# Population stratification using a statistical model on hypergraphs

Alexei Vazquez
*The Simons Center for Systems Biology Institute for Advanced Study, Einstein Drive, Princeton, New Jersey 08540, USA*
(Received 10 December 2007; published 10 June 2008)

Population stratification is a problem encountered in several areas of natural science, engineering, and public health. We tackle this problem by mapping a population and its element attributes onto a hypergraph, a natural extension of the concept of graph or network to encode associations among any number of elements. On this hypergraph, we construct a statistical model reflecting our intuition about how the element attributes can emerge from a postulated population structure. Finally, we introduce the concept of stratification representativeness as a mean to identify the simplest stratification already containing most of the information about the population structure. We demonstrate the power of this framework stratifying an animal and a human population based on phenotypic and genotypic properties, respectively.

DOI: 10.1103/PhysRevE.77.066106 PACS number(s): 89.75.Hc, 89.75.Fb, 02.50.Tt

## I. INTRODUCTION

A population stratification problem consists of uncovering the structure of a population of individuals, samples, or elements given a list of attributes characterizing them. For example, the design of a zoo requires us to understand what is the best way to allocate different animals in different zoo locations depending on their habitat, behavior, and other properties. The traditional approach to tackle this problem is based on a mapping onto a network problem [Refs. [1–5]], where nodes or vertices represent the population elements, the links or edges represent pairwise relations between the elements, and the edge weights account for the degree of similarity or dissimilarity between the corresponding elements.

In several population stratification problems it is clear, however, that the system under consideration is characterized by relationships involving more than two elements. For example, the property "mammal" divides the animal population into two groups: nonmammals and mammals, each containing several elements. Hypergraphs can be used to represent associations beyond pairwise relations. A hypergraph is an intuitive extension of the concept of graph or network where the edges are sets of any number of elements. For example, in an animal population, an edge can represent an association between all animals with a given property—all airborne animals, for example.

We consider hypergraphs as a suitable mathematical structure to represent a population of elements and their attributes. We introduce a statistical model on the population attributes hypergraph as a mean to solve the inverse problem, finding the population stratification given the population elements and their associations according to certain attributes. We go over technical issues associated with the framework and its application to real examples as well.

## II. HYPERGRAPH REPRESENTATION

A *hypergraph* is an intuitive extension of the concept of a graph or network where the nodes represent the systems elements and the edges (also called hyperedges) are sets of any number of elements [Fig. 1(a)]. This mathematical construction is very useful to represent a population of elements and their attributes. For example, consider the animal population in Fig. 2(a) together with their attributes: habitat, nutrition behavior, etc. In this case the hypergraph nodes represent animals. Furthermore, we can use an edge to represent the association between all animals with a given attribute: edge 1, all nonairborne animals; edge 2, all airborne animals; and so on [Fig. 2(b)].

This mapping is applicable when the attributes are given by genetic information as well. For example, consider a human population for which we know which nucleotides (represented by the letters A, C, G, and T) are present at specific chromosomes and chromosome positions. Since humans have two copies of each gene, we have two letters for each position. A scenario could be the presence of one of the letters A or G at a given position, resulting in the combinations AA, AG, and GG. When these combinations appear in a significant frequency in the population they are referred as

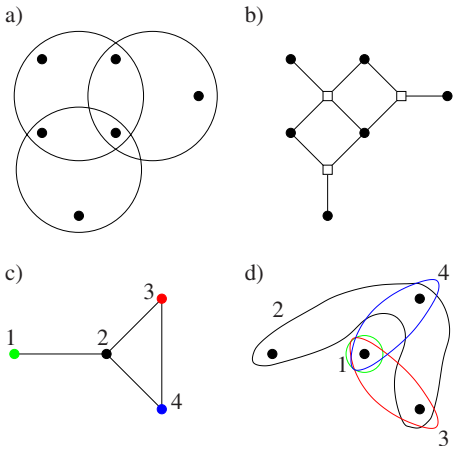


FIG. 1. (Color online) Hypergraph: (a) A hypergraph with three edges. Each edge is represented by a circle and its composed by the nodes within the circle. (b) Bipartite graph representation of the hypergraph in (a), the squares representing the hypergraph edges. (c) and (d) Nearest-neighbor mapping of a graph onto a hypergraph. The graph in (c) is mapped onto the hypergraph in (d), where each hyperedge represents a set of nearest neighbors of a node in graph (c), as indicated by the enumeration.

