Inference of Population Structure with Optimal Number of Ancestral Groups

Daegun Won
Advisor: Eric Xing
Graduate Advisor: Suyash Shringarpure

SCS Senior Thesis
Spring 2010

Abstract

In this paper we study the problem of population structure inference using multi-locus genotype data. Traditional methods for inferring population structure such as Structure program or *mStruct* does not present a good way to optimize the number of ancestral population groups by including the number in the model and inferring from the model itself. In this paper we present a model that will have the ability to infer the optimal number intrinsically. We tested the model against a number of simulated dataset and the number of 'dominant' ancestral population groups was identical to the optimal number, while keeping the admixture accuracy in a reasonable level.

1 Introduction

There are many scenarios that the population structure can be useful in the field of population genetics. It can provide valuable information in evolutionary history and migrations of human population. Sometime it can also be used to classify individuals of unknown origin. However, in any case, there needs to be a set of predefined populations to use. Therefore, identifying the population structure has been a very traditional problem in population genetics.

Some of the approaches to this problem assume that each modern individual originated from a single ancestral population. This approach transforms the problem into a clustering problem. Then a natural way of solving this problem will be using distance-based methods. This type of approaches has an advantage that it is very easy to apply and visually attractive. But it involves constructing both meaningful and reasonable distance metric and graphical representation. Although this might be a good way of solving the problem on a high level due to its simplicity and visual interpretability, it is not the best approach to solve the problem in a fine scale. Furthermore, the assumption that each individual originated from a single population is not a reasonable assumption considering mating.

Thus many of recent works assume that individuals originated from a number of populations. *Structure*, implemented by Pritchard *et al*., proposes a model called allele-frequency admixture model. *Structure* assumes that each allele at each locus is an independent draw from the appropriate population-specific multinomial distribution of marker allele. The model identifies each ancestral population (AP) by its allele-frequency profile, which is a vector of allele frequencies of each allele in the ancestral population, and the fraction of contribution from each AP in a modern individual by an admixing vector. Overall, this model is essentially a direct application of Latent Dirichlet Allocation (LDA).

An improvement of *Structure*, called *mStruct*, has been proposed recently by incorporating the possibility of mutation into the probabilistic model. Due to the newly introduced mutation parameter, the model is more expressive than *Structure* and provides a better result. However, it is necessary to define the number of ancestral populations to use either *Structure* or *mStruct*. So the existing methods try a number of values and pick the one that gives the highest BIC value to find the optimal number of APs. Although this is a reasonable way, it would be more desirable to get the optimal value from the model itself rather than trying a number of values. Also this way involves multiple inferences, which may slow down the calculation dramatically if the inference process is slow. However, it is often very challenging to get rid of this trial stage and automatically infer the optimal value.

Our goal in solving this problem is to get rid of the search for the optimal number of APs. To make the problem simple to start with, we focus on extending *Structure*. Specifically, we aim to extend this LDA-based model to a hierarchical dirichlet process (HDP) based model so that we can get the optimal number of APs from the posterior mean of the admixture vector.

2 Previous models

In this section we describe the representation of the two previous models, *Structure* and *mStruct*. They are applications of latent dirichlet allocation. Briefly, the generative process can be described as the following:

- Draw an admixing vector an individual $j : \vec{\theta}_j \sim P(\cdot \mid \alpha)$
- For each allele X_{ji}
  - Draw the ancestral population origin indicator $Z_{ji} \sim Multinomial(\cdot \mid \vec{\theta}_j)$
  - (for *Structure*)
    Draw an allele $X_{ji} \mid Z_{ji} = k \sim P(\cdot \mid \Theta^k)$ for some population-specific parameters Θ^k
  - (for *mStruct)*
    Draw a founder allele indicator $C_{ji} \mid Z_{ji} = k \sim Multinomial(\cdot \mid \vec{\beta}^k)$
    and an allele $X_{ji} \mid C_{ji} = l, Z_{ji} = k \sim P(\cdot \mid \mu^k_{jl}, \delta^k_{jl})$

As we can see in Figure 1, both of these models need the number of ancestral groups (K) specified. *Structure* uses a Monte Carlo Markov Chain (MCMC) method to get the posterior mean of the admixing vector for each individual and *mStruct* uses a variational method that converges faster than the MCMC method. These models determine the optimal number by maximizing likelihood or Bayesian Information Criterion after a number of trials.

