# Inference of Population Structure in the Presence of Geographical Migration

**Jing Xiang**
Department of Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213
jingx@cs.cmu.edu

## Abstract

Knowledge about the genetic structure of modern human populations can be useful in inferring human evolutionary history and is important in many medical contexts. Genomic polymorphism data has facilitated our ability to reconstruct the ancestral structures of human populations. Because of the complexity of modern populations, an individual often does not belong to a single population, but belongs to many populations in varying degrees. In addition, population structure is influenced by geographical information. In order to infer population structure that accounts for mixed-membership and geographical effects, we adapt a Latent Variable Model for geographic lexical variation for use on genetic polymorphism data.

## 1 Introduction

Population genetics is concerned with how the genetic content of populations varies over time and the factors that facilitate these changes. The study of population structure, which is the genetic composition of populations can have many useful applications. From it, we can infer the history of modern human populations to understand human evolution and migration. In addition, the differences in genetic structure between populations can be important in the medical context. For example, in pharmacogenetics, population genetic structure is used as an predictor of drug response, such as a drug's safety and efficacy for that ethnic group [6].

The availability of large-scale genomic polymorphism data has greatly benefitted the field of human population genetics. In particular, the work by Rosenberg *et al.* marked the first time that the scientific community could access roughly 400 autosomal microsatellite markers typed on the Human Diversity Project (HGDP-CEPH) cell line panel [4]. This data set included over a thousand individuals from over 50 globally distributed populations. Although human genetic variation is largely influenced by geography, many population genetic models do not use geographical information. The emergence of these kinds of resources has given us the ability to do inference based on models that incorporate geographical data.

Previous work done by Pritchard *et al.*[3] uses a model-based clustering method for inferring population structure and assigning individuals to populations on the basis of their genotypes. It incorporates prior information about the geographic location of individuals. Because modern populations are complex, an individual is not simply a member of one population, but often belongs to several populations. Thus, Pritchard *et al.* use a mixed-membership model and a Markov chain Monte Carlo (MCMC) algorithm to approximate the posterior distribution for inference.

Similar to the previous study, we intend to model populations using a mixed-membership framework and account for geographical effects. However, a different approach will be taken, one that has been applied to text data. Recently, Eisenstein *et al.* [2] focused on investigating geographic linguistic

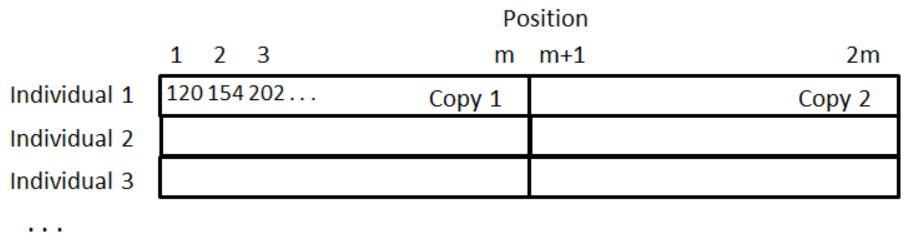



Figure 1: Diagrammatic illustration of the microsatellite data. Each individual has 2 copies (autosomal) and the number of repeats at each position are shown. The number m represents the number of different microsatellites, in this case, m = 377.

variation using text data from the microblogging website Twitter with geographical information in the form of latitude and longitude coordinates. The model used is an extension of Latent Dirichlet Allocation (LDA) [1] where each document is composed of observed words. In addition, each document belongs to a latent topic and originates from an author in a latent region. The approximate inference techniques are based on variational methods, and empirical Bayes parameter estimation is performed with the EM algorithm. The method is described in detail by Eisenstein *et al.* [2]. In this paper, we modify this approach so that it can be used to infer population structure amongst individuals from genetic polymorphism data.

