*

Separation of overlapping subpopulations by mutual information

Sean O'Rourke* seano@cs.ucsd.edu
Gal Chechik† gal@stanford.edu
Eleazar Eskin* eeskin@cs.ucsd.edu

Abstract

Identifying ancestral sequences is an important first step in understanding population history and dynamics. However, several interesting cases including human genetic variation feature highly overlapping subpopulations, whose ancestral subpopulations differ by fewer than the average number of mutations between members of the same population. In such cases both flat and hierarchical distance-based clustering methods optimize the wrong objective function, and will therefore fail to recover the correct population structure. We present an algorithm for identifying substructure in overlapping populations based on the observation that variations at different sequence positions are uncorrelated. While motivated by a particular biological application, the algorithm can be applied to both discrete and continuous clustering problems for which distance-based measures are inappropriate. We demonstrate the algorithm's effectiveness in reconstructing human subpopulations.

1 Introduction

Many biological processes can be characterized as forms of sequence evolution, in which a set of individual sequences are generated by a process of alternating replication and random mutation. When different sequences replicate at different rates, such data can be modeled as a number of ancestral sequences, each generating offspring by mutation. One well-studied problem, *population substructure*, aims to recover the underlying ancestral sequences from the entire population and to cluster individuals by ancestral sequence. Examples of this problems include determining human population substructure[6] and reconstructing Alu repeat insertion history[7].

If the sequences are long relative to their mutation rate, the probability of the same mutation occurring independently in unrelated individuals is vanishingly small. Viewing subpopulations as clusters of points in sequence space, the mutation distance between points within a single cluster will be small relative to that between cluster means, and the substructure problem will be amenable to distance-based clustering. However, if the sequences are short relative to the mutation rate, inter- and intra-population mutation distances are on the same

*Department of Computer Science and Engineering, University of California San Diego
†Department of Computer Science, Stanford University
*Thanks to Noah Zaitlen and Robin Friedman for providing and preparing SNP data.

order, and the populations largely overlap. In this situation distance-based clustering algorithms (such as agglomerative linkage clustering) perform poorly, since an individual may often be closer to another population's ancestral sequence than to its own, and the resulting clusters will reflect local irregularities instead of the actual generating process[7].

But since mutations are independent in our model, correlation between two or more positions therefore suggests that a population was actually generated by two or more ancestral sequences differing at those positions, since deviations from a subpopulation's ancestral sequence at different positions should be uncorrelated. For example, consider a population of 10 instances each of the sequences "0...1" and "1...0", where "..." represents a large number of unrelated intervening positions. Looking only at individual positions, all 20 sequences appear to form a single group with a single ancestral sequence and per-position mutation rate 0.5 on either end. Mutations in the intervening sequence will make the two halves indistinguishable by Hamming distance. However, under the assumption of uncorrelated mutations, the probability of observing two such correlated positions in a single population is vanishingly small. By dividing the population into two subpopulations with ancestral sequences "0...1" and "1...0", we eliminate this correlation and increase the population's likelihood.

Here we propose a top-down clustering algorithm for reconstructing population substructure for evolutionary processes with similar mutation and replication rates. Our algorithm, sequential shape-constrained clustering (SSCC), finds clusters that maximize information about sequence identity while minimizing mutual information between positions within each cluster. SSCC extends standard top-down clustering with a penalty for multi-information between positions within a subpopulation. We present here results for discrete sequence clustering with a specific penalty function, but SSCC is more general, and can be applied with other penalty functions to both discrete and continuous multivariate data.

Section 2 describes the formulation of the problem and the objective function, and section 3 presents an efficient sequential iterative algorithm to minimize this objective. In section 4 we illustrate the objective's behavior on artificial data. We then apply SSCC to recover population substructures from human single nucleotide polymorphisms (SNPs) and Alu repeat elements, comparing the results to clusters obtained both by K-means and by state-of-the-art population substructure methods. In particular, we show that our method is competitive with, and in some cases outperforms, the STRUCTURE program [8], one of the most widely-used tools for this problem. Finally in section 5 we suggest several extensions and briefly describe other problems for which our algorithm is well-suited.