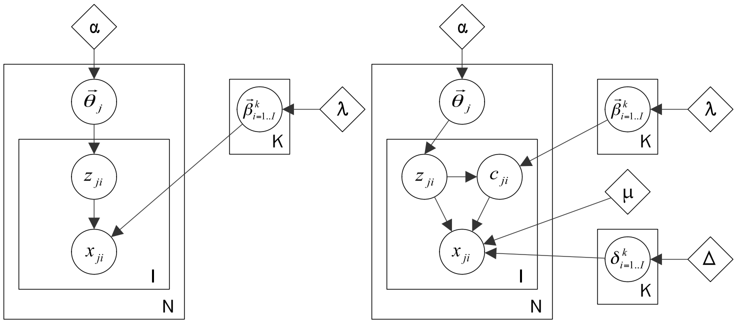

Figure 1: (Left) The model used by Structure. (Right) The model used by mStruct.

3 Our Model

3.1 Generative Process

Let x_{ji} be the observed data for individual j at locus i. Then the hierarchical dirichlet process is then used generatively for the jth individual as follows:

$$\begin{aligned} G_0 \mid \gamma, H &\sim DP(\gamma, H) \\ G_j \mid \alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ \theta_{ji} \mid G_j &\sim G_j \\ x_{ji} \mid \theta_{ji} &\sim multinomial(\theta_{ji}) \end{aligned}$$

Note that θ is now indicating something different than what it indicated in *Structure*. Here θ is the prior for multinomial distribution that draws the observed allele x.

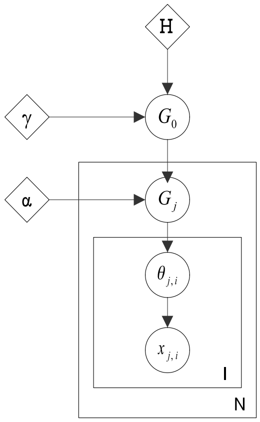

Figure 2: The proposed model

3.2 Chinese Restaurant Franchise Analogy

Our construction follows the Chinese Restaurant Franchise (CRF) described in [2] almost exactly. The analogy to the CRF is shown in Figure 3. Each individual can be considered to be a restaurant and the loci are customers. The dish served to a particular customer in a restaurant is the population of origin for the locus in that individual. Our construction departs from the CRF only in that the support for the data distribution is the different at each locus. In other words, the space of the observed alleles is different at different loci. However, this difference can be eliminated by considering that the support at each locus is the union of the supports at all loci. Also to deal with loci having alleles, we differentiate the same allele at different loci. For example, allele 1 at locus 1 and allele 1 at locus 2 are considered to be different.

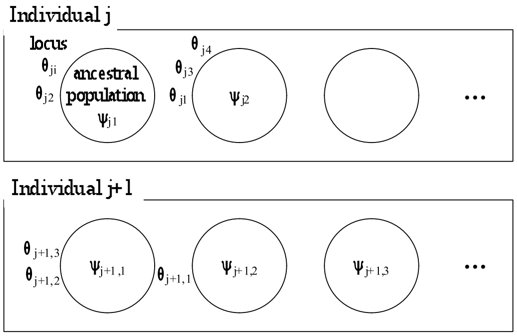

Figure 3. Analogy to CRF

4 Inference

We use one of the three Markov Chain Monte Carlo methods derived by Teh et al [2] to approximate the posterior distribution. Specifically, the method used is ‘posterior sampling by direct assignment’. This scheme directly maps each allele at locus *i* of an individual *j* to an ancestral group *k* by introducing a variable z_{ji}. The inference process has the following steps:

```
• Initialize random values for all variables
repeat
    • Sample z given all other variables
    • Sample m given all other variables and updated value of z
```

- Sample β given all other variables and updated value of **z** and **m**
- Sample α_0
- Sample γ

until convergence

- Calculate the admixture vector for each individual by inspecting **z**

4.1 Updating Variables

(1) Sample Z:

$$p(z_{ji} = k \mid \mathbf{z}^{-ji}, \mathbf{m}, \beta) = (n_{j.k}^{-ji} + \alpha_0 \beta_k) f_k^{-x_{ji}}(x_{ji}) \quad \text{(for previously used k)}$$

$$= \alpha_0 \beta_\mu f_k^{-x_{ji}}(x_{ji}) \quad \text{(for new k)}$$

$$f_k^{-x_{ji}}(x_{ji}) = \frac{B(h_1 + \sum_{j'i', z_{j'i'}=k} I(x_{j'i'} = a_1), \ldots, h_P + \sum_{j'i', z_{j'i'}=k} I(x_{j'i'} = a_P))}{B(h_1 + \sum_{j'i' \neq ji, z_{j'i'}=k} I(x_{j'i'} = a_1), \ldots, h_P + \sum_{j'i' \neq ji, z_{j'i'}=k} I(x_{j'i'} = a_P))}$$

where $a_i$is each observed allele and $n_{j.k}$ is number of alleles of the individual *j* assigned to the ancestral group *k*. h_i's are priors set for each allele observed and each superscript represents a variable that should be skipped when calculating the function.