## 2 Methods

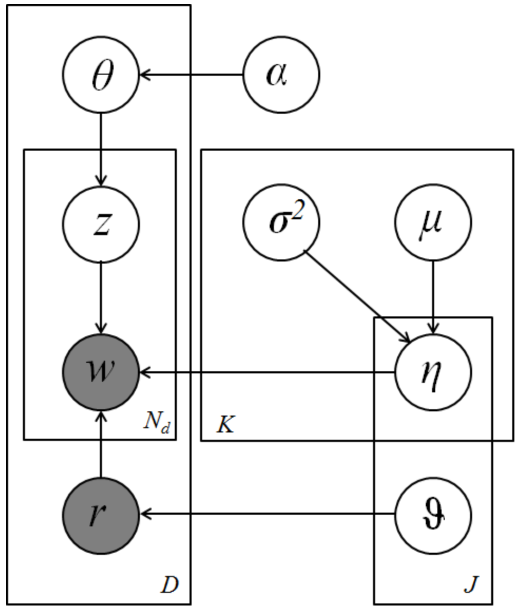
### 2.1 Data

Because the method originally developed by Eisenstein *et al.* [2] was designed for text data with words, documents and topics, the genetic data had to be preprocessed into a suitable format. The genetic data obtained is microsatellite data which includes 377 autosomal microsatellites in 1056 individuals from 52 populations. This was taken from Rosenberg *et al.* [4] and typed on the Human Genome Diversity Project (HGDP-CEPH) cell line panel. Microsatellites are simply repeating sequences of DNA. Figure 1 depicts the format of the data before preprocessing. Because humans are diploid, they have two sets of chromosomes, resulting in two copies of their genetic data. The figure shows each individual with two copies of the 377 microsatellites. Each position contains the number of repeats of that particular microsatellite. Note that the copies from the same individual are different, which is why both must be included.

The existing software for text data assumes a bag of words LDA model. However, with microsatellite data, the position or loci is significant. Thus, in order to create the genetic vocabulary that is independent of position, we concatenate the microsatellite position with the number of repeats. For instance, if individual 1 has microsatellite 1 with 120 repeats, then 1:120 is added to the vocabulary. The vocabulary is then reindexed and the data is relabeled using the new indices. The relabeled data set can then use the bag of words LDA model. It can be deduced that individuals with the same genetic ”words”, the same number of repeats at the same microsatellite positions, will be clustered together.

### 2.2 Model

Because genetic polymorphism data is being used instead of text data, we will need different terms to refer to the elements of the model. For simplicity, the topics are the populations and the documents are the individuals. The words will be referred to as markers, which abbreviates genetic markers. The markers in this case are the microsatellites and the number of repeats observed. The basic model applied is described by Eisenstein *et al.* [2] as the Geographic Topic Model. However, a few modifications have been made to accommodate the application to genetic marker data. First, the individuals are not tagged with latitude and longitude information, instead the geographical information provided is the individual’s continent. Thus, the model has been adjusted such that the




| | |
|---|---|
| $\mu_k$ | mean of base population *k*'s distribution over markers |
| $\sigma^2{}_k$ | variance parameter for regional variants of population *k* |
| $\eta_{jk}$ | region *j*'s variant of the base population $\mu_k$ |
| $\theta_d$ | individual *d*'s population proportions |
| $r_d$ | individual *d*'s region |
| $z_n$ | marker *n*'s population assignment |
| $w_n$ | marker *n*'s observed value |
| $\alpha$ | global prior over individual-population proportions |
| $\vartheta$ | global prior over region classes |

Figure 2: Plate Diagram for the geographic topic model with a table of all random variables. Summarizes the adaptation of the original topic model to genetic data.

regions are observed instead of the latitudes and longitudes. The plate diagram displayed in Figure 2 illustrates the new model.

The modified generative process is as follows:

- **Generate base populations**: for each population k < K,
  - Choose the base population from a normal distribution with a uniform diagonal covariance: $\mu_k \sim \mathcal{N}(a, bI)$.
  - Choose the regional variance from a Gamma distribution $\sigma_k^2 \sim \Gamma(c, d)$.
  - Generate regional variants: for each region j < J, choose the region-population $\eta_{jk} \sim \mathcal{N}(\mu_k, \sigma_k^2 I)$. To convert $\eta_{jk}$ into a multinomial distribution over words, we exponentiate and normalize $\beta_{jk} = \sum_{i=1}^{W} exp(\eta_{jk}^{(i)})$.
- **Generate markers and locations**: for each individual d,
  - Choose population proportions from $\theta \sim$ Dir($\alpha$).
  - Choose region *r* from the multinomial distribution $\vartheta$.
  - For each marker token, choose the population indicator *z* $\sim \theta$ and choose the marker *w* $\sim \beta_{rz}$.

