf{a})}$
1	$(1,2,3)$	0
2	$(1,3,2)$	$(a_2 - a_3)^2/(a_1 + a_2 + a_3)^2$
3	$(3,2,1)$	$(a_1 - a_3)^2/(a_1 + a_2 + a_3)^2$
4	$(2,1,3)$	$(a_1 - a_2)^2/(a_1 + a_2 + a_3)^2$
5	$(2,3,1)$	$(a_1^2 + a_2^2 + a_3^2 - a_1a_2 - a_1a_3 - a_2a_3)/(a_1 + a_2 + a_3)^2$
6	$(3,1,2)$	$(a_1^2 + a_2^2 + a_3^2 - a_1a_2 - a_1a_3 - a_2a_3)/(a_1 + a_2 + a_3)^2$

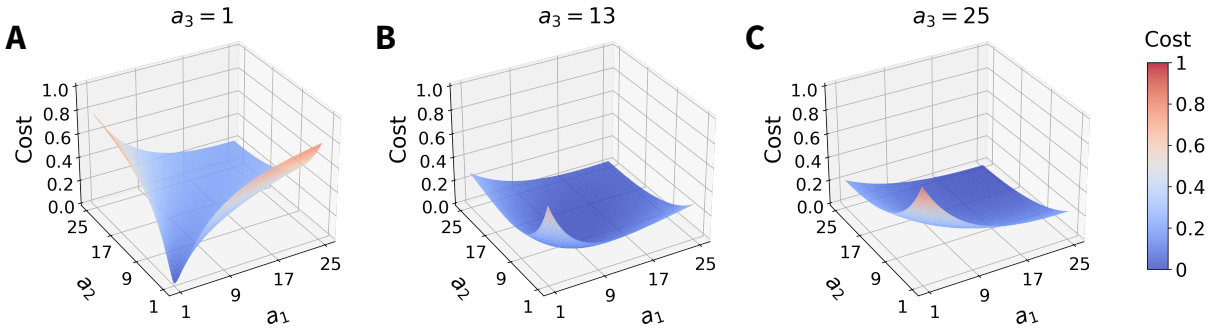

Figure 3: Alignment cost (Table 1) for permutation $\phi = (2,3,1)$ as a function of Dirichlet parameters a_1 and a_2 for $K = 3$ clusters and fixed a_3. (A) $a_3 = 1$. (B) $a_3 = 13$. (C) $a_3 = 25$. Parameters are varied in $[1, 25]$.

Figure 3 plots the alignment cost from Table 1 for label-switching with a specific permutation $(2, 3, 1)$ as a function of a_1 and a_2, for three fixed values of a_3. In each panel, varying the three parameters in $[1, 25]$, for $a_1 = a_2 = a_3$, the alignment cost has the minimum value of 0. In Figure 3A, the largest alignment cost is reached when one of the three parameters has value 25 and the other two are equal to 1; in Figure 3C, the maximum occurs when two parameters equal 1 and the third is 25. In the intermediate Figure 3B, large values occur both in the case that two values equal 1 and the third is 13, and in the case that one value is 1, one is 25, and the third is 13. These maxima reflect the intuition that when memberships differ substantially across clusters, the identity permutation has substantially lower cost than do permutations that represent misalignments.

3.4 Arbitrary K

We now generalize the calculation of the individual mean contribution to distance between replicates (Eq. 3) and alignment cost (Eq. 4) from the $K = 2$ case to arbitrary K.

Theorem 3.2. *Consider a population of individuals with membership coefficients in $K \geq 2$ clusters. Suppose that in one replicate, the membership coefficients of the individuals follow a Dirichlet model with parameters $\mathbf{a} = (a_1, a_2, \ldots, a_K)$, and in a second replicate, they follow a Dirichlet model with parameters $\mathbf{b} = (b_1, b_2, \ldots, b_K)$. The mean contribution of a randomly chosen individual to the distance between replicates is*

$$A_{\mathbf{a},\mathbf{b}} = 2\left[\frac{\sum_{i=1}^{K-1}(a_i+1)a_i + \sum_{i=1}^{K-2}\sum_{j=i+1}^{K-1} a_i a_j}{(a_0+1)a_0} + \frac{\sum_{i=1}^{K-1}(b_i+1)b_i + \sum_{i=1}^{K-2}\sum_{j=i+1}^{K-1} b_i b_j}{(b_0+1)b_0} - \frac{\sum_{i=1}^{K-1} a_i b_i + (\sum_{i=1}^{K-1} a_i)(\sum_{i=1}^{K-1} b_i)}{a_0 b_0}\right]. \quad (10)$$

Recall that $a_0 = \sum_{i=1}^{K} a_i$ and $b_0 = \sum_{i=1}^{K} b_i$; the proof appears in Appendix C. Note that in the case of $K = 2$, Eq. 10 reduces to Eq. 5. If two replicates are aligned, then we can derive the mean contribution of an individual to the distance between replicates by substituting $b_i = a_i$ into Eq. 10 for all i.

Corollary 3.3. *The mean contribution to the distance between two aligned replicates*

of an individual whose membership coefficients follow a $\mathrm{Dir}(\mathbf{a})$ *distribution, where* $\mathbf{a} = (a_1, a_2, \ldots, a_K)$*, is*

$$A_{\mathbf{a},\mathbf{a}} = \frac{4\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} a_i a_j}{(a_0+1)a_0^2}. \tag{11}$$

We obtain this result by applying Theorem 3.2 to the identity permutation ϕ_0. For $K = 2$, Eq. 11 reduces to Eq. 6. For a general permutation ϕ, supposing that the two replicates are not necessarily correctly aligned, we calculate the contribution of a randomly chosen individual to the cost using Eqs. 10 and 11, following Eq. 4.

Corollary 3.4. *Suppose the membership coefficients of the individuals in a Dirichlet model follow* $\mathrm{Dir}(\mathbf{a})$ *where* $\mathbf{a} = (a_1, a_2, \ldots, a_K)$*. The mean contribution of an individual to the alignment cost for a second replicate whose parameters follow a Dirichlet model with permutation* $\phi(\mathbf{a})$ *is:*

$$C_{\mathbf{a},\phi(\mathbf{a})} = \frac{1}{a_0^2}\sum_{i=1}^{K} a_i(a_i - b_i). \tag{12}$$

Once again, for $K = 2$, Eq. 12 reduces to Eq. 8. For $K = 3$, Eq. 12 gives Table 1. With these general results on the alignment cost under label-switching now established, we proceed to analyze the effects of the parameters on the alignment cost.

4 Effect of the parameters under label-switching

4.1 Effect of the Dirichlet parameters with fixed permutation

For label-switching with a fixed permutation, the value of $C_{\mathbf{a},\phi(\mathbf{a})}$ in Eq. 12 is only affected by the values of the Dirichlet parameters. In general, as the difference between the parameter of a cluster and its corresponding parameter under permutation—the difference between a_i and $a_{\phi(i)}$—increases, $C_{\mathbf{a},\phi(\mathbf{a})}$ also increases.

We have already examined the effect of the Dirichlet parameters in cases with $K = 2$ and $K = 3$. In Figure 2, we examined the relationship between $C_{\mathbf{a},\phi(\mathbf{a})}$ and the a_i for the only non-identical permutation with $K = 2$, $\phi(1, 2) = (2, 1)$. The alignment cost was equal to 0 for $a_1 = a_2$, that is, when both clusters have mean membership 0.5. The cost increases as a_1 and a_2 become increasingly different.

For $K = 3$, similar relationships appear in Figure 3 for the permutation $(2, 3, 1)$. For $a_1 = a_2 = a_3$, the alignment cost reaches the minimum of zero. In panels A and C, with domain $[1, 25]$ for each of the three parameters, the alignment cost is maximal when one of the three is 25 and the other two are 1; in panel B, with a_3 fixed, the maximum occurs at $(a_1, a_2, a_3) = (1, 1, 13)$. The more diverged the values of three parameters, the higher the alignment cost. This result corresponds to the intuition that when clusters have distinct membership patterns, they are less easily mistaken for each other.

4.2 Effect of the permutation with fixed Dirichlet parameters

The cost for a permutation increases as its parameters increasingly diverge from the starting permutation. Suppose now that we consider the effect only of the permutation. The minimum possible value of $C_{\mathbf{a},\phi(\mathbf{a})}$ in Eq. 12 is 0. Clearly, the cost is 0 for two replicates in which the second replicate is unpermuted in relation to the first.