(2) Sample M:

$$p(m_{jk} = m \mid \mathbf{z}, \mathbf{m}^{-jk}, \beta) = \frac{\Gamma(\alpha_0 \beta_k)}{\Gamma(n_{j.k} + \alpha_0 \beta_k)} s(n_{j.k}, m)(\alpha_0 \beta_k)^m$$

Here, s(n,m) is an unsigned stirling number of the first kind.

(3) Update β:

$$(\beta_1, \ldots, \beta_k, \beta_\mu) \mid \mathbf{m} \sim Dir(m_{.1}, \ldots, m_{.K}, \gamma)$$

$m_{.i}$ represents sum of all m_{ji}'s.

4.2 Updating Hyperparameters

(1) Update α

$$\alpha_0 \mid \mathbf{w}, \mathbf{s} \sim Gamma(a_0 + m.. - \sum_{j=1..J} s_j, b_0 - \sum_{j=1..J} \log(w_j))$$

$$w_j \mid \alpha_0 \sim Beta(\alpha_0, n_{j..} - 1)$$

$$s_j \mid \alpha_0 \sim Bernouli(n_{j..} / \alpha_0)$$

(2) Update γ

$$\gamma \mid \eta, K \sim \pi_\eta Gamma(a_1 + k, b_1 - \log(\eta)) + (1 - \pi_\eta) Gamma(a_1 + k - 1, b_1 - \log(\eta))$$

$$\eta \mid \gamma, K \sim Beta(\gamma + 1, m..)$$

$$\left(\frac{\pi_\eta}{1 - \pi_\eta} = \frac{a_1 + k - 1}{m..(b_1 - \log(\eta))} \right)$$

In this step, $m_{..}$ represents sum of all $m_{.i}$'s and K represents the number of different values of z_{ji}.

5 Experimental Results

To see the correctness of the result, we tested *HDPStructure* against a number of simulated data sets generated by the program used in Shringarpure et al [5]. First, we tested on 4 different simulated datasets and then on a single dataset with different initializations. The goal of this experiment is to show that this model achieves the main objective, which is gettingg the optimal number of ancestral groups from the model without human intervention.

In the figures, each vertical line represents an individual, each color represents an ancestral group, and the length of each color means the amount of the contribution of the ancestral group.

5.1 Validation on Coalescent Simulation

To verify the correctness of the estimation of *HDP-Structure*, we first simulated a number of data sets, using coalescent techniques used in Shringarpure et al [5]. Due to the heavy calculations and slow convergence of the inference steps, the test sets were generated in a small scale with two optimal ancestral populations. To estimate the error of the admixture vector, we calculated the average of the differences of population 1's contributions. Table 1 presents the specification and the summary of each dataset, and Figure 3 shows the estimations from *HDP-Structure* compared against the estimations of *mStruct.*

Dataset	# Individuals	# Loci	Ploidy	$\text{Error}_{pop(1)}$
1	75	20	2	0.128
2	60	15	2	0.115
3	50	12	2	0.124
4	50	10	2	0.101

Table 1: Summary of datasets

The estimation results show that the estimation of *HDP-Structure* makes a reasonably good estimation of admixture, around 10~12% error in terms of the contribution of the first population group. Also, the number of 'dominant' or 'significant' ancestral groups matches the optimal number of the ancestral groups. The actual number of ancestral groups varied around 3 to 10, but all of them do not have enough significance as shown in the graphs.

5.2 Convergence to the optimal number of ancestral groups

Although the correctness of the number of ancestral groups was shown in the previous experiment, we tested *HDP-Structure* and *mStruct* on one dataset with different settings of number of populations. This test was necessary because if extra ancestral groups inferred by *mStruct* are not significantly affecting the modern population so that it is almost negligible, it greatly reduces the meaning of this project. The dataset had 50 people with alleles observed at 10 loci from each of two sets of chromosome and the number of populations was set to 2, 3, 5 and 7 respectively. The estimation results are shown in figure 4.

As we can see in the figure, *HDP-Structure* is not highly affected by the initial number of ancestral groups. It still keeps the number of dominant ancestral population groups to two and the compositions stay consistent. However, *mStruct* gives a noticeable change in the composition as the number increases. At the beginning it seems like the optimal ancestral groups split into multiple subgroups but this trend does not last long and gives a completely different estimate soon.