### 2.3 Inference

The inference procedure uses mean-field variational inference which minimizes the Kullback-Leibler divergence between the variational distribution Q and the true distribution. Variational distributions are placed over all the latent variables $\theta, z, r, \vartheta, \eta, \mu$, and $\sigma^2$ and updated until convergence. The details of the inference procedure including the variational distribution updates, mean-field variational inference, and updating the region-population distributions are described fully by Blei *et al.* [1], Wainwright and Jordan [5], and Eisenstein *et al.* [2] respectively.

## 3 Experiments

After the software was adjusted for this application, the HGDP data set taken from Rosenberg *et al.* [4] was used for testing. The data set consisted of 1056 individuals. After it had been relabeled, the number of unique genetic markers was approximately 7000. The most common 3000 markers were used in the vocabulary. The method was executed 20 times for each value of K from 2 to 9. Here we present the results for K = 5. The data set was run with selecting the number of fixed regions to be 1 initially. This was to facilitate comparison with the existing software Structure, developed by Pritchard et al. [3]. Because Structure does not have regional parameters, it is equivalent to using this method with J = 1. The model and inference procedure was also performed with J = 5 to demonstrate the effect of allowing regional variants of populations on prediction.

## 4 Simulation

A simulated data set was also generated from the model. Simulated data can be generated by fixing the parameters, $\mu$, $\sigma^2$, and $\vartheta$. Specifically, $\mu_k$ and $\sigma_k^2$ for each population were fixed to 0 and 1 respectively. The multinomial $\vartheta$ was fixed to the proportions observed in the original HGDP data set. The data was then simulated according to the model described above in section 2.2. For each *r* and $\theta$, draw $\eta_{jk} \sim \mathcal{N}(\mu_k, \sigma_k^2 I)$. For each individual, sample $r \sim \vartheta$ and $\theta \sim \mathrm{Dir}(\alpha)$. The topic indicators and words can then be drawn. The vocabulary generated consisted of 1000 different genetic markers and 500 individuals were simulated. This simulated data was then used to evaluate the inference procedure.

## 5 Results and Discussion

The following describe the outcome of the experiments using the geographical population model and inference procedure.

### 5.1 Evaluation on HGDP Data and Comparison with Structure.

For the diagrams of population structure that follow, each vertical bar represents an individual. Each colour denotes a different population. Thus, the coloured bars show the extent to which each individual belongs to that particular population. This set of results was generated with setting the number of populations to 5.

If we assume that all individuals came from the same region, then the population structure is shown in Figure 5. The continents Africa, America, East Asia and Oceania are distinct, while the populations from Central South Asia, Europe and the Middle East are more similar according to this model. To check whether the model was making reasonable predictions, the results from using the same data set with the existing software, Structure, was used for comparison. As shown in Figure 6, Structure generates similar results.

However, when the number of regions is fixed to 7, a different plot is observed. As compared with Figure 5, Figure 7 exhibits regions with individuals that are a mixture of different populations. This mixing effect is a caused by the representation of regional variants of populations in the model. This indicates that the individuals of within a region are a mixture of populations. This suggests that populations have migrated across geographical regions.