Consider a permutation cycle $\mathcal{PC}$: a subset of elements in the permutation ϕ that are permuted among themselves, so that $\phi(i) \in \mathcal{PC}$ for all $i \in \mathcal{PC}$ and $\phi(i) \notin \mathcal{PC}$ for all $i \notin \mathcal{PC}$. We interpret a *permutation cycle* as minimal in the sense that none of its proper subsets is a permutation cycle. Suppose ϕ is decomposed into $N_{\mathcal{PC}}$ permutation cycles. The set of all elements $I = \{1, 2, \ldots, K\}$ is the disjoint union of all permutation cycles: $I = \dot{\bigcup}_{h=1}^{N_{\mathcal{PC}}} \mathcal{PC}_h$. We have the following result.

Proposition 4.1. *Consider a permutation ϕ. $C_{\mathbf{a},\phi(\mathbf{a})} = 0$ if and only if for each permutation cycle $\mathcal{PC}_h$ in ϕ, there exists a constant $c_h > 0$ such that $a_i = c_h$ for all $i \in \mathcal{PC}_h$.*

The proposition states that the cost associated with ϕ is zero if and only if for all permutation cycles, all clusters in the permutation cycle have the same Dirichlet parameter.

Proof of Proposition 4.1. We prove the "if" direction first. Index the permutation cycles by h. Suppose for each h that $a_i = c_h$ for all $i \in \mathcal{PC}_h$. For each i, suppose that when ϕ is decomposed into permutation cycles, $\phi(i)$ is in permutation cycle $\mathcal{PC}_h$. Then $a_i = c_h$. Because cluster i and cluster $\phi(i)$ are in the same permutation cycle, $b_i = a_{\phi(i)} = a_i = c_h$. Therefore, $C_{\mathbf{a},\phi(\mathbf{a})} = (1/a_0^2) \sum_{i=1}^{K} a_i(a_i - a_i) = 0$.

For the "only if" direction, suppose $C_{\mathbf{a},\phi(\mathbf{a})} = (1/a_0^2)\sum_{i=1}^K a_i(a_i - b_i) = 0$. We have $\sum_{i=1}^K a_i b_i = \sum_{i=1}^K a_i^2$. Equivalently, $(\sum_{i=1}^K a_i b_i)^2 = (\sum_{i=1}^K a_i^2)^2 = (\sum_{i=1}^K a_i^2)(\sum_{i=1}^K b_i^2)$ because $\mathbf{b} = \phi(\mathbf{a})$ and all the Dirichlet parameters are positive. By the Cauchy-Schwarz inequality, $(\sum_{i=1}^K a_i b_i)^2 \leq (\sum_{i=1}^K a_i^2)(\sum_{i=1}^K b_i^2)$, with equality if and only if $\mathbf{a} = \alpha\mathbf{b}$ for some constant α. Because $\sum_{i=1}^K a_i = \sum_{i=1}^K b_i$, the only value α can take is 1, so that $a_i = b_i = a_{\phi(i)}$ for all $i = 1, 2, \ldots, K$. Note that i and $\phi(i)$ are in the same permutation cycle, say, $\mathcal{PC}_h$, by definition. We can denote by c_h the value $a_i = a_{\phi(i)}$; the value c_h applies for all i in $\mathcal{PC}_h$. Thus, assuming $C_{\mathbf{a},\phi(\mathbf{a})} = 0$ produces the conclusion that $a_i = c_h$ for all $i \in \mathcal{PC}_h$. □

Note that the proposition applies to the identity permutation $\phi = \phi_0$. The identity permutation places each cluster in its own permutation cycle, so that $\phi(i) = i$, $b_i = a_{\phi(i)} = a_i$, and $(1/a_0^2)\sum_{i=1}^K a_i(a_i - b_i) = 0$.

Consider an example of Proposition 4.1 with permutation $\phi_1(1, 2, 3, 4, 5) = (2, 1, 4, 5, 3)$. This permutation has two permutation cycles: $\{1, 2\}$ and $\{3, 4, 5\}$. If $a_1 = a_2$ and $a_3 = a_4 = a_5$, then $C_{\mathbf{a},\phi_1(\mathbf{a})} = 0$. This can be easily seen from the fact that in Eq. 12, $C_{\mathbf{a},\phi(\mathbf{a})} = 0$ if $a_i = b_i$ for all $i = 1, 2, \ldots, K$. When two clusters have the same parameter, they assign the same mean membership value, so that they are indistinguishable. Although ϕ is not the identity, it produces cost 0 because it only permutes indistinguishable clusters.

We also report the maximum possible cost as a function of the Dirichlet parameters, together with the permutation that gives this cost. This upper bound on the cost provides information on the worst-case misalignment possible given two replicates. Examining the form of Eq. 12, for fixed $(a_1, a_2, \ldots, a_K)$, this maximum can be calculated by minimizing $\sum_{i=1}^K a_i b_i$, where $\mathbf{b} = \phi(\mathbf{a})$, over permutations ϕ.

Proposition 4.2. *Fix Dirichlet parameters* $\mathbf{a}$. *Let* σ *describe a permutation that orders* $a_1, a_2, \ldots, a_K$ *with* $a_{\sigma(1)} \leq a_{\sigma(2)} \leq \ldots \leq a_{\sigma(K)}$. *Considering all permutations* ϕ, *the cost* $C_{\mathbf{a},\phi(\mathbf{a})}$ *is maximized when the permutation* ϕ *satisfies* $\phi(i) = \sigma(K - \sigma^{-1}(i) + 1)$ *for* $i = 1, 2, \ldots, K$. *The maximum cost is* $(1/a_0^2)\sum_{i=1}^K a_{\sigma(i)}\left[a_{\sigma(i)} - a_{\sigma(K-i+1)}\right]$.

The proposition states that the maximal value of $C_{\mathbf{a},\phi(\mathbf{a})}$ is attained when ϕ matches the largest parameter in $\mathbf{a}$ to the smallest value in $\mathbf{b}$, the second-largest value in $\mathbf{a}$ to the second-smallest value in $\mathbf{b}$, and so on.

Proof of Proposition 4.2. We make use of the rearrangement inequality (Steele, 2004, p. 78). Write $a_{\sigma(1)} \leq a_{\sigma(2)} \leq \ldots \leq a_{\sigma(K)}$, and write $r = \sigma^{-1}(i)$, denoting the rank order of a_i in the list $(a_1, a_2, \ldots, a_K)$, with 1 smallest and K largest.

By the rearrangement inequality, the minimum of $\sum_{i=1}^{K} a_i b_i = \sum_{i=1}^{K} a_i a_{\phi(i)} = \sum_{i=1}^{K} a_{\sigma(i)}\, a_{\phi(\sigma(i))}$ is attained when $\phi(i) = \sigma(K - r + 1)$ for each i. The value of $\sum_{i=1}^{K} a_i b_i$ at the minimum is

$$\min_{\phi} \sum_{i=1}^{K} a_i b_i = \sum_{i=1}^{K} a_i\, a_{\sigma(K-r+1)} = \sum_{i=1}^{K} a_{\sigma(i)} a_{\sigma(K-i+1)}.$$

□

With the permutations that produce minimal and maximal cost established, we can examine the effect of the permutation on cost more generally. Figure 4 shows the cost contributed by individuals in each of four populations, for all 24 permutations of four clusters with fixed Dirichlet parameters. With the parameters fixed, the contribution to alignment cost in general increases as more clusters are misaligned. This result can be seen in the fact that permutations with 3 or 4 misaligned clusters tend to lie toward the right side of the figure, which is ordered left to right by increasing cost; permutations with only 2 misaligned clusters tend to lie near the left side. The maximal cost follows Proposition 4.2: in panels A-C, the highest-cost permutation reverses the order of the mean memberships, as do the four highest-cost permutations with equal cost in panel D.

Although the general pattern is that an increase in the number of misaligned clusters increases the alignment cost, many counterexamples exist. For example, in Figure 4A, $\phi_1(1, 2, 3, 4) = (4, 2, 3, 1)$, with two clusters misaligned, has greater cost than $\phi_2(1, 2, 3, 4) = (4, 3, 1, 2)$, with all four clusters misaligned. Permutation ϕ_1 exchanges the two clusters with the greatest difference in mean, whereas permutation ϕ_2, while assigning cluster 1 to the distant cluster 4, performs a less costly exchange among clusters 2, 3, and 4 than mapping cluster 4 to cluster 1.