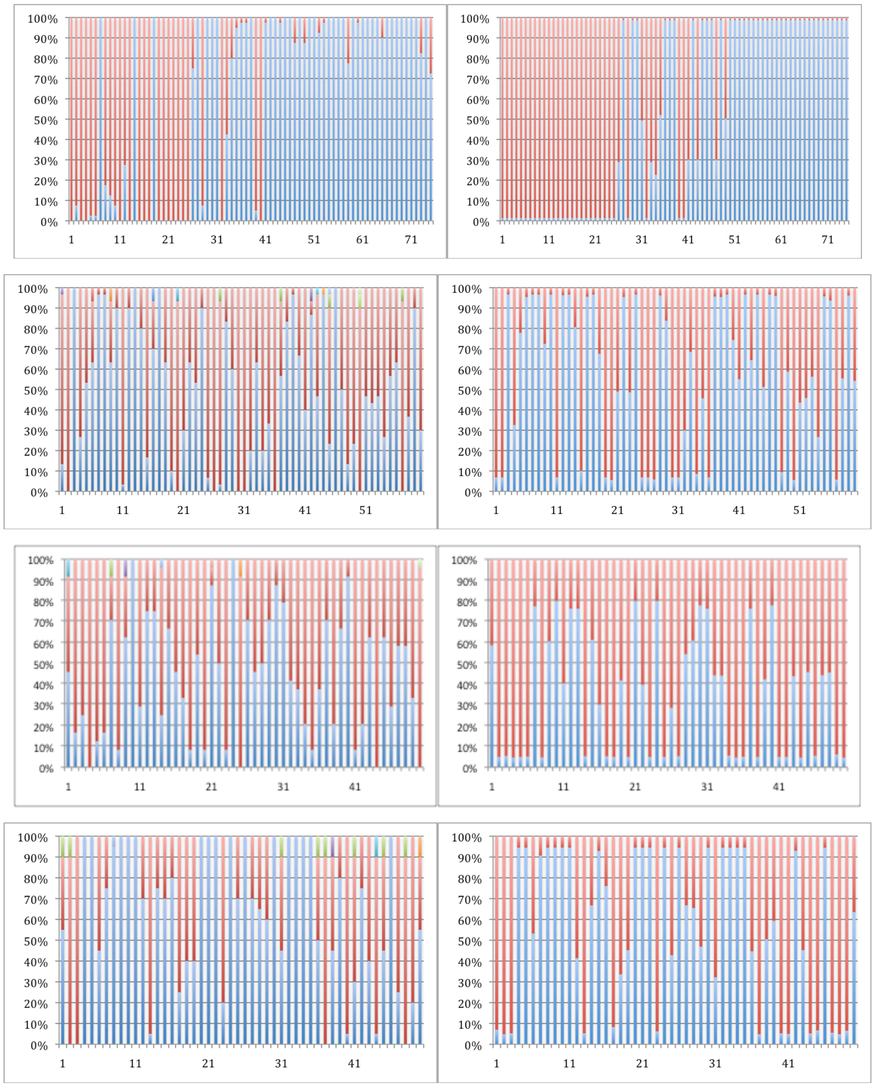

Figure 3: Inference results of *HDP-Structure* (left) and *mStruct* (right) against four datasets (each row)