 

a

| | hair | feathers | eggs | milk | airborne | aquatic | predator | toothed | backbone | breathes | venomous | fins | legs | tail | domestic | catsize |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aardvark | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 4 | 0 | 0 | 1 |
| antelope | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 |
| bass | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| bear | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 4 | 0 | 0 | 1 |
| boar | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 |
| buffalo | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 |
| calf | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 1 | 1 |
| carp | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| catfish | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| cavy | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 0 | 1 | 0 |
| cheetah | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 |
| chicken | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 1 | 1 | 0 |
| chub | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| clam | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| crab | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| crayfish | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 |
| crow | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | 0 |
| deer | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 |
| dogfish | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| dolphin | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| dove | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 1 | 1 | 0 |
| duck | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | 0 |
| elephant | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 |
| flamingo | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | 1 |
| flea | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 6 | 0 | 0 | 0 |
| frog | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 4 | 0 | 0 | 0 |
| fruitbat | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | 0 |
| giraffe | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 |
| | . | | | | | | | | | | | | | | | |
| | . | | | | | | | | | | | | | | | |
| | . | | | | | | | | | | | | | | | |

b

| | |
|---|---|
| airborne-0 | aardvark antelope bass bear boar buffalo calf carp catfish cavy ... |
| airborne-1 | chicken crow dove duck flamingo fruitbat gnat gull hawk honeybee ... |
| aquatic-0 | aardvark antelope bear boar buffalo calf cavy cheetah chicken clam ... |
| aquatic-1 | bass carp catfish chub crab crayfish dogfish dolphin duck frog gull ... |
| backbone-0 | clam crab crayfish flea gnat honeybee housefly ladybird lobster moth ... |
| backbone-1 | aardvark antelope bass bear boar buffalo calf carp catfish cavy ... |
| breathes-0 | bass carp catfish chub clam crab crayfish dogfish haddock herring ... |
| breathes-1 | aardvark antelope bear boar buffalo calf cavy cheetah chicken crow ... |
| catsize-0 | bass carp catfish cavy chicken chub clam crab crayfish crow dove duck ... |
| catsize-1 | aardvark antelope bear boar buffalo calf cheetah deer dogfish dolphin ... |
| domestic-0 | aardvark antelope bass bear boar buffalo catfish cheetah chub clam ... |
| domestic-1 | calf carp cavy chicken dove goat hamster honeybee parakeet pony ... |
| eggs-0 | aardvark antelope bear boar buffalo calf cavy cheetah deer dolphin ... |
| eggs-1 | bass carp catfish chicken chub clam crab crayfish crow dogfish dove ... |
| feathers-0 | aardvark antelope bass bear boar buffalo calf carp catfish cavy ... |
| feathers-1 | chicken crow dove duck flamingo gull hawk kiwi lark ostrich parakeet ... |
| fins-0 | aardvark antelope bear boar buffalo calf cavy cheetah chicken clam ... |
| fins-1 | bass carp catfish chub dogfish dolphin haddock herring pike piranha ... |
| hair-0 | bass carp catfish chicken chub clam crab crayfish crow dogfish dolphin ... |
| hair-1 | aardvark antelope bear boar buffalo calf cavy cheetah deer elephant ... |
| legs-0 | bass carp catfish chub clam dogfish dolphin haddock herring pike ... |
| legs-2 | chicken crow dove duck flamingo fruitbat gorilla gull hawk kiwi lark ... |
| legs-4 | aardvark antelope bear boar buffalo calf cavy cheetah crab deer ... |
| legs-5 | starfish |
| legs-6 | crayfish flea gnat honeybee housefly ladybird lobster moth termite ... |
| legs-8 | octopus scorpion |
| milk-0 | bass carp catfish chicken chub clam crab crayfish crow dogfish dove ... |
| milk-1 | aardvark antelope bear boar buffalo calf cavy cheetah deer dolphin ... |
| | . |
| | . |
| | . |

c

Group 1 antelope buffalo calf cavy deer elephant fruitbat giraffe goat gorilla hamster hare oryx pony reindeer squirrel vampire vole wallaby
Group 2 aardvark bear boar cheetah leopard lion lynx mole mongoose opossum polecat puma housecat raccoon wolf
Group 3 dolphin mink platypus porpoise seal sealion
Group 4 chicken crow dove duck flamingo gull hawk kiwi lark ostrich parakeet penguin pheasant rhea skimmer skua sparrow swan vulture wren
Group 5 bass carp catfish chub dogfish haddock herring pike piranha seahorse seasnake sole stingray tuna
Group 6 frog newt pitviper slowworm toad tortoise tuatara
Group 7 flea gnat honeybee housefly ladybird moth slug termite wasp worm
Group 8 clam crab crayfish lobster octopus scorpion seawasp starfish

FIG. 2. Stratification according to phenotypic attributes: (a) A list of animals is given together with certain attributes characterizing them. The complete data set is available from Ref. [6]. Except for the attribute "legs," 1 and 0 indicate possession or not, respectively, of the corresponding attribute. The problem consists in determining the optimal stratification of the animal population based on the provided attributes. (b) Hypergraph representing the zoo data. Each line corresponds with an edge, whose elements are specified within the right column. (c) ML stratification for the case of eight groups.

a single nucleotide polymorphism (SNP). This genetic information can be mapped onto a hypergraph. The vertices in the hypergraph represent individuals, and the edges now represent groups of individuals with the same genetic information at a given position: edge 1, all individuals with call AA for SNP 1; edge 2, all individuals with call AG for SNP 1; and so on (Fig. 3).