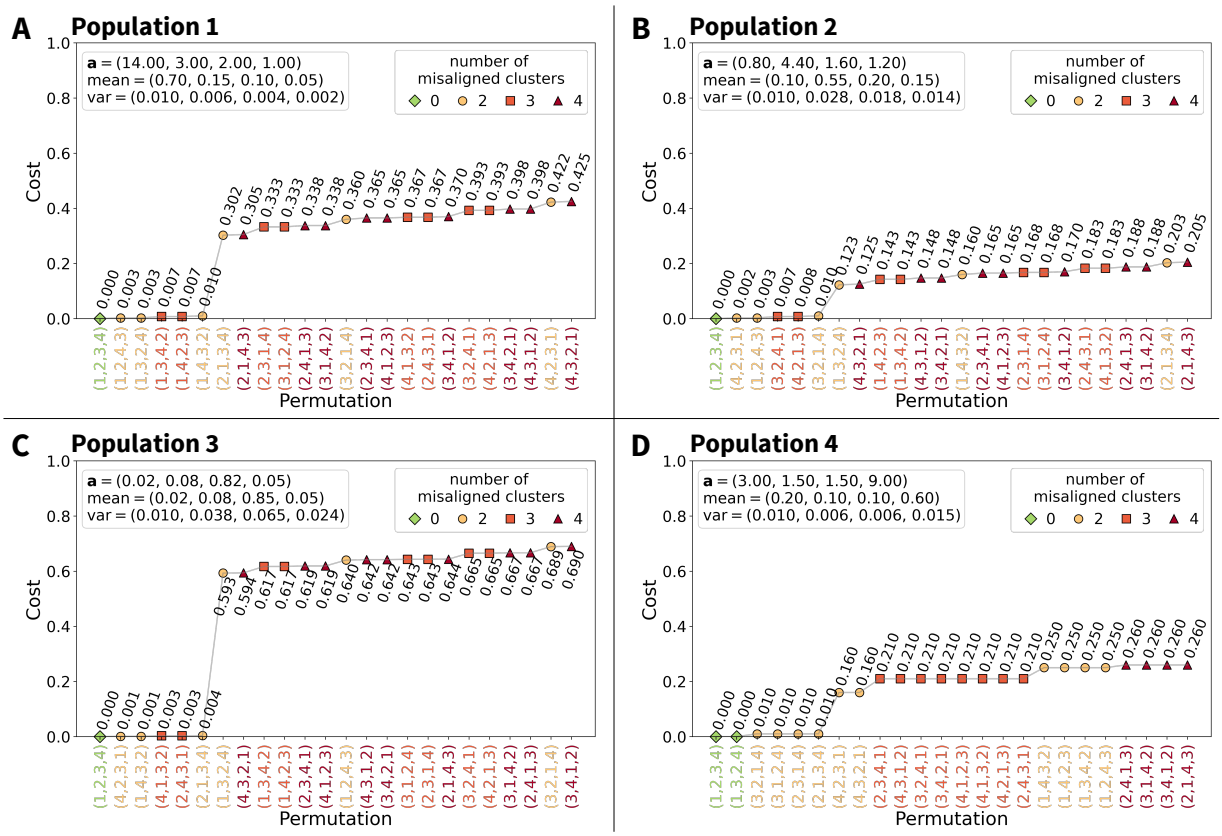

Figure 4: Alignment cost as a function of permutations. The figure considers an example cluster analysis in which four populations are placed into four clusters; each population has a different cluster in which membership predominates. In each panel, the 24 permutations of the four clusters are ordered by the alignment cost associated with an individual from a specific population. (A) Population 1. (B) Population 2. (C) Population 3. (D) Population 4. Simulation parameters appear in Appendix A. Permutations are labeled with respect to the original clusters $(1, 2, 3, 4)$, representing $\phi(1, 2, 3, 4)$ for each of the 24 possible choices of ϕ. In panel D, clusters 2 and 3 have the same parameter values; we do not count as misaligned clusters that map within a permutation cycle to a cluster with the same parameter value. The number of individuals per population is 100.

5 Multiple individuals

The theoretical contributions to distance between replicates (Theorem 3.2) and to alignment cost (Corollary 3.4) are both derived as expectations of random variables for a single individual in one population. We provide a simple extension to multiple individuals from multiple populations.

For multiple individuals from multiple populations, we can obtain an expected total contribution to distance between replicates and an expected total contribution to the alignment cost. We treat all individuals in a population as independent and identically distributed draws from the population. The expected mean total contribution of multiple individuals can be calculated by the linearity of expectation.

Proposition 5.1. *Consider L populations, in which population ℓ has N_ℓ individuals and membership coefficients $\mathbf{a}^{(\ell)} = (a_1^{(\ell)}, a_2^{(\ell)}, \ldots, a_K^{(\ell)})$ that follow a $\mathrm{Dir}(\mathbf{a}^{(\ell)})$ distribution, with $a_0^{(\ell)} = \sum_{k=1}^{K} a_k^{(\ell)}$. The total distance between two replicates under label-switching with permutation ϕ, which maps $a_i^{(\ell)}$ to $a_{\phi(i)}^{(\ell)}$, for $i = 1, 2, \ldots, K$, is*

$$A_{\phi,total} = \sum_{\ell=1}^{L} N_\ell A_{\mathbf{a}^{(\ell)},\phi(\mathbf{a}^{(\ell)})}, \tag{13}$$

and the total alignment cost is

$$C_{\phi,total} = \sum_{\ell=1}^{L} N_\ell C_{\mathbf{a}^{(\ell)},\phi(\mathbf{a}^{(\ell)})}. \tag{14}$$

Proof of Proposition 5.1. The proof is trivial. For a population with N_ℓ individuals and Dirichlet parameters $\mathbf{a}^{(\ell)}$, suppose two replicates follow a permutation ϕ for the second replicate in relation to the first. The membership coefficients of these individuals are independently drawn from the same Dirichlet distribution. Hence, by the linearity of expectation, the expected total contributions to distance and cost are sums across individuals, $N_\ell A_{\mathbf{a}^{(\ell)},\phi(\mathbf{a}^{(\ell)})}$ and $N_\ell C_{\mathbf{a}^{(\ell)},\phi(\mathbf{a}^{(\ell)})}$, respectively. For multiple populations, we simply sum expectations across populations. □

6 Example

We use data from the human-genetic ancestry inference study of Fortier et al. (2020) to illustrate the alignment cost in a practical setting. For the data, we assume that membership coefficients follow the Dirichlet model, whose parameters we then estimated. We then measure empirical alignment costs between pairs of replicates, comparing them to theoretical costs that result from using the estimated parameters of the Dirichlet distribution.

6.1 Data

Fortier et al. (2020) conducted clustering using *STRUCTURE* applied to 978 sampled individuals from $L = 53$ human populations, with $K = 4$. They performed analyses using a larger data set of 791 loci genotyped in the individuals and a less informative smaller subset containing, among the 791 loci, only 13 that are used in forensic genetics. The 53 populations vary in sample size, from 1 to 51 individuals.

For each analysis, 10 clustering replicates were performed, so that we have two sets of 10 replicates from Fortier et al. (2020), each with a 978×4 membership coefficient matrix. The individual membership coefficients of all replicates appear in Figure 5A (all 791 loci) and Figure 6A (13-locus subset). Fortier et al. (2020) summarized these replicates; we show all 10. With all 791 loci, most individuals are placed predominantly in one cluster; with the 13-locus subset, membership is more evenly distributed across clusters. We use the 791-locus analysis as an example of replicates with lower variability across individuals in membership coefficients within populations, and the 13-locus analysis as an example with greater variability, interpreted in this case as more "noise" in membership estimates.