2 Formulation

Let $X = (X_1, \ldots, X_l)$ be a vector random variable over the m-letter alphabet $\Sigma = \{a_1 \ldots a_m\}$. Each instance x of X is a vector of length l, and we model the process that generates it as follows. First, one of n subpopulations $C \in \{1, \ldots, n\}$ is drawn with probability $p(C = c)$, and x is assigned the consensus sequence of the subpopulation $\mu_c = (\mu_{c1}, \ldots, \mu_{cl}) \in \Sigma^l$. Then the sequence is mutated independently at each position according to a subpopulation specific mutation rate $\sigma_c = (\sigma_{c1} \ldots \sigma_{cl})$, which determines the probability of mutating from the consensus value μ_{cj} to a_k. Given subpopulation-specific consensus sequences and mutation rates (μ_i, σ_i), one can calculate the probability $p(x|c)$ that a given sequence x comes from a subpopulation c. The pair (μ_i, σ_i) can be regarded as the discrete analogue of the mean and (spherical) covariance matrix in a Gaussian mixture model. The goal of subpopulation identification is to find subpopulations of the sequence set, such that sequences in each subpopulation are similar. Instead of estimating the parameters μ and σ directly, we choose here to use a nonparametric model and measure within-subpopulation similarity by the homogeneity of the conditional entropy $H(X|C)$ [5], as compared with the overall homogeneity $H(X)$. This reduces to using the mutual

information $I(X;C) = H(X) - H(X|C)$ as a measure for the goodness of dividing sequences into subpopulations.

Since mutations occur independently, sequences x drawn from the same subpopulation c have $p(x_i|\{x\}_{j\neq i}, C = c) = p(x_i|C = c)$, where $\{x\}_{j\neq i}$ is any set of positions excluding the position i. As a result of this independence, the conditional mutual information $I(X_i;X_j|C)$ for every pair of positions X_i, X_j within a cluster should vanish. For subsets of higher order, the independence of all l variables can be quantified by an extension of the mutual information, the *multi-information* $MI(X_1;\ldots;X_l) \stackrel{\text{def}}{=} D_{KL}(p(x_1,\ldots,x_l)||p(x_1)\ldots p(x_l))$ [10], which is again non-negative and vanishes if and only if all X_i are independent.

We propose using the additional knowledge about conditional independence within subpopulations to devise a better clustering procedure for finding subpopulations when many clusters are overlapping. For such data, distance-based clustering methods like K-means and Gaussian mixtures tend to find wrong clusters that are more concentrated in space, but have non-independent positions. This happens because when the distances between consensus vectors μ_i, μ_j are small relative to the mutation rates σ_i, σ_j the probability of cluster c_i generating a sequence x such that $p(x|c_{j\neq i}) > p(x|c_i)$ becomes significant.

The goal of subpopulation identification is therefore to find subpopulations of the sequences, such that sequences in each subpopulation are similar, but positions are independent. This can be formally quantified by maximizing the mutual information $I(X;C)$ while removing the position dependence $MI(X_1;\ldots;X_l|C)$, $\min_{\mu,\sigma} S(C) = \beta\ MI(X_1;\ldots;X_l|C) - I(X;C)$, where β controls the trade-off between the two objective factors. Since with limited data it is often difficult to estimate the high-dimensional joint distribution $p(X_1,\ldots,X_l)$ needed for estimating the multi-information, one can replace the above independence condition with a weaker, pairwise one

$$\min_{\mu,\sigma} S(C) = \beta \sum_{i<j} I(X_i;X_j|C) - I(X;C) \tag{1}$$

The two conditions become equivalent up to constants for Bethe-type distributions, where no high-order correlations above second order exist.

Our method is related to previous models of clustering with side information. Shental *et al.* [9] showed how to use equivalence (and non-equivalence) constraints to learn the components of the within-cluster covariance matrix, so these can be normalized out when learning a Gaussian mixture model. Our problems differ from those discussed in these two papers in several aspects: First, rather than learning a whitening transformation from the data, we enforce spherical clusters. Second, we operate in a discrete space over a given alphabet rater than a continuous space. Finally, rather than focusing on second order relations (covariance matrices), we force high-order independencies using the mutual information measure. Clustering categorical variables while preserving information about another variable was addressed in the Information Bottleneck (IB) framework [11]. There, information preservation $I(X;C)$ is traded for a compression term that determines the loss of information due to clustering. Adapting their notation to the one used here, the IB functional is $\min_{p(c|z)} I(Z;C) - \beta I(C;X)$, where Z is a random variable that holds the identity of the sequence, rather than its value. Chechik *et al.* [3] discussed clustering while preserving information about one variable X^+ but removing information about a second variable X^-. In the limit of infinite β, this is formalized as finding $\min I(C;X^-) - \gamma I(C;X^+)$. The method described here aims to cluster while preserving information about the sequences $I(C;X)$, but at the same time removing conditional information per cluster, that is, finding $\min I(X|C) - \gamma I(C;X)$.

3 Iterative algorithm

We find local optima of $S(C)$ using a two-phase iterative algorithm combining sequential updates and top-down splitting. The first phase splits each existing cluster c if doing so improves $S(C)$. Since the number of possible splits is exponential in cluster size, we perform a greedy step, and choose to split the cluster into two families that differ at a single position i of the sequence. We choose the position which reduces the dependencies the most, namely, the position $i = \text{argmax}_i \sum_{j \neq i} I(X_j; X_i | C = c)$. We then divide c into those sequences with the most-frequent value at position i, and those with all other values. In practice, we find that this criterion yields good splits at a small computation time cost. The second phase iterates through the clusters, sequentially updating each of their sequences' cluster assignments until no more improvement is possible. Since each step in the algorithm decreases the objective function, the cluster distributions are determined by hard assignments of a finite number of instances, and equation (1) is bounded below by $-I(X;C)$, the two-phase updates must converge.