As we mention above, hypergraphs are a generalization of graphs to allow for connections beyond two elements, graphs being the particular case where edges contain only two elements. Yet it is worth mentioning that we could imagine other mappings between a graph and a hypergraph. For example, let the vertices of the graph be the vertices of the hypergraph as well, but now let the hyperedges be the sets of nearest neighbors of the vertices [Figs. 1(c) and 1(d)]. As shown in the next section, this mapping allows us to make a connection with the clustering algorithm on graphs introduced in Ref. [4].

## III. STATISTICAL MODEL

After identifying hypergraphs as a suitable mathematical structure to represent a population and their attributes we focus on determining how to use this information to solve the inverse problem, finding the population stratification given the population elements and their associations according to certain attributes. Our working hypothesis is that (i) the population is divided in groups and (ii) the elements of each group are characterized by a different combination of attributes. The latter do not exclude the possibility that two groups exhibit one same attribute, being different according to others. These hypotheses are the bases for the following statistical model on hypergraphs.

*Data.* Consider a population of $n$ individuals and a hypergraph with $m$ edges characterizing the relationships among them. The hypergraph can be specified, for example, using the adjacency matrix $a$, where $a_{ij}=1$ if element $i$ belongs to edge $j$ and it is zero otherwise.

a

| | rs1380576 | rs12039365 | rs4951393 | rs11586387 | rs12041243 | rs3789052 | rs3789051 | rs4252675 | rs4252685 | rs4252686 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | CC | AA | CC | AT | GG | GG | TT | CC | AA | AG |
| 1 | CG | AA | AC | AA | AA | AG | AG | TT | AC | AG |
| 2 | GG | AA | AA | AA | AA | AA | AA | TT | AA | GG |
| 3 | CC | AA | CC | AT | AA | GG | GG | TT | CC | AA |
| 4 | CG | AG | AC | AT | GG | GG | TT | AC | AG | AA |
| 5 | GG | AA | AA | AA | AA | AA | AA | TT | AA | GG |
| 6 | GG | AA | AA | AA | AA | AA | AA | TT | AA | GG |
| 7 | CC | AA | CC | AT | AA | GG | GG | TT | CC | AA |
| 8 | GG | AA | AA | AA | AA | AA | AA | TT | AA | GG |
| 9 | GG | AG | AA | AA | AG | AG | AG | TT | AA | GG |
| 10 | GG | AG | AA | AA | AG | AG | AG | TT | AA | GG |
| 11 | CG | AA | AC | AT | AA | AG | AG | TT | AC | AG |
| 12 | GG | AG | AA | AA | AG | AG | AG | TT | AA | GG |
| 13 | GG | AA | AA | AA | AA | AA | AA | TT | AA | GG |
| 14 | GG | AA | AA | AA | AA | AA | AA | TT | AA | GG |

b

| | |
|---|---|
| rs1380576-CC | 0 3 7 24 25 27 32 33 ... |
| rs1380576-CG | 1 4 11 18 19 21 22 ... |
| rs1380576-GG | 2 5 6 8 9 10 12 13 ... |
| rs12039365-AA | 0 1 2 3 5 6 7 8 11 ... |
| rs12039365-AG | 4 9 10 12 16 17 21 ... |
| rs12039365-GG | 20 56 59 |
| rs4951393-AA | 2 5 6 8 9 10 12 13 ... |
| rs4951393-AC | 1 4 11 18 19 21 22 ... |
| rs4951393-CC | 0 3 7 25 27 33 37 50 ... |
| rs11586387-AA | 1 2 5 6 8 9 10 12 13 ... |
| rs11586387-AT | 0 3 4 7 11 24 34 |
| rs12041243-AA | 1 2 3 5 6 7 8 11 13 ... |
| rs12041243-AG | 9 10 12 16 17 21 28 ... |
| rs12041243-GG | 20 56 59 |
| rs3789052-AA | 2 5 6 8 13 14 23 26 ... |

FIG. 3. Mapping genotypic information into a hypergraph: (a) A population of individuals, labeled by $1,2,3,\ldots$, is given together with their genotype for specific DNA positions within. These positions have been selected because they exhibit significant variation across the human population, referred to as single nucleotide polymorphisms (SNPs), and are labeled using the standard SNP notation: rsNUMBER. The letters A, C, G, and T represent nucleotides, and two letters are reported because each DNA position appears in two different chromosome copies. (b) Hypergraph representing the genotypic data. Each line corresponds with an edge, whose elements are specified within the right column.

*Model.* The population is divided into $n_g$ groups and let $g_i$, $i=1,\ldots,n$, denote the group to which node $i$ belongs. With probability $\theta_{ij}$ an element of group $i$ belongs to edge $j$.