6.2 Maximum likelihood estimation of Dirichlet parameters

Consider a membership matrix from population ℓ of sample size N_ℓ. The matrix has size $N_\ell \times K$, and it can be written $Q^{(\ell)} = (q_1^{(\ell)}, q_2^{(\ell)}, \ldots, q_{N_\ell}^{(\ell)})^T$, where $q_1^{(\ell)}, q_2^{(\ell)}, \ldots, q_{N_\ell}^{(\ell)}$ denote membership vectors for the N_ℓ individuals. If we assume that each of the N_ℓ vectors represents an independent multivariate draw from an underlying Dirichlet distribution with

parameter vector $\mathbf{a}$, then we can obtain a maximum likelihood estimate of $\mathbf{a}$ by maximizing log-likelihood $L(\mathbf{a})$. Taking the likelihood as a product of Eq. 1 across individuals, we have

$$L(\mathbf{a}) = \log \prod_{i=1}^{N_\ell} \mathbb{P}[q_i^{(\ell)}|\mathbf{a}] = N_\ell \left[\log \Gamma\left(\sum_{k=1}^{K} a_k \right) - \sum_{k=1}^{K} \log \Gamma(a_k) \right] + N_\ell \sum_{k=1}^{K} (a_k - 1) \log \bar{q}_k^{(\ell)}, \quad (15)$$

where $\log \bar{q}_k^{(\ell)} = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \log q_{ik}$.

This objective of maximizing $L(\mathbf{a})$ is equivalent to minimizing $-L(\mathbf{a})$, a convex function in $\mathbf{a}$. The minimization problem has no closed-form solution, but can be solved numerically. We use fixed-point iteration (Minka, 2000). The update step in the iteration is

$$a_k^{\text{new}} = \Psi^{-1}\left[\Psi\left(\sum_{k=1}^{K} a_k^{\text{old}} \right) + \log \bar{q}_k^{(\ell)} \right], \quad (16)$$

where $\Psi(x) = d \log \Gamma(x)/dx$ is the digamma function. The algorithm is guaranteed to converge to the maximizing $\mathbf{a}$ for the Dirichlet distribution (Minka, 2000, Section 1).

To use the fixed-point iteration method to numerically find the maximum likelihood estimate of $\mathbf{a} = (a_1, a_2, \ldots, a_K)$, we follow the method of Minka (2000, eqs. 19-21) to start the iteration from an initial guess for $\mathbf{a}$; this method relies on empirical computations of the means and variances of the $q_{ik}^{(\ell)}$ across individuals i. To obtain the update in Eq. 16, we apply Newton's method for solving $\Psi(x) = y$, following Minka (2000, Appendix C).

6.3 Empirical and theoretical alignment cost calculations

For both the 791-locus and 13-locus cases, we estimated the Dirichlet parameters for each of the 53 populations and each of the 10 replicates. For the single-individual group, because no variance among individuals is available, we cannot estimate the Dirichlet parameters, and we simply used the membership coefficients of the individual as the parameter estimates.

To examine the performance of the Dirichlet model in measuring alignment costs, we computed empirical and theoretical alignment costs between pairs of replicates. For each pair of replicates, we computed the total empirical alignment cost using Eq. 2 to obtain the sum of squared differences between their membership matrices. For individual i and

cluster k in a population ℓ with sample size N_ℓ individuals, denote by $q_{ik}^{(R_1,\ell)}$ and $q_{ik}^{(R_2,\ell)}$ the membership coefficients in replicates R_1 and R_2. The sum is

$$D_{R_1,R_2} = \sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell}\sum_{k=1}^{K}(q_{ik}^{(R_1,\ell)} - q_{ik}^{(R_2,\ell)})^2. \tag{17}$$

For the theoretical computation, we first used the inferred permutation between replicate 1 and subsequent replicates, as provided by CLUMPP and reported by Fortier et al. (2020), as the "correct" alignments. We next computed the theoretical contribution to alignment cost for each of the 53 populations based on the inferred pairwise permutation (Eq. 12), aggregating the contributions from all populations following Eq. 14.

More precisely, suppose that for a pair of replicates (R_1, R_2), the inferred Dirichlet parameters for the L populations are $\{\mathbf{a}^{(R_1,\ell)}\}_{\ell=1,2,\ldots,L}$ and $\{\mathbf{a}^{(R_2,\ell)}\}_{\ell=1,2,\ldots,L}$. First, we choose R_1 as the "base" replicate, and denote the permutation in replicate R_2 with respect to R_1 as $\phi^{R_1\to R_2}$. The total theoretical cost in this situation is

$$C_{\phi^{R_1\to R_2},\text{total}} = \sum_{\ell=1}^{L} N_\ell C_{\mathbf{a}^{(R_1,\ell)},\mathbf{a}^{(R_2,\ell)}}. \tag{18}$$

Next, we use R_2 as the base, and $\phi^{R_2\to R_1}$ is the permutation in replicate R_1 with respect to R_2. The cost is

$$C_{\phi^{R_2\to R_1},\text{total}} = \sum_{\ell=1}^{L} N_\ell C_{\mathbf{a}^{(R_2,\ell)},\mathbf{a}^{(R_1,\ell)}}.$$

Note that in general, $C_{\mathbf{a}^{(R_1,\ell)},\mathbf{a}^{(R_2,\ell)}} \neq C_{\mathbf{a}^{(R_2,\ell)},\mathbf{a}^{(R_1,\ell)}}$, because the inferred Dirichlet parameters differ for R_1 and R_2. To account for this asymmetry, we take as the theoretical alignment cost the mean of the two values:

$$C_{\text{total}(R_1,R_2)} = \frac{1}{2}(C_{\phi^{R_1\to R_2},\text{total}} + C_{\phi^{R_2\to R_1},\text{total}}). \tag{19}$$

6.4 Data analysis

The greater variability of membership coefficients in the 13-locus case compared to the 791-locus case is depicted in Figure 7, both on the basis of the empirical variance in mem-

bership coefficients (Figure 7A and C) and using the theoretical variance computed from the estimated Dirichlet parameters (Figure 7B and D). We interpret the alignment costs in relation to this observation concerning variability in the two cases.

For the 791-locus case, averaging across individuals, Figure 5B reports pairwise empirical costs between replicates and Figure 5C reports theoretical costs. The relative difference between empirical and theoretical costs, computed with respect to the theoretical cost as

$$\frac{\left|C_{\text{total}(R_1,R_2)} - D_{R_1,R_2}\right|}{C_{\text{total}(R_1,R_2)}}, \tag{20}$$

appears in Figure 5D. The theoretical cost in Figure 5C generally accords with the empirical cost in Figure 5B. The relative difference in Figure 5D is small for most pairs of replicates.

For another assessment of the agreement of theoretical and empirical alignment costs, using replicate 1 as the base, we computed the theoretical alignment cost for all 24 permutations of the $K = 4$ clusters (Figure 5E). If a permutation was observed among replicates 2 to 10, then its empirical cost with respect to replicate 1 is also shown; if multiple replicates possess the same permutation, then we take their mean cost. This analysis finds that empirical and theoretical costs agree across permutations with a wide range of cost values.

Comparing Figures 6 and 5, Figure 6 reports corresponding quantities for the 13-locus case. The empirical (Figure 6B) and theoretical (Figure 6C) alignment costs have lower values than in the 791-locus case. In Figure 6A, in which individuals possess high variability within populations, comparing to the low-variability replicates of Figure 5, it is less easily discerned that an alignment is suboptimal; the lower alignment costs for the high-variability 13-locus case compared to the low-variability 791-locus case reflect this observation.

The agreement between theoretical and empirical alignment costs is reduced for Figure 6 compared to Figure 5, with a substantial difference between the theoretical costs in Figure 6C and the empirical costs in Figure 6B. The relative difference is high in Figure 6D, and the empirical costs differ from the theoretical costs for many permutations in Figure 6E. The greater disagreement between theoretical and empirical costs suggests that for the high-variability 13-locus case, the Dirichlet model provides a poorer fit to the replicates than in the low-variability 791-locus case.

7 Discussion

We have used a Dirichlet model to study the membership coefficients produced by mixed-membership unsupervised clustering algorithms (Section 2.2). Under the Dirichlet model, using a theoretical measure for the alignment cost between clustering replicates (Eq. 4), we have evaluated the alignment cost for a pair of clustering replicates as a function of the model parameters. The model provides tools for use in evaluating clustering replicates, both in analyses of specific data sets and in assessing the performance of clustering algorithms.

Under the model, Corollary 3.4 describes the cost of one replicate in relation to another, making use of the general Theorem 3.2. A replicate with N individuals and K clusters—and hence, NK data entries—is summarized with K parameters, one for each cluster. Theorem 3.2 and Corollary 3.4 provide relatively simple expressions in terms of the K parameter values for each of two replicates. We have evaluated these expressions for the special cases of $K = 2$ (Section 3.2) and $K = 3$ (Section 3.3), for which they reduce further.