Since the algorithm performs a large number of sequential updates, computing the score of a category when considering reassigning an instance becomes a major limiting factor. We show here how the value of $S(C)$ can be updated efficiently after adding or removing a single instance. Consider the case where we add a single instance $x = (x_1, \ldots, x_l)$ to a category c with a distribution $p(X|c)$. This creates a new category c' with probability $p(c') = (p(c) + 1/|X|)$ and sequence distribution $p(X|c') = (1-\alpha)p(X|c) + \alpha[X = x]$, where $\alpha = 1/(|c|+1)$ and $[X = x]$ is an indicator function. Separating the $X_i = x_i$ terms from the rest and summing over sequence positions i yields

$$I(X;c') = I(X;c) - I(x;c) + I(x;c') + p(c) \log \bar{\alpha} \sum_{i=1}^{l} (1 - p(x_i|c)) \tag{2}$$

where $\bar{\alpha} = 1 - \alpha$, $I(x;c) \stackrel{\text{def}}{=} \sum_i p(x_i, c) \log \left(p(x_i|c)/p(x_i) \right)$ is the symbol-specific information that the category c provides about the specific sequence x, and $I(X;c) \stackrel{\text{def}}{=} \sum_x p(x, c) \log \left(p(x|c)/p(x) \right)$ is the category-specific information that the category c provides about the sequences X.

A similar transformation can be applied to the second term in equation (1). Using $I(X_i; X_j | C = c) = H(X_i | C = c) + H(X_j | C = c) - H(X_i, X_j | C = c)$, the above substitution yields

$$\begin{aligned} H(X_i | C = c') &= \bar{\alpha} H(X_i | C = c) - (1 - p(x_i|c)) \bar{\alpha} \log \bar{\alpha} \\ &\quad + \bar{\alpha} p(x_i|c) \log p(x_i|c) - p(x_i|c') \log p(x_i|c') \qquad (3) \\ H(X_i, X_j | C = c') &= \bar{\alpha} H(X_i, X_j | C = c) - (1 - p(x_i, x_j|c)) \bar{\alpha} \log \bar{\alpha} \\ &\quad + \bar{\alpha} p(x_i, x_j|c) \log p(x_i, x_j|c) - p(x_i, x_j|c') \log p(x_i, x_j|c') \qquad (4) \end{aligned}$$

where $p(x_i, x_j|c') = \bar{\alpha} p(x_i, x_j|c) + \alpha$. Note that since both of these terms depend only upon the previous values of $H(X_i; X_j | C = c)$, $H(X_i; c)$, and $I(X; c)$, and upon the probabilities of the instance s being considered, the score update can be computed in $O(l^2)$ time rather than $O(l^2(|X| + |\Sigma|^2))$.

4 Results

To elucidate SSCC's behavior, we generated three 50-character bit-strings with relative Hamming distances of 2, 3, and 5, then generated 20 copies of each with mutation rate 0.05, yielding a 60-member population with expected Hamming distance of 2.5 between a

mutant and its ancestral sequence. When the stopping criterion is chosen to recover three clusters, SSCC recovers the ancestral sequences exactly, and assigns 93% of sequences to their generating clusters. On the other hand, iterative K-means, even when initialized with the correct parent sequences, typically converges to a degenerate solution in four or fewer iterations, demonstrating the inappropriateness of distance-based clustering for overlapping populations.

Note that because the clusters overlap, the most probable assignment of points to clusters may not be the one by which the data were generated. Because of this, and because we want to recover population substructure, recovery of ancestral sequences or cluster means is often a more appropriate performance measure than correct sequence labeling. By this measure both STRUCTURE and SSCC perform perfectly on this generated dataset. However, when the mutation rate is increased to 0.16, STRUCTURE finds a degenerate solution, while SSCC still recovers the correct ancestral sequences despite only correctly classifying 71% of individuals.

4.1 Human population substructure

Genetic association studies are heralded as a powerful tool for discovering the genetic basis of human disease [2]. These studies analyze the genetic sequences of sets of healthy and diseased individuals to identify sequence variation associated with one or the other. Association studies assume that all subjects are from the same population and, when performed on populations with substructure, may produce many spurious associations. When one subpopulation has a higher incidence of the disease, any variation specific to that population will appear to cause the disease. Techniques for identifying mixed substructure within a sample are therefore an important part of genetic associations studies [8].