*Likelihood.* The likelihood to observe the data given this model is

$$P(a|g,\theta) = \prod_{i=1}^{n}\prod_{j=1}^{m} \theta_{g_i j}^{a_{ij}}(1-\theta_{g_i j})^{1-a_{ij}}. \quad (1)$$

In essence the likelihood (1) is a mathematical representation of our intuition about the observation of the hypergraph given a population stratification; i.e., elements of the same group have the same probability to exhibit a certain attribute and thus to belong to the edge representing that attribute. In the following we discuss how to determine the best choice of model parameters $(g,\theta)$ and $n_g$.

The likelihood (1) resembles that introduced in Ref. [4] in the context of finding communities on graphs. Despite the similarity and being a source of inspiration, they are quite different in their interpretation. A hypergraph can be indeed represented by a bipartite graph, with one type of nodes corresponding to the hypergraph nodes and another representing the hypergraph edges [Fig. 1(b)]. In this work we focus, however, on clustering the original hypergraph nodes alone. Therefore, the likelihood in Eq. (1) represents a statistical model on a hypergraph. In contrast, a true statistical model on a bipartite graph should attend to cluster both types of nodes—the original hypergraph nodes and the attribute nodes. There are other technical differences. Here we model the stratification encoded in $g$ as parameters, while they were modeled as hidden variables in [4]. Hence, although similar in form, the likelihood in Eq. (1) is different from that in Ref. [4].

The connection between the statistical model on graphs introduced in Ref. [4] and Eq. (1) is established by the following mapping between a graph and a hypergraph: Let the vertices of the graph be the vertices of the hypergraph as well, but now let the hyperedges be the sets of nearest neighbors of the vertices [Figs. 1(c) and 1(d)]. For this particular hypergraph, the likelihood in Eq. (1) coincides with that introduced in Ref. [4]. Yet the likelihood in Eq. (1) can be applied to a larger spectrum of problems.

## IV. MAXIMUM-LIKELIHOOD STRATIFICATION

The model defined above belongs to the class of finite mixture models [2]. In general a mixture model assumes that the observed variables come from a mixture of distributions. In our case the observed variables are the adjacency matrix elements $a_{ij}$, representing the participation or not of vertex $i$ on the edge $j$, and the mixture comes from the subdivision of vertices in groups. Because the adjacency matrix elements can take the values 0 and 1 only, we have a mixture model on Boolean variables [see Eq. (1)]. Thus, we can obtain the optimal stratification using techniques applicable to finite mixture models in general. In particular, we use the well-established expectation maximization (EM) algorithm [7] to determine the maximum-likelihood (ML) stratification given a fixed number of groups.

*ML stratification.* First, we compute the expectation of the log-likelihood $\mathcal{L}=\ln P(a|g,\theta)$ with respect to the probability $q_{ir}$ that element $i$ belongs to group $r$, obtaining

$$E[\mathcal{L}] = \sum_{i=1}^{n}\sum_{r=1}^{n_g}\sum_{j=1}^{m} q_{ir}[a_{ij}\ln\ \theta_{rj} + (1-a_{ij})\ln(1-\theta_{rj})]. \quad (2)$$

Second, we compute the parameters $\theta$ that maximize (2), resulting in

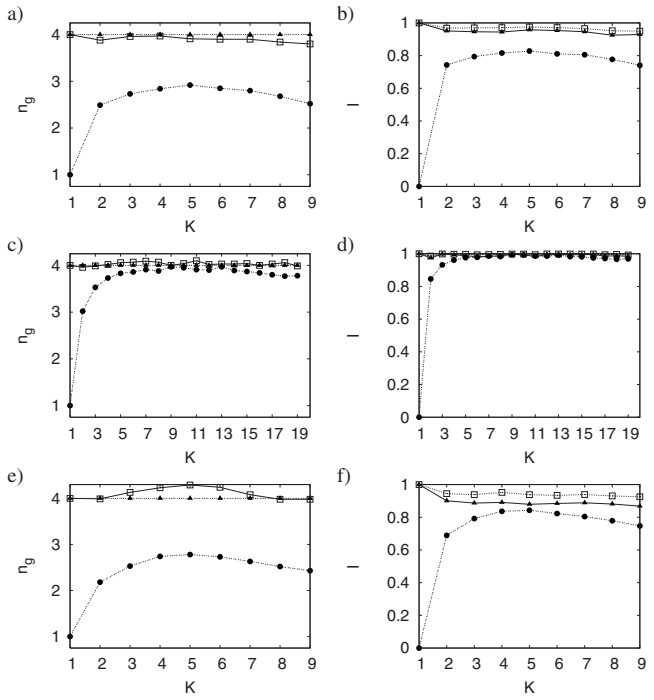


FIG. 4. Test example: The best choice of $n_g$ and normalized mutual information $I$ between the predicted population optimal stratification and the original stratification as a function of degree $K$. These results are obtained computing the optimal stratification for $n_g=1,\ldots,20$ using the EM algorithm with one initial condition. The optimal $n_g$ was obtained using the AIC (solid circles), the representativeness criterion (open squares), and assuming it equal to 4 (solid triangles). In (a) and (b) the case study hypergraphs have $n=100$ nodes divided into four groups of equal size and $m=10$ edges, while in (c) and (d) the number of edges is doubled to $m=20$. In (e) and (f) the case study hypergraphs are similar to those in (a) and (b) except that the groups have different sizes: 40, 30, 20, and 10 nodes.