In analyzing the properties of the theoretical cost as a function of the Dirichlet parameters, we have seen that for a fixed permutation between a pair of replicates, the cost increases as the Dirichlet parameters of the two replicates diverge (Section 4.1). We have also seen that when the Dirichlet parameters are fixed, the cost increases with the number of misaligned clusters (Section 4.2). However, this result depends in part on the specific permutation, as certain permutations might produce lower cost than others with fewer misaligned clusters. When all clusters within the same permutation cycle share common Dirichlet parameters, none of these clusters are "misaligned," and the cost is zero (Proposition 4.1). We have also described the maximal cost across permutations (Proposition 4.2), potentially enabling cost functions to be normalized by the maximum across permutations.

In an example data analysis, we have found that the Dirichlet model closely fits data with low variability in estimated cluster memberships across individuals within populations (Section 6.4). The fit is not as close for data with high variability in cluster memberships, and hence with more "noise." However, alignment costs are smaller for such cases; in noisy data, the model is poorer but the distinction between properly aligned and misaligned replicates is less consequential.

We envision several applications for the Dirichlet model and its associated results. First, the model can be used to provide summary statistics for replicate mixed-membership cluster analyses. The model would first be used to estimate Dirichlet parameters for replicates. Theoretical alignment costs of permutations of those replicates could then be calculated from the parameters, measuring the cost difference between the optimal permutation and suboptimal permutations. As functions of the estimated parameters, the cost distribution of permutations, the cost difference between optimal and suboptimal permutations, and the maximal cost across permutations can all provide informative summaries.

Such summary statistics could potentially be applied in diagnostics for alignments. The cost difference between the optimal and least suboptimal permutation can measure the extent to which the optimal permutation of a replicate is evident—the "noise" in the replicate—guiding computational decisions for identifying optimal alignments. In particular, if noise is low and the cost for suboptimal permutations is high, then the optimal replicate is likely to be relatively easy to identify, and choices that prioritize speed rather than comprehensive searches in existing alignment algorithms might be suitable.

Next, using the model, clustering alignment methods could adopt heuristic threshold values to decide when to stop the search for a better alignment, or to decide if two replicates represent substantially different modes or merely represent label-switching (Jakobsson and Rosenberg, 2007; Kopelman et al., 2015). Such threshold values could potentially be tuned prior to application of the alignment methods, employing our maximal cost computation. The theoretical alignment cost can thus provide an automated method of choosing threshold values suited to particular data sets, as the theoretical calculation of costs associated with label-switching would be performed in place of more computationally intensive empirical calculations.

Finally, and potentially most significantly, methods based on the model have the potential to contribute to new alignment algorithms. Existing algorithms rely on empirical cost calculations between pairs of replicates. Using the model, however, once the Dirichlet parameters have been estimated, theoretical alignment costs calculated from the estimated parameters potentially reduce computation time. In particular, when it makes sense to

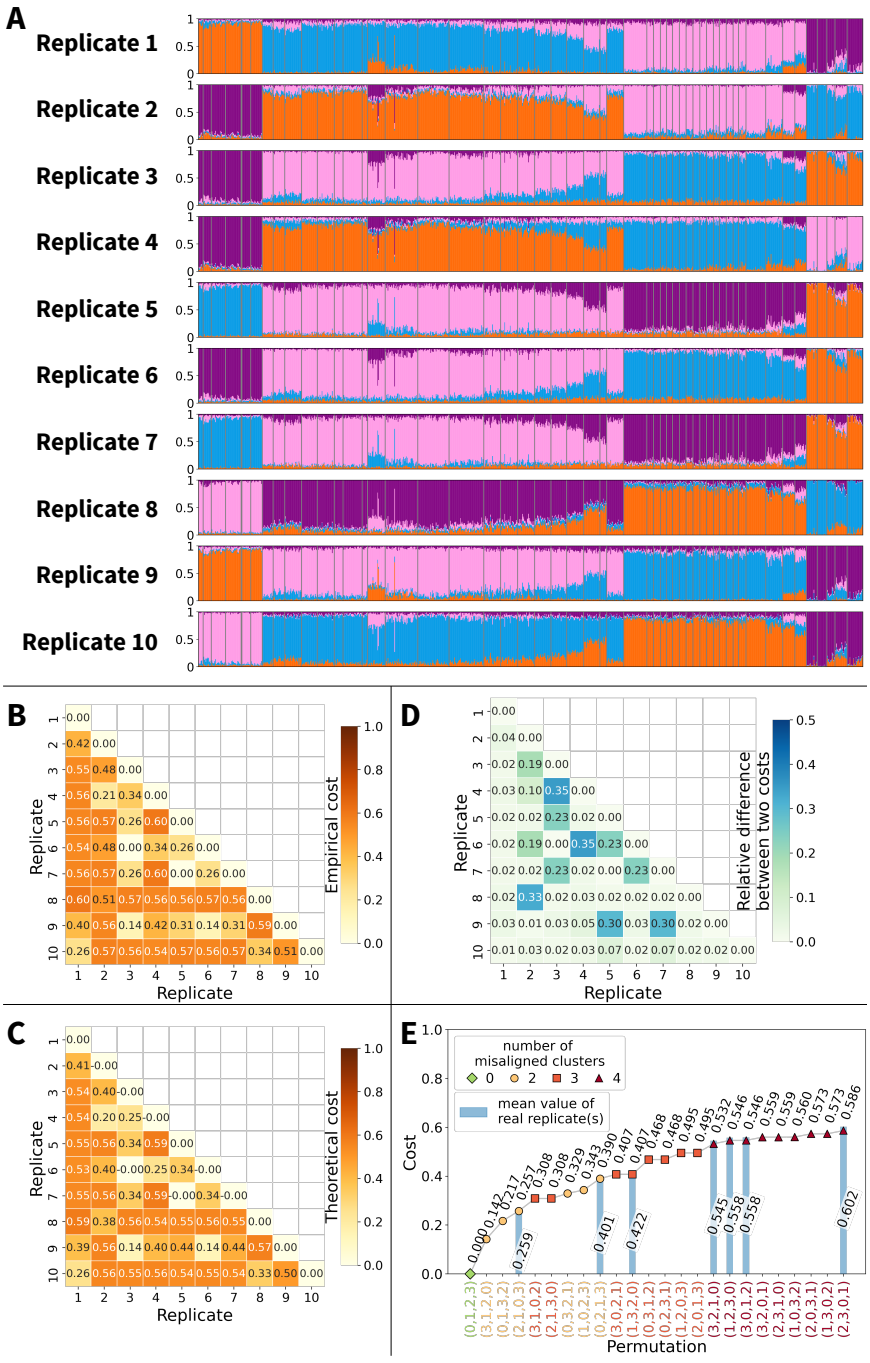

Figure 5: Application of the Dirichlet model to data from 791 loci. (A) 10 clustering replicates. (B) Empirical alignment cost between pairs of replicates, following Eq. 17, divided by the total number of individuals $N = \sum_{\ell=1}^{L} N_\ell$. (C) Theoretical alignment cost between pairs of replicates, divided by the total number of individuals. The symmetric Eq. 19 is used for the computation. (D) Relative difference between empirical and theoretical alignment cost for pairs of replicates, evaluated using Eq. 20. (E) Theoretical alignment costs for all possible permutations of replicate 1 and empirical alignment costs for replicates 2 to 10 in relation to replicate 1 (blue lines). The theoretical computation uses Eq. 18 and the empirical computation uses Eq. 17.