Another way of examining the results is analyzing how the population proportions that are predicted differ between individuals of a particular region. Thus, the Euclidean distance between the population proportions of each individual from one continent and every individual in another continent is calculated. The entries in the following tables report the mean of the individual distances, shown in the equation below. $\theta$ is a vector of the population proportions, n is the number of individuals in the first continent and m is the number of individuals in the continent being compared with.

$$D = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \|\theta_i - \theta_j\|; \tag{1}$$

It can be observed by comparing Figures 3 and 4 that many of the distances between continental groups of individuals are smaller in the model allowing for regional variants (J = 7). If you recall from Figure 7, more mixture of populations is present within the individuals of several continents. For example, if you compare Oceania and Central South Asia in Figures 5 and 7, it is clear that the individuals in each continent are more similar in the latter. This is reflected in the distance between them, 0.64 in Figure 3 as compared to 1.02 in Figure 4. This can also be explained by the $\eta_{jk}$'s in the model which signifies that each region has a variant of the base population. Thus, the continental groups are not independent of each other.

| | Africa | America | Central South Asia | East Asia | Europe | Middle East | Oceania |
|---|---|---|---|---|---|---|---|
| Africa | 0.10 | 1.26 | 0.96 | 1.14 | 1.05 | 0.95 | 1.29 |
| America | | 0.14 | 1.01 | 1.11 | 1.11 | 1.10 | 1.30 |
| C/S Asia | | | 0.23 | 0.80 | 0.26 | 0.27 | 1.02 |
| East Asia | | | | 0.15 | 0.98 | 0.96 | 1.14 |
| Europe | | | | | 0.13 | 0.18 | 1.15 |
| Mid East | | | | | | 0.18 | 1.13 |
| Oceania | | | | | | | 0.05 |

Figure 3: The table shows the average Euclidean distance between population proportions $\theta$ of every individual in one continent measured against every individual in another. The $\theta$'s are calculated from the model where (K,J) = (5,1).

| | Africa | America | Central South Asia | East Asia | Europe | Middle East | Oceania |
|---|---|---|---|---|---|---|---|
| Africa | 0.10 | 1.28 | 0.93 | 1.05 | 1.01 | 0.95 | 1.09 |
| America | | 0.09 | 0.94 | 0.54 | 1.06 | 1.08 | 1.03 |
| C/S Asia | | | 0.29 | 0.58 | 0.36 | 0.40 | 0.64 |
| East Asia | | | | 0.14 | 0.77 | 0.80 | 0.50 |
| Europe | | | | | 0.28 | 0.34 | 0.84 |
| Mid East | | | | | | 0.36 | 0.86 |
| Oceania | | | | | | | 0.15 |

Figure 4: The table shows the average Euclidean distance between population proportions $\theta$ of every individual in one continent measured against every individual in another. The $\theta$'s are calculated from the model where (K,J) = (5,7).

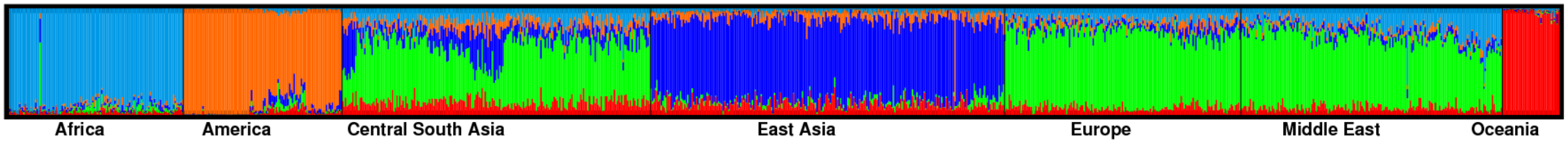



Figure 5: Illustrates the population structure predicted with 5 populations (K = 5) and 1 region (J = 1).

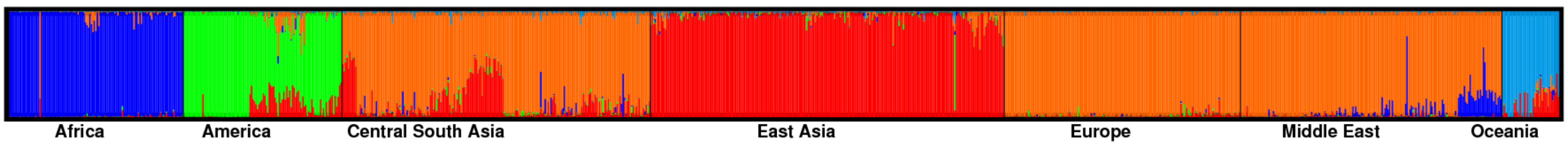



Figure 6: Displays the populations predicted by the software Structure with K = 5.