$$\theta_{rj} = \frac{\sum_{i=1}^{n} q_{ir} a_{ij}}{\sum_{i=1}^{n} q_{ir}}. \tag{3}$$

Finally, $q$ is estimated using

$$q_{ir} = \frac{P(a|g, g_i = r, \theta)}{\sum_{s=1}^{n_g} P(a|g, g_i = s, \theta)}. \tag{4}$$

Starting from an initial condition, we iterate Eqs. (3) and (4) until the change of all $q$ elements is smaller than a predefined precision. The EM algorithm always converges to a local maximum of the likelihood, which may or may not coincide with the global maximum. One approach to explore different local maxima, in case they exist, consists in generating different initial conditions [2]. Here we explore different initial conditions by assigning to the $q$ elements the random initial values

$$q_{ir} = \frac{x_{ir}}{\sum_{s=1}^{n_g} x_{is}}, \tag{5}$$

where $x_{ir}$ is a random number between 0 and 1. Putting all together, starting from each initial condition, we iterated Eqs. (3) and (4) until the change of all $q$ elements is smaller than $10^{-6}$.

## V. BEST CHOICE OF $n_g$

A more subtle issue is to determine the optimal number of groups. The standard approach to solve this problem is based on the Occam's razor principle: provided different models describing the reality with similar accuracies we should select the simplest. In other words, we accept an increase in model complexity only provided we obtain a significantly better description accuracy or predictive power. We use the Akaike information theoretical criterion (AIC) [8] to quantify the model complexity. According to this criterion, the complexity of a model is determined by the number of independent parameters and the best choice of $n_g$ is the one minimizing

$$\chi_{\mathrm{AIC}}(ng) = -\max_{\{g,\theta\}} \mathcal{L} + (n+m)(n_g - 1), \tag{6}$$

where $(n+m)*(n_g-1)$ is the number of independent parameters in our statistical model. The first term on the right-hand side of Eq. (6) quantifies the goodness of the fit and it decreases with increasing $n_g$. On the other hand, the second term on the right-hand quantifies the model complexity and increases with increasing $n_g$. The optimal choice of $n_g$ results from the balance between these two opposite contributions.

It becomes clear below that the AIC criterion can result in too conservative estimates of $n_g$, forcing us to consider a different approach. Instead of focusing on model complexity, we ask the following question: given the ensemble of all models with different $n_g$, which is the most representative among them? To be more precise we need a measure to compare the degree of similarity between two different population stratifications $S_i$ and $S_j$, corresponding to models with $i$ and $j$ groups, respectively. We consider the normalized mutual information [3]

$$I(S_i, S_j) = \frac{-2\sum_{k=1}^{n_g^{(i)}} \sum_{l=1}^{n_g^{(j)}} \rho_{kl}^{(i,j)} \ln \rho_{kl}^{(i,j)} / \rho_k^{(i)} \rho_l^{(j)}}{\sum_{k=1}^{n_g^{(i)}} \rho_k^{(i)} \ln \rho_k^{(i)} + \sum_{l=1}^{n_g^{(j)}} \rho_l^{(j)} \ln \rho_l^{(j)}}, \tag{7}$$

where

$$\rho_{kl}^{(i,j)} = \frac{1}{n} \sum_{s=1}^{n} q_{sk}^{(i)} q_{sl}^{(j)}, \tag{8}$$

$$\rho_k^{(i)} = \frac{1}{n} \sum_{s=1}^{n} q_{sk}^{(i)}. \tag{9}$$

The normalized mutual information equals 0 when the stratification $S_i$ does not contain any information about the stratification $S_j$, becomes 1 when the two stratifications are iden-