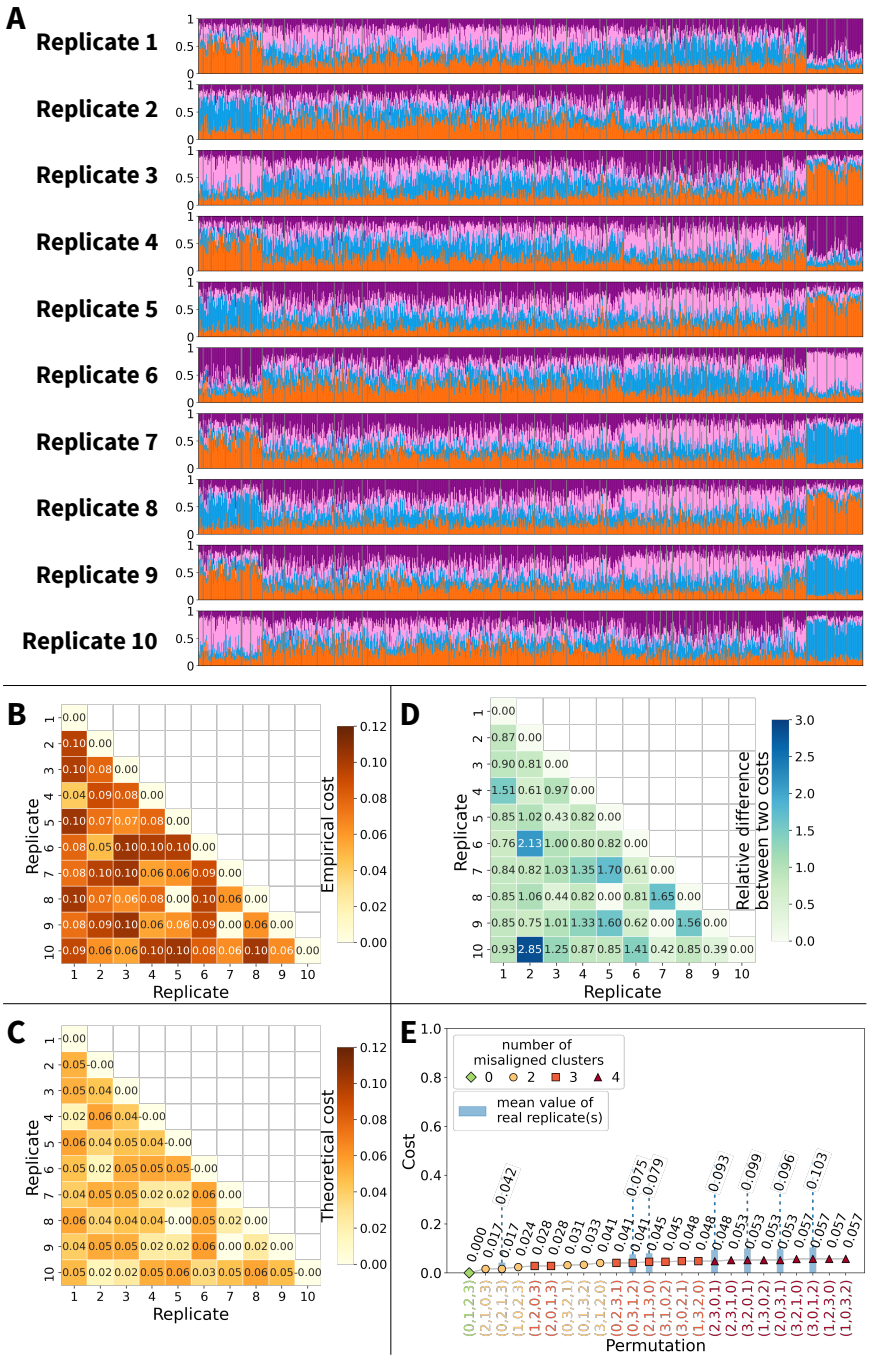

Figure 6: Application of the Dirichlet model to data from 13 loci. (A) 10 clustering replicates. (B) Empirical alignment cost between pairs of replicates, following Eq. 17, divided by the total number of individuals $N = \sum_{\ell=1}^{L} N_\ell$. (C) Theoretical alignment cost between pairs of replicates, divided by the total number of individuals. The symmetric Eq. 19 is used for the computation. (D) Relative difference between empirical and theoretical alignment cost for pairs of replicates, evaluated using Eq. 20. (E) Theoretical alignment costs for all possible permutations of replicate 1 and empirical alignment costs for replicates 2 to 10 in relation to replicate 1 (blue lines). The theoretical computation uses Eq. 18 and the empirical computation uses Eq. 17.

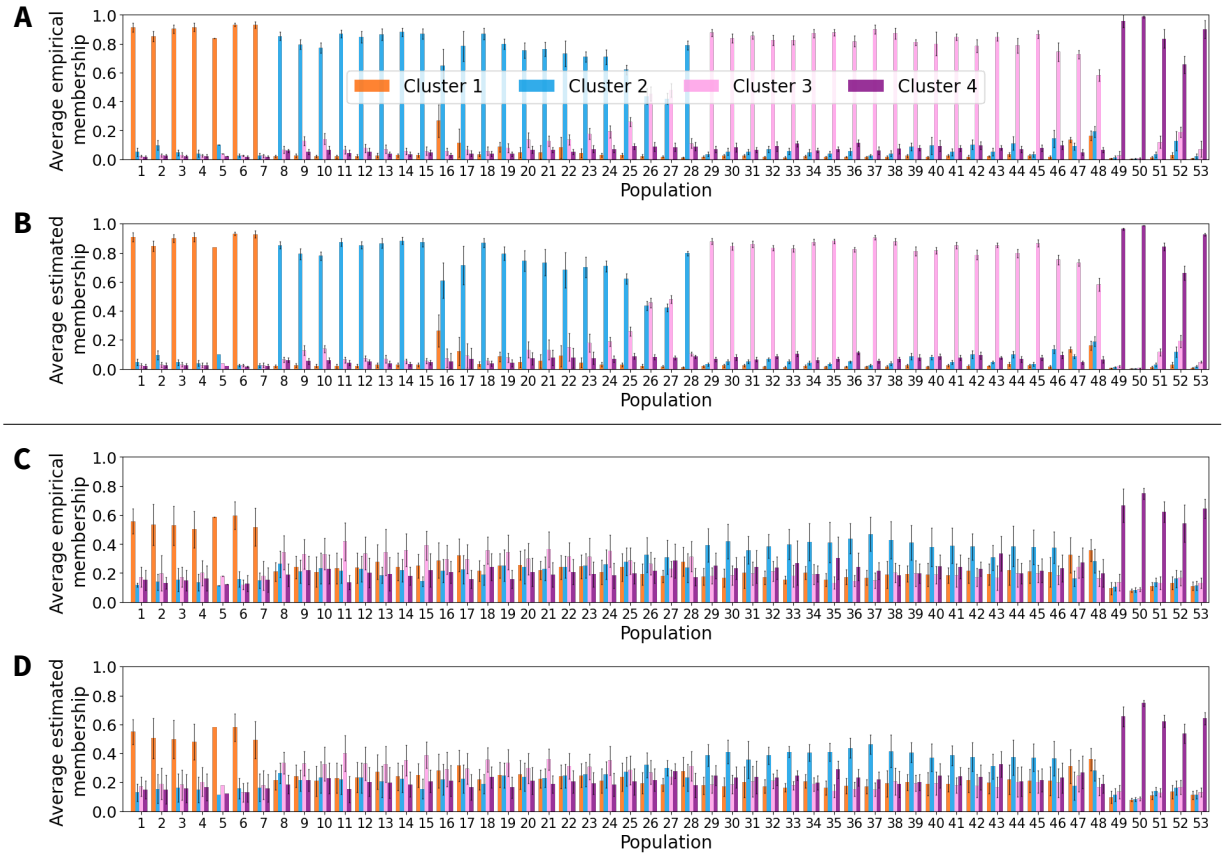

Figure 7: Mean and standard deviation of membership coefficients in 53 populations. (A) Empirical, 791 loci. (B) Theoretical, 791 loci. (C) Empirical, 13 loci. (D) Theoretical, 13 loci. Panels A and B consider replicate 1 from Figure 5; panels C and D consider replicate 1 from Figure 6. Standard deviations appear as error bars. Population 34 has only one individual and no empirical standard deviation; a theoretical standard deviation is induced by the choice to equate Dirichlet parameters $\mathbf{a}$ with the membership coefficient vector for the individual.

treat individual members of predefined populations as identically distributed, the membership coefficients for the N_ℓ members of a population can be summarized with K parameters in place of $N_\ell K$ data points. Alignment can then proceed based on computations involving K theoretical values rather than $N_\ell K$ empirical values, reducing computation time compared to use of empirical costs. Notably, as the problem of identifying optimal permutations of replicates can be understood as an example of a general class of *assignment problems* (Burkard et al., 2012) seen in combinatorial optimization and operations research, this use of the model can potentially contribute to alignment problems beyond the genetics context.

We note a number of limitations. First, the utility of the Dirichlet model is more limited in cases in which the model provides a poor fit to the data. However, we have seen that the fit can be assessed by comparing empirical and theoretical alignment costs, so that applicability of the model can be assessed for a particular data set.