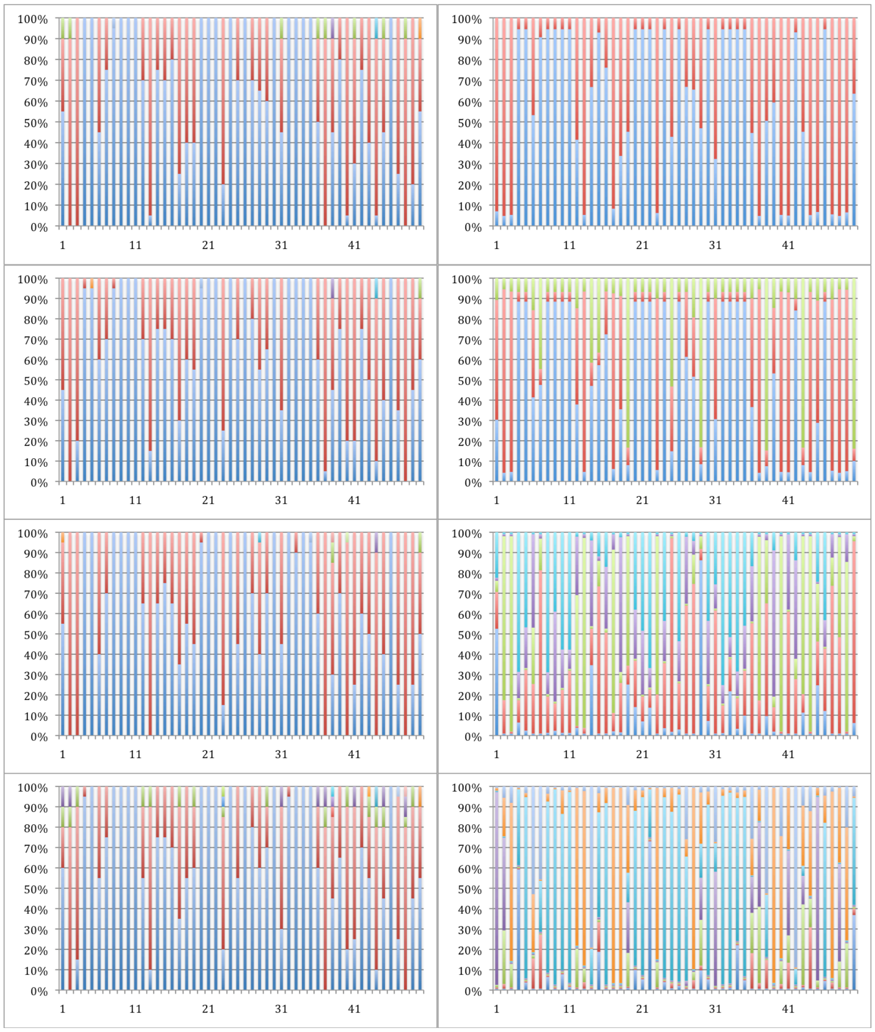

Figure 4: Inference results of *HDP-Structure* (left) and *mStruct* (right) with (initial) number of populations set to 2, 3, 5, 7 respectively in each row

6 Conclusion and Future Works

From the tests on the simulated datasets, we confirmed that the model picked up the optimal number of population correctly. Initial settings with higher number of populations introduced more noises. However, this result is expected given that the number of iterations was the same for each initial setting. There might be multiple ways of removing or minimizing the noise: one could be taking empirical posterior mean. Currently *HDP-Structure* takes only one posterior sample due to the nature of HDP adding and removing mixture components. However we could still take the posterior mean by 'deactivating' mixture components instead of just removing ones. This will minimize the contribution of each noise component, although it would not reduce the number of ancestral groups. But we can easily handle this once we set a threshold of contribution.

Another big issue that should be improved is its speed. Compared to *mStruct*, the inference step presented in this project took much more iterations to converge. For instance, the variational inference method used in *mStruct* converged within 10~30 iterations, but the MCMC method we used here took at least around 3000 iterations to get stable. Furthermore, each iteration was much slower as well. Considering the slow convergence of MCMC methods, other inference methods using techniques such as variational inference or mean field approximation should be developed. Speed improvement is very necessary because testing on human datasets or larger sets are missing because of the slow speed.

Since this model is an extension of *Structure*, which does not take the mutation process into consideration, another possible extension is considering the mutation process as *mStruct* does.

In summary, recent population stratification methods such as *Structure* and *mStruct* require human belief and a post inference process to get the optimal number of ancestral groups. By extending the LDA based models to a HDP mixture model, the *HDP-Structure* approach presented in this project attempts to achieve a better justification of the optimal number while keeping almost the same level of accuracy of admixture vectors each individual.

Acknowledgements

I would like to thank my advisor Professor Eric P. Xing for letting me have this great research opportunity. He encouraged me to have more passion in academics and research and gave me valuable advices on academics and my career as a computer scientist.

Also I want to thank my graduate student advisor Suyash Shringarpure for his suggestions and guidance throughout the project. Without his help, I would have had much more difficult time with numerous misunderstanding, questions, and program errors I had.

References

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

[2] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, (2006) Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101(476):1566–1581.

[3] J. Pritchard, M. Stephens, and P. Donnelly, (2000) Inference of Population Structure using Multilocus Genotype Data. *Genetics* 155:945–959

[4] R. Neal, (2000) Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, Vol.9, No.2. (Jun.,2000):249–265.

[5] S. Shringarpure, and E. Xing, (2009) mStruct:Inference of Population Structure in Light of Both Genetic Admixing and Allele Mutations. *Genetics*

[6] E. Xing, R. Sharan. and M. Jordan, (2004) Bayesian Haplotype Inference via the Dirichlet Process. *Proceedings of the 21st International Conference on Machine*, ACMPress, 879–886