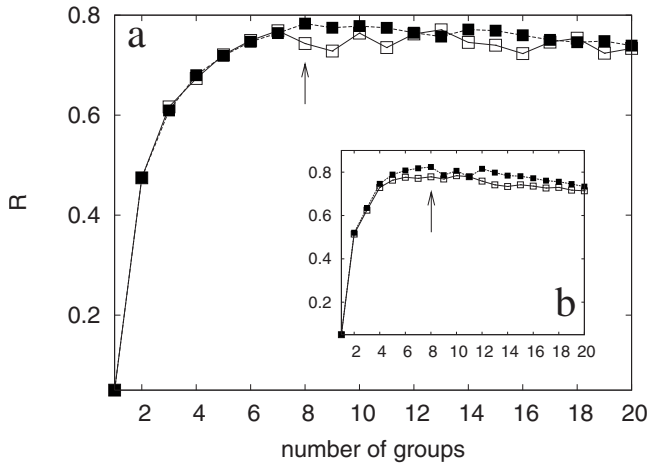


FIG. 5. Representativeness plot: Representativeness as a function of the number of groups for the (a) zoo and (b) MDM4 problems. Different symbols indicate different numerical accuracies of the numerical algorithm to find the ML stratification. The arrow indicates the number of groups maximizing the representativeness. The different symbols indicate different numbers of initial conditions for the EM algorithm, from 100 (open symbols) to 10 000 (solid symbols).

tical, and interpolates between 0 and 1 for intermediate scenarios.

For each stratification $S_i$ we define stratification *representativeness*

$$R(S_i) = \frac{\sum_j I(S_i,S_j)}{\sum_j 1}, \tag{10}$$

the average of the normalized mutual information of all stratifications $S_j$ with respect to a given stratification $S_i$. The larger $R(S_i)$ is, the more the stratification $S_i$ represents the stratification ensemble. Furthermore, we define the most representative stratification among an ensemble of stratifications as the stratification maximizing $R$. In case there is more than one stratification satisfying this criteria we invoke the Occam's razor principle and select the one with the lowest number of groups.

## VI. TEST EXAMPLES

To test the population stratification framework introduced above we need hypergraph examples for which the stratification is already known. The statistical model defined by Eq. (1) provides a straightforward method to generate an ensemble of hypergraphs. Indeed, provided $g$ and $\theta$ we can generate realizations of the hypergraph adjacency matrix using Eq. (1). We consider the following ensemble of hypergraphs with $n$ nodes and $m$ edges: (i) The population is divided into $n_g$ groups of equal size. (ii) All nodes have the same degree $K$, where the *degree* is the number of edges to which a node belongs. (iii) The edges to which the elements of a given group belong are selected at random among the $m$ edges, making it such that every pair of groups differs in at least one edge. Provided $m > n_g$, the latter is possible only for $1 \leq K \leq m-1$, defining our working range for $K$.

Using this hypergraph ensemble, we generate hypergraphs with $n$ nodes, $m$ edges, and degree $K$. For each hypergraph we determine the best choice of $n_g$ and the corresponding population stratification, using both the AIC and representativeness criteria. To compare the predicted optimal stratification and the original subdivision of the population we use the normalized mutual information (7) [3]. Finally, the results are averaged over 100 hypergraphs for each set of $(n,m,m)$.

Figure 4 shows the results for $n=100$, $m=10$, and $m=20$ as a function of degree $K$. When we fix, *a priori*, the number of groups to 4, the stratification method based on Eq. (1) almost finds the right subdivision. Indeed, the normalized mutual information between the predicted stratification in four groups and the original subdivision is very close to 1, indicating that most nodes have been allocated to their original groups [solid triangles in Figs. 4(b) and 4(d)]. While this observation does not exclude the existence of hypergraph instances where the method can fail, it supports its use in real cases.

Next we test the best choice of $n_g$ when it is not known *a priori*. For $m=10$ edges the AIC underestimates $n_g$, particularly for small $K$ [Fig. 4(a)]. Consequently, the normalized mutual information between the predicted and original subdivision of the population is quite small [Fig. 4(b)]. This disagreement persists for $m=20$ and small values of $K$, but gets significantly improved for $K$ larger than 4 [Figs. 4(c) and 4(d)]. In contrast, the representativeness criterion performs quite well for all the tested parameter combinations. On average, it predicts the right number of groups, 4 [Fig. 4(a)] and the normalized mutual information is very close to 1 [Fig. 4(b)]. Taken together, these results indicate that the representativeness criterion performs as well if not better than the AIC. Hence, in the following we restrict ourselves to the former approach to select the best choice of $n_g$.

To test the performance of the clustering algorithm in the context of variable group sizes, we consider the case where the population is divided as before into $n_g$ groups, but now each group contains a variable number of vertices. We obtain similar results as for the case of equal group sizes [Figs. 4(e) and 4(f)], indicating that the representativeness approach works when groups have variable sizes as well.

## VII. REAL EXAMPLES

Now we proceed to apply the population stratification framework to real examples. The first example is the zoo problem [Fig. 2(a)], requiring us to group different animals according to their habitat, nutrition behavior, and other properties [Fig. 2(a)]. In this case the hypergraph nodes represent animals and each edge represents an association between animals exhibiting a given phenotypic attribute [e.g., edge 1, all nonairborne animals; edge 2, all airborne animals, Fig. 2(b)].