A second limitation is that the theoretical cost measure does not consider the scenario in which replicates possess different numbers of clusters. The approach, however, can be extended. Consider two replicates, replicate 1 with K_1 clusters and replicate 2 with $K_2 > K_1$ clusters. In principle, it is possible to sum membership coefficients in each of K_1 disjoint subsets of the K_2 clusters and to then evaluate alignment cost between the K_1 clusters of replicate 1 and the K_1 subsets for replicate 2. This computation can be performed in principle for each way of distributing the K_2 clusters over K_1 subsets. The Stirling number of the second kind, $S_2(K_2, K_1)$, counts the partitions of K_2 labeled objects into K_1 unlabeled classes, in such a way that each of the K_1 classes contains at least one of the K_2 objects; the number of scenarios that must be considered is the number of ways of distributing the K_2 labeled clusters over K_1 *labeled* subsets, or $S_2(K_2, K_1)\, K_1!$.

A third limitation comes from our choice of distance measure. The derivation of the theoretical alignment cost relied on the squared 2-norm of the difference between membership vectors as the distance between replicates. Various distances have previously been used to compare pairs of replicates (Rosenberg et al., 2002; Jakobsson and Rosenberg, 2007; Kopelman et al., 2015; Behr et al., 2016). Other measures, including other p-norms, could

be used in Eq. 3 to enlarge small differences between vectors (lower p) or to reduce them (higher p); a new cost computation under the Dirichlet model would then be required.

Although our study has been motivated by the setting of unsupervised clustering in population genetics, the Dirichlet model applies to mixed-membership clustering more generally. Hence, our analysis of the model and its performance can contribute to other fields where cluster analysis—and particularly unsupervised cluster analysis—is used.

Funding. We acknowledge National Institutes of Health grant R01 HG005855 and National Science Foundation grant NSF BCS-2116322.

Disclosure. The authors declare that they have no competing interests.

Supplemental material

Title: AlignmentCost

Python package "AlignmentCost" for analyzing alignment cost. Python package "AlignmentCost" contains code to perform the empirical data analysis described in the article (Figures 5 and 6), including functions for computing the alignment cost (eqs. 10-12) and estimating the Dirichlet parameters (Section 6.2). The package also contains the empirical datasets used as examples in the article. (AlignmentCost.zip, ZIP file)

References

Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Feinberg, S. E. (2015a), "Introduction to mixed membership models and methods," in Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Feinberg, S. E. (editors), *Handbook of Mixed Membership Models and their Applications*, Boca Raton: Chapman & Hall/CRC, 3–13.

Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (2015b), *Handbook of Mixed Membership Models and Their Applications*, CRC press.

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

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

Burkard, R., Dell'Amico, M., and Martello, S. (2012), *Assignment Problems*, Philadelphia: Society for Industrial and Applied Mathematics.

Corander, J., Waldmann, P., and Sillanpää, M. J. (2003), "Bayesian analysis of genetic differentiation between populations," *Genetics*, 163, 367–374.

Fortier, A. L., Kim, J., and Rosenberg, N. A. (2020), "Human-genetic ancestry inference and false positives in forensic familial searching," *G3: Genes, Genomes, Genetics*, 10, 2893–2902.

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

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005), "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, 20, 50–67.

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

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2004), *Continuous Multivariate Distributions Vol 1: Models and Applications*, New York: John Wiley & Sons.

Minka, T. (2000), "Estimating a Dirichlet distribution," *MIT Technical Report*.

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

Rosenberg, N. A. (2004), "DISTRUCT: a program for the graphical display of population structure," *Molecular Ecology Notes*, 4, 137–138.

Rosenberg, N. A., Burke, T., Elo, K., Feldman, M. W., Freidlin, P. J., Groenen, M. A., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., et al. (2001), "Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds," *Genetics*, 159, 699–713.

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

Steele, J. M. (2004), *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*, Cambridge University Press.

Stephens, M. (2000), "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62, 795–809.

Appendix A Dirichlet parameters used in generating figures

Using the Dirichlet model, we simulated membership coefficients of N_ℓ individuals for each of four populations, $\ell = 1, 2, 3, 4$, using 50 and 100 for N_ℓ in the various analyses. Dirichlet parameters $\mathbf{a}^{(\ell)} = (a_1^{(\ell)}, a_2^{(\ell)}, a_3^{(\ell)}, a_4^{(\ell)})$ were chosen so that membership coefficients had specified values for expectations and variances. The variance of Dirichlet variables for the first cluster was set at $c = 0.001$ for Figure 1 and $c = 0.01$ for Figure 4. That is,

$$\frac{a_k^{(\ell)}(a_0^{(\ell)} - a_k^{(\ell)})}{(a_0^{(\ell)})^2(a_0^{(\ell)} + 1)} = c \quad \text{for} \quad k = 1, \tag{21}$$

where $a_0^{(\ell)} = \sum_{k=1}^4 a_k^{(\ell)}$, for $\ell = 1, 2, 3, 4$. With the four means specified and the variance specified for the membership coefficient of the first cluster, the Dirichlet parameters

are uniquely specified. The parameters and the corresponding mean and variance of the Dirichlet distributions appear in Table 2.

Appendix B Proof of Theorem 3.1

When $K = 2$, we have

$$\begin{aligned} A_{\mathbf{a},\mathbf{b}} &= \int_{p_1=0}^{1}\int_{q_1=0}^{1}\left((p_1-q_1)^2+\left[(1-p_1)-(1-q_1)\right]^2\right)\frac{p_1^{a_1-1}(1-p_1)^{a_2-1}}{\Gamma(a_1)\,\Gamma(a_2)/\Gamma(a_1+a_2)} \\ &\quad\times\frac{q_1^{b_1-1}(1-q_1)^{b_2-1}}{\Gamma(b_1)\,\Gamma(b_2)/\Gamma(b_1+b_2)}\,dq_1\,dp_1 \\ &= 2\frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)}\frac{\Gamma(b_1+b_2)}{\Gamma(b_1)\Gamma(b_2)}\times\int_{p_1=0}^{1}p_1^{a_1-1}(1-p_1)^{a_2-1} \qquad (22) \\ &\quad\times\int_{q_1=0}^{1}\left[p_1^2q_1^{b_1-1}(1-q_1)^{b_2-1}-2p_1q_1^{b_1}(1-q_1)^{b_2-1}+q_1^{b_1+1}(1-q_1)^{b_2-1}\right]dq_1\,dp_1. \end{aligned}$$

We now apply the beta integral $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \Gamma(a)\,\Gamma(b)/\Gamma(a+b)$ sequentially to the inner integral and then the outer integral of Eq. 22, obtaining

$$\begin{aligned} A_{\mathbf{a},\mathbf{b}} &= 2\frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)}\frac{\Gamma(b_1+b_2)}{\Gamma(b_1)\Gamma(b_2)}\int_{p_1=0}^{1}p_1^{a_1-1}(1-p_1)^{a_2-1}\left[p_1^2\frac{\Gamma(b_1)\,\Gamma(b_2)}{\Gamma(b_1+b_2)}\right. \\ &\quad\left.-2p_1\frac{\Gamma(b_1+1)\,\Gamma(b_2)}{\Gamma(b_1+b_2+1)}+\frac{\Gamma(b_1+2)\,\Gamma(b_2)}{\Gamma(b_1+b_2+2)}\right]dp_1 \\ &= 2\frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\,\Gamma(a_2)}\left[\frac{\Gamma(a_1+2)\,\Gamma(a_2)}{\Gamma(a_1+a_2+2)}-\frac{2b_1}{b_1+b_2}\frac{\Gamma(a_1+1)\,\Gamma(a_2)}{\Gamma(a_1+a_2+1)}\right. \\ &\quad\left.+\frac{(b_1+1)b_1}{(b_1+b_2)(b_1+b_2+1)}\frac{\Gamma(a_1)\,\Gamma(a_2)}{\Gamma(a_1+a_2)}\right]. \qquad (23) \end{aligned}$$

Finally, we simplify using $\Gamma(x+1) = x\Gamma(x)$ to obtain Eq. 5.