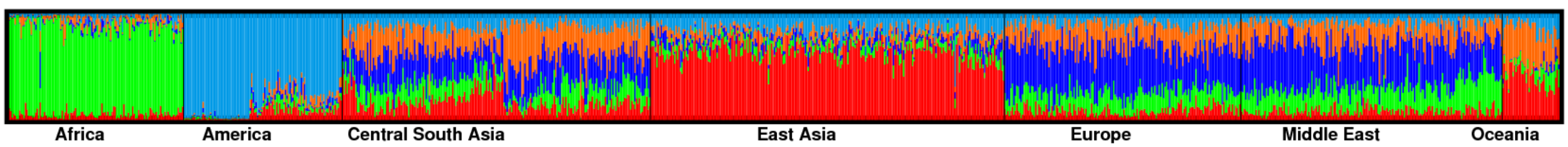



Figure 7: Displays the population structure predicted with 5 populations (K = 5) and 7 regions (J= 7).

### 5.2 Evaluation on Simulated Data Set

As described previously, a simulated data set was generated from the model with fixed mean, variance and population proportions. Testing the model and inference procedure on simulated data is another avenue to check that it is generating correct results. Figure 8 shows the results for 500 simulated individuals. In general, the population structure looks similar to that which has been observed previously such as in Figure 7. However, there is one difference in that America and Africa seem to be grouped in the same population. This is not the case in other results provided by the model or Structure. It could be a result of different means and variances. However, this needs to be investigated further.

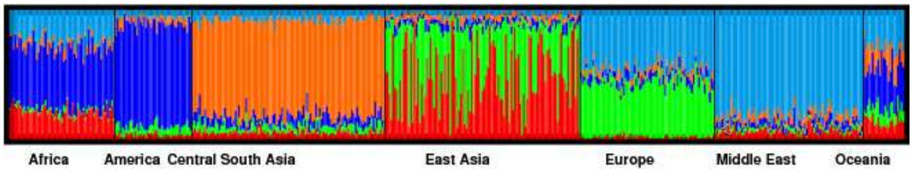



Figure 8: Genetic marker data for 500 individuals is simulated from the model. The geographical population model is then used to predict the proportions of K populations to which each individual contains. The number of regions J is fixed to 7.

## 6 Conclusion

We present a model that identifies the relationship between geography and population variation. The model is modified from an existing geographical topical model originally used for linguistic variation. It has been demonstrated that if we assume a single fixed region, then this model will obtain similar results to the existing software Structure. However, when multiple regions are specified, the individuals within each region are predicted to be more of a mixture of populations. This results from the regional variants described by the model and is suggestive of the migration of humans across geographical regions.

A possible extension of the model is to incorporate latitude and longitude information so that the regions can be latent. Since more specific regional information can be obtained, it can be translated into known latitudes and longitudes. While the results shows that each region consists of many populations, the limitation of the model is that it cannot identify the location where each population originates and where they migrate to. We can think of a graph where the nodes are the regions and the edges represent migration of populations between the two regions. The goal of future work is to do structured learning of the graph and identify which edges are present. This will then help us understand the pattern of human migration.

### References

[1] D.M. Blei, A.Y. Ng and M.I. Jordan. Latent Dirichlet Allocation. (2003). *Journal of Machine Learning Research*, 3: 993-1022.

[2] J. Eisenstein, Brendan O'Connor, N.A. Smith, and E.P. Xing. (2010). A Latent Variable Model for Geographic Lexical Variation. *Proceedings of EMNLP*.

[3] J.K. Pritchard, M. Stephens, and P. Donnelly. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2): 945-59.

[4] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman. (2002). Genetic structure of human populations. *Science*, 298: 2381-2385.

[5] M.J. Wainwright and M.I. Jordan. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(12): 1305.

[6] J.F. Wilson, M.E. Weale, A.C. Smith, F. Gratrix, B. Fletcher, M.G. Thomas, N. Bradman and D.B. Goldstein. (2001). Population genetic structure of variable drug response. *Nature Genetics*, 29: 265-269.