Figure 2(c) shows the animal stratification for the zoo problem for the case of eight groups. A quick inspection shows that elements within the same group have indeed the sense of a group. The first three groups contain all mammals subdivided by their habitat and feeding behavior. The remaining groups represent birds, fishes, amphibian-reptiles,

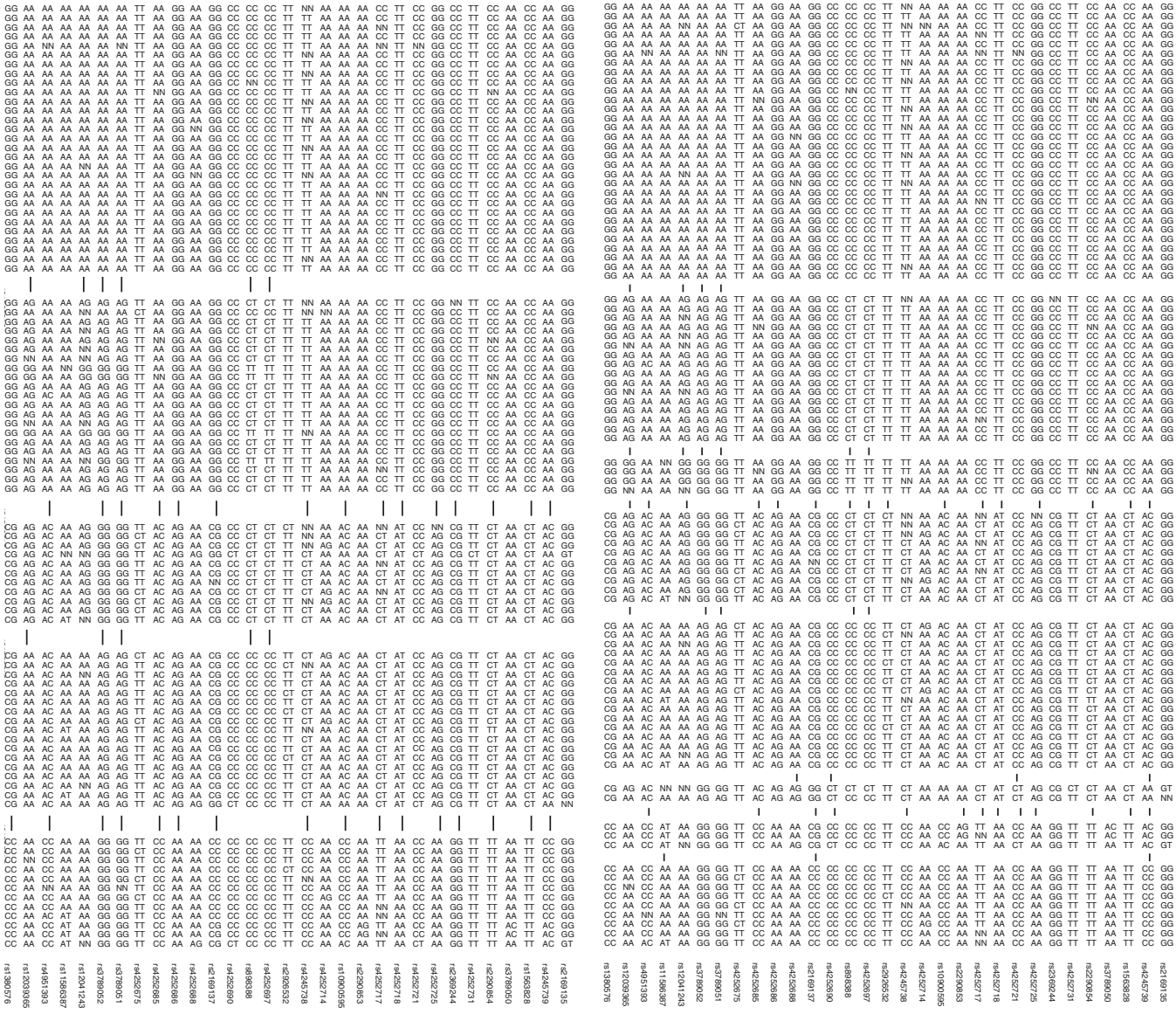

FIG. 6. Stratification according to genotypic attributes: A population of 90 Caucasians is studied, focusing on SNPs within the MDM4 gene, as reported by the HapMap project [9]. SNPs with no variation within this particular subpopulation have been excluded. A, C, G, and T represent the different nucleotides, and NN represents data that are not available. The specific SNPs under consideration are indicated by the bottom labels, using the standard SNP notation. The figure shows the ML stratification for the case of five (left) and eight (right) groups, the latter corresponding with the best choice of $n_g$ according to the representativeness criterion [Fig. 5(b)]. The vertical lines in between indicate the SNPs at which the adjacent groups differ significantly.

terrestrial arthropods, and aquatic arthropods (except the scorpion), in that order. A similar stratification is obtained for the case of seven groups, except for groups 1 and 3, which are merged into one group. On the other hand, a stratification into nine groups further splits the birds into two groups.