Table 2: Model parameters for simulating clustering examples in Figures 1 and 4

Figure(s)	Population	Cluster	a_1	a_2	a_3	a_4
Figure 1A, 1B, and 1C replicate 1	1	Parameter	146.3	31.35	20.9	10.45
		Mean	0.7	0.15	0.1	0.05
		Variance	0.001	6×10^{-4}	4×10^{-4}	2×10^{-4}
	2	Parameter	24.65	135.575	49.3	36.975
		Mean	0.1	0.55	0.2	0.15
		Variance	3×10^{-4}	0.001	6×10^{-4}	5×10^{-4}
	3	Parameter	2.53	10.12	107.525	6.325
		Mean	0.02	0.08	0.85	0.05
		Variance	1×10^{-4}	5×10^{-4}	0.001	3×10^{-4}
	4	Parameter	47.8	23.9	23.9	143.4
		Mean	0.2	0.1	0.1	0.6
		Variance	7×10^{-4}	4×10^{-4}	4×10^{-4}	0.001
Figure 1C replicate 2	1	Parameter	95.6	47.8	23.9	71.7
		Mean	0.4	0.2	0.1	0.3
		Variance	0.001	6×10^{-4}	4×10^{-4}	9×10^{-4}
	2	Parameter	62.7	62.7	41.8	41.8
		Mean	0.3	0.3	0.2	0.2
		Variance	0.001	0.001	8×10^{-4}	8×10^{-4}
	3	Parameter	37.35	49.8	124.5	37.35
		Mean	0.15	0.2	0.5	0.15
		Variance	5×10^{-4}	6×10^{-4}	0.001	5×10^{-4}
	4	Parameter	12.65	5.06	1.265	107.525
		Mean	0.1	0.04	0.01	0.85
		Variance	7×10^{-4}	3×10^{-4}	8×10^{-5}	0.001
Figures 4	1	Parameter	14	3	2	1
		Mean	0.7	0.15	0.1	0.05
		Variance	0.01	0.006	0.004	0.002
	2	Parameter	0.8	4.4	1.6	1.2
		Mean	0.1	0.55	0.2	0.15
		Variance	0.01	0.028	0.018	0.014
	3	Parameter	0.019	0.077	0.816	0.048
		Mean	0.02	0.08	0.85	0.05
		Variance	0.01	0.038	0.065	0.024
	4	Parameter	3	1.5	1.5	9
		Mean	0.2	0.1	0.1	0.6
		Variance	0.01	0.006	0.006	0.015

Appendix C Proof of Theorem 3.2

The proof entails a calculation of the multiple integral in Eq. 3. This integral can be rearranged in a nested way:

$$\begin{aligned} A_{\mathbf{a},\mathbf{b}} &= 2\frac{\Gamma(\sum_{i=1}^K a_i)}{\prod_{i=1}^K \Gamma(a_i)}\frac{\Gamma(\sum_{i=1}^K b_i)}{\prod_{i=1}^K \Gamma(b_i)} \int_{p_1=0}^1 p_1^{a_1-1} \int_{p_2=0}^{1-p_1} p_2^{a_2-1} \cdots \int_{p_{K-1}=0}^{1-\sum_{i=1}^{K-2} p_i} p_{K-1}^{a_{K-1}-1} \\ &\times \left(1-\sum_{i=1}^{K-1} p_i\right)^{a_K-1} \int_{q_1=0}^1 q_1^{b_1-1} \int_{q_2=0}^{1-q_1} q_2^{b_2-1} \cdots \int_{q_{K-1}=0}^{1-\sum_{i=1}^{K-2} q_i} q_{K-1}^{b_{K-1}-1} \\ &\times \left(1-\sum_{i=1}^{K-1} q_i\right)^{b_K-1} \left(\sum_{i=1}^{K-1} p_i^2 + \sum_{i=1}^{K-1} q_i^2 - \sum_{i=1}^{K-1}\sum_{j=1}^{K-1} p_i q_j - \sum_{i=1}^{K-1} p_i q_i + \sum_{i=1}^{K-2}\sum_{j=i+1}^{K-1} p_i p_j \right. \\ &\left. + \sum_{i=1}^{K-2}\sum_{j=i+1}^{K-1} q_i q_j \right) dq_{K-1} \cdots dq_2\, dq_1\, dp_{K-1} \cdots dp_2\, dp_1. \end{aligned} \tag{24}$$

We make use of the mean and variance of the Dirichlet distribution, so that for $(p_1, p_2, \ldots, p_K)$ Dirichlet-distributed with parameters $(a_1, a_2, \ldots, a_K)$ and $(q_1, q_2, \ldots, q_K)$ Dirichlet-distributed with parameters $(b_1, b_2, \ldots, b_K)$, with $a_0 = \sum_{i=1}^K a_i$ and $b_0 = \sum_{i=1}^K b_i$, we have (Kotz et al., 2004, eq. 49.9)

$$\mathbb{E}[p_i] = \frac{a_i}{a_0} \tag{25}$$

$$\mathbb{E}[q_i] = \frac{b_i}{b_0} \tag{26}$$

$$\mathrm{Var}[p_i] = \frac{a_i(a_0-a_i)}{{a_0}^2(a_0+1)} \tag{27}$$

$$\mathrm{Var}[q_i] = \frac{a_i(a_0-a_i)}{{a_0}^2(a_0+1)} \tag{28}$$

$$\mathrm{Cov}(p_i, p_j) = -\frac{a_i a_j}{a_0^2(a_0+1)},\ i \neq j \tag{29}$$

$$\mathrm{Cov}(q_i, q_j) = -\frac{b_i b_j}{b_0^2(b_0+1)},\ i \neq j. \tag{30}$$

Using these results, we have

$$\mathbb{E}[p_i^2] = \mathrm{Var}[p_i] + \mathbb{E}[p_i]^2 = \frac{a_i(a_i+1)}{a_0(a_0+1)} \tag{31}$$

$$\mathbb{E}[q_i^2] = \mathrm{Var}[q_i] + \mathbb{E}[q_i]^2 = \frac{b_i(b_i+1)}{b_0(b_0+1)} \tag{32}$$

$$\mathbb{E}[p_i p_j] = \mathrm{Cov}(p_i, p_j) + \mathbb{E}[p_i]\,\mathbb{E}[p_j] = \frac{a_i a_j}{a_0(a_0+1)},\ i \neq j \tag{33}$$

$$\mathbb{E}[q_i q_j] = \mathrm{Cov}(q_i, q_j) + \mathbb{E}[q_i]\,\mathbb{E}[q_j] = \frac{b_i b_j}{b_0(b_0+1)},\ i \neq j. \tag{34}$$

Because $\mathbf{p}$ and $\mathbf{q}$ are independent, we also have

$$\mathbb{E}[p_i q_j] = \mathbb{E}[p_i]\,\mathbb{E}[q_j]. \tag{35}$$

In the integral in Eq. 24, each term in the sum $\sum_{i=1}^{K-1} p_i^2 + \sum_{i=1}^{K-1} q_i^2 - \sum_{i=1}^{K-1}\sum_{j=1}^{K-1} p_i q_j - \sum_{i=1}^{K-1} p_i q_i + \sum_{i=1}^{K-2}\sum_{j=i+1}^{K-1} p_i p_j + \sum_{i=1}^{K-2}\sum_{j=i+1}^{K-1} q_i q_j$ is integrated with respect to the Dirichlet density over two simplices, one for $\mathbf{p}$ and one for $\mathbf{q}$. Hence, each term can be integrated by one of Eqs. 25-35: Eq. 31 for terms p_i^2, Eq. 32 for terms q_i^2, Eq. 35 for terms $p_i q_j$ ($j = i$ and $j \neq i$), Eq. 33 for terms $p_i p_j$ ($j \neq i$), and Eq. 34 for terms $q_i q_j$ ($j \neq i$).

The integral becomes:

$$A_{\mathbf{a},\mathbf{b}} = 2\left(\sum_{i=1}^{K-1}\mathbb{E}[p_i^2] + \sum_{i=1}^{K-1}\mathbb{E}[q_i^2] - \sum_{i=1}^{K-1}\sum_{j=1}^{K-1}\mathbb{E}[p_i q_j] - \sum_{i=1}^{K-1}\mathbb{E}[p_i q_i] + \sum_{i=1}^{K-2}\sum_{j=i+1}^{K-1}\mathbb{E}[p_i p_j] + \sum_{i=1}^{K-2}\sum_{j=i+1}^{K-1}\mathbb{E}[q_i q_j]\right).$$

Simplifying using Eqs. 25-35, we obtain Eq. 10, concluding the proof.