In our setting, we have information on a set of single nucleotide polymorphisms (SNPs), or individual nucleotides that vary across a population. An individual's SNPs can be represented as an l-bit string encoding the variant at each position. We apply our method to a whole genome map of 1.5 million SNPs from 23 African Americans, 24 Asian Americans and 24 European Americans [6]. To avoid linkage disequilibrium from proximity in the genome we use only every thousandth SNP, leaving a total of 1598 markers.

SSCC correctly labels all individuals using 1598 SNPs. We therefore compared SSCC to STRUCTURE on the harder problem of labeling individuals using only 80 of these SNPs. Our method correctly stopped at 3 clusters for 15 of 19 disjoint subsets examined, achieving an average classification accuracy of 91.8% ($\sigma = 0.085$). We then ran STRUCTURE on each of these subsets with standard parameters and 3 clusters, achieved 90.1% average accuracy ($\sigma = 0.10$). This demonstrates that SSCC is competitive with current state-of-the-art methods for this problem.

4.2 Alu structure

Alus are short interspersed nucleotide elements (SINEs) of about 280 nucleic acids that have, in the last 55 million years, been copied about 1 000 000 times, and now make up about 10% of our genome [1]. While it was once believed that a few "master Alus" gave birth to all Alu repeats, recent analyses have shown that a much larger number of Alu elements may be active [7, 4]. Alu families are highly overlapping and therefore resistant to traditional substructure methods. For example, the AluSx, AluSz, AluSp and AluYa5 subfamilies we consider have consensus sequences differing by an average of 12.8 mutations, while the average member of AluSx differs from its consensus sequence by an average of 34.8 mutations ($\sigma = 4.88$).

We compare STRUCTURE to SSCC on a dataset from Price *et al.* [7] of 4000 instances of the

four above-mentioned families. Both methods very nearly recover the subfamily consensus sequences, finding ancestral sequences with Hamming distances $(2, 2, 1, 0)$ and $(3, 1, 1, 0)$, respectively. This small discrepancy may represent a real shortfall of both algorithms, but may also reflect the fact that our dataset contains a non-representative subset of the above-mentioned families.

5 Discussion

We have described a clustering objective function that combines information maximization with constraints on cluster shape, and a top-down sequential algorithm to optimize it. To address the problem of inferring population substructure in the presence of significant overlapping mutation, we define a version of our model which penalizes mutual information between sequence positions within each cluster. This constraint corresponds to the assumption that subpopulations are generated by random, uncorrelated mutations from an ancestral sequence. Our algorithm matches the popular STRUCTURE program in inferring human population structure from SNPs, and significantly outperforms it on the much harder problem of clustering Alu repeat elements.

There remain a number of avenues for further theoretical and algorithmic development. First, one could extend SSCC to the continuous case, using a parametric model of sequence distribution per class. Second, an iterative Blahut-Arimoto style update rule may provide better scalability. Third, it is possible to considered other forms of inter- and intra-cluster penalties. Finally, SSCC can be extended to perform hierarchical clustering by extending C in equation (1) to range over parent populations derived from the cluster splitting rule.

References

[1] MA Batzer and PL Deininger. Alu repeats and the human genomic diversity. *Nature reviews genetics*, 3:370–9, May 2002.

[2] CS Carlson, MA Eberle, L Kruglyak, and DA Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–52, May 2004.

[3] G Chechik and N Tishby. Extracting relevant structures with side information. In *Proc. NIPS*, Cambridge, MA, 2002. MIT Press.

[4] R Cordaux, DJ Hedges, and MA Batzer. Retrotransposition of Alu elements: how many sources? *Trends in Genetics*, 20(10):464–7, October 2004.

[5] TM Cover and JA Thomas. *The elements of information theory*. Plenum Press, New York, 1991.

[6] DA Hinds, LL Stuve, GB Nilsen, E Halperin, E Eskin, DG Ballinger, KA Frazer, and DR Cox. Whole genome patterns of common DNA variation in diverse human populations. *Science*, 307:1072–9, 2005.

[7] AL Price, E Eskin, and PA Pevzner. Whole genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Research*, 14:2245–52, 2004.

[8] JK Pritchard, M Stephens, and PJ Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–59, 2000.

[9] N Shental, A Bar-Hillel, T Hertz, and D Weinshall. Computing gaussian mixture models with EM using equivalence constraints. In S Thrun, L Saul, and B Schölkopf, editors, *Proc. NIPS 16*. MIT Press, Cambridge, MA, 2004.

[10] M Studenty and J Vejnarova. The multi-information function as a tool for measuring stochastic dependence. In MI Jordan, editor, *Learning in Graphical Models*, pages 261–297. MIT Press, Cambridge, MA, 1998.

[11] N Tishby, F C Pereira, and W Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–77, 1999.