Figure 5(a) shows the representativeness as a function of the number of groups for the zoo problem. For a small number of groups $R$ increases monotonically with increasing the number of groups, saturating to an approximate plateau at large group numbers. In the latter region, there are small variations determined by the numerical accuracy of the algorithm computing the ML stratification for a fixed number of groups. A model with eight groups provide the highest degree of representativeness [Fig. 2(c)]. Once again, a quick inspection is sufficient to realize that, indeed, this represents a natural subdivision of the animal population.

The second real example concerns stratification according to genetic information. It consists of a population of 90 Caucasians and the genotype at different SNPs within the MDM4 gene, as reported by the HapMap project [9]. The MDM4 gene plays a key role in the p53 stress response pathway, and genetic variations within this gene could potentially result in different predispositions to cancer and/or response to cancer drug therapy [10]. Focusing on SNPs with variation among this particular subpopulation, we stratify its elements using the method described above. Figure 5(b) shows the representativeness of the ML stratification as a function of the number of groups. As for the zoo problem,

the representativeness increases monotonically for a small number of groups and saturates to a plateau with some variations determined by the numerical accuracy. At five groups we already observe a high degree of representation and eight groups represent the best choice of $n_g$ according to the representativeness criterion.

The genetic information for all individuals is shown in Fig. 6 stratified into five and eight groups, the latter corresponding with the highest representativeness stratification. The top and bottom groups are almost entirely homozygous (same letter) at every position. In contrast, all the intermediate groups are heterozygous (different letter) at several positions, which do not overlap between them in at least one position. A visual inspection of both stratifications indicates that they are very similar, as anticipated by the close values of representativeness between five and eight groups [Fig. 5(b)].

## VIII. DISCUSSION

The mapping of either phenotypic or genetic information onto a hypergraph offers significant advantages over the current reductionist mapping of the stratification problem onto a network problem. First, the hypergraph contains all the information provided by the original data. Second, it allows us to introduce an intuitive statistical model for the observed phenotypic and genotypic variations based on a postulated population stratification and the tendency of individuals within a group to exhibit certain phenotypic and genotypic features. Finally, the generalization to problems dealing with both phenotypic and genotypic variation is straightforward, after introducing a hypergraph with two edge types.

The representativeness measure introduced here can be used as an alternative to model complexity when selecting the optimal number of groups given the available information. It is based on the interpretation of statistical significance in terms of information content, a philosophy with increasing recognition among the statistical modeling community [3,11]. This measure allows us to focus our analysis on a stratification obtained for a characteristic number of groups, with a high information content about stratifications with a different number of groups. Indeed, we believe that the representativeness approach works because it chooses the clustering with best consensus agreement among an ensemble of clusterings with a variable number of groups.

Hypergraph partitioning has been already studied with applications to numerical linear algebra and logic circuit design [12]. The focus has been, however, on balance clustering, which imposes stratifications on groups of similar size. In contrast, the framework developed here is more suitable to determine a natural partition of the population (or the hypergraph representing it), potentially resulting in clusters of different sizes [see Fig. 1(c), for example]. It is worth noticing that our framework can be adapted to balance clustering as well, after adding the constraint that all groups have the same size to the starting statistical model.

Finally, it is worth mentioning that the present method is not the only approach one can follow to stratify the population. As is known from the data clustering literature, there is often more than one method to cluster some elements based on some attributes. Particularly, in the context of genotypic information one can also introduce a distance metric and then build a statistical model for this new variable [13].

[1] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).
[2] G. McLachlan and D. Peel, *Finite Mixture Models* (Wiley, New York, 2000).
[3] A. Fred and A. Jain,*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, New York, 2003), Vol. II, pp. 128–133.
[4] M. Newman and E. Leicht, Proc. Natl. Acad. Sci. U.S.A. **104**, 9564 (2007).
[5] B. Frey and D. Dueck, Science **315**, 972 (2007).
[6] A. Asuncion and D. Newman, http://www.ics.uci.edu/~mlearn/MLRepository.html
[7] A. Dempster, N. Laird, and D. Rubin, J. R. Stat. Soc. Ser. B (Stat. Methodol.) **39**, 1 (1977).
[8] H. Akaike, IEEE Trans. Autom. Control **19**, 716 (1974).
[9] The International HapMap Consortium, Nature (London) **426**, 798 (2003).
[10] S. L. Harris and A. J. Levine, Oncogene **24**, 2899 (2005).
[11] N. Slonim, G. Atwal, G. Tkacik, and W. Bialek, Proc. Natl. Acad. Sci. U.S.A. **102**, 18297 (2005).
[12] D. Papa and I. Markov, *Approximation Algorithms and Metaheuristics* (CRC Press, Boca Raton, FL, 2007), pp. 1–19.
[13] J. K. Pritchard, M. Stephens, and P. Donnelly, Genetics **155**, 945 (2000).