Inference of Population Structure with Optimal Number of Ancestral Groups

Daegun Won
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
daegunw@andrew.cmu.edu

Abstract

In this project I 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 I present a model that will have the ability to infer the optimal number intrinsically. I tested the model against a number of simulated dataset and the number of 'dominant' ancestral population groups were identical to the optimal number, while keeping the admixture accuracy in a reasonable level.

1 Introduction

Identifying the population structure given the genetic information of individuals has been a very traditional problem in population genetics. Recently there have been several approaches trying to solve this problem by assigning individuals to populations. *Structure*, implemented by Pritchard *et al.*, proposes a model called allele-frequency admixture model that is similar to Latent Dirichlet Allocation. *Structure* assumes that each allele at each locus in each genotype is an independent draw from the appropriate distribution. An improvement of *Structure*, which is called *mStruct*, has been proposed recently by incorporating the possibility of mutation into the probabilistic model. While mStruct provides a good result, the optimal choice of the number of ancestral groups remains uncertain because the model itself did not make the choice but an approximation not from the model did. With this ability missing, the inference result of population structure is a bit questionable although it should be fairly reasonable. In the next few sections, I first explain the background information such as models for the population structure and two major previous works project, I aim to develop a probabilistic model that has the ability to infer the optimal number of ancestral groups from the model itself to offer a better justification of the choice of the number of ancestral groups.

2 Previous models

The previous models are essentially applications of Latent Dirichlet Allocation (LDA). Briefly, the generative process is the following:

1. Draw an admixing vector an individual *n*: $\vec{\theta}_n \sim P(\cdot \mid \alpha)$

2. For each allele X_{i,n_e},

 2.1 Draw the ancestral population origin indicator $Z_{i,n_e} \sim Multinomial(\cdot \mid \vec{\theta}_n)$

2.2
(for *Structure*) Draw the allele $X_{i,n_e} | Z_{i,n_e} = k \sim P(\cdot | \Theta^k)$ for some
population-specific parameters Θ^k.
(for *mStruct)* Draw a founder allele indicator $C_{i,n_e} | Z_{i,n_e} = k \sim Multinomial(\cdot | \vec{\beta}_i^k)$
and the allele $X_{i,n_e} | C_{i,n_e} = l,\ Z_{i,n_e} = k \sim P(\cdot | \mu_{i,l}^k, \delta_{i,l}^k)$,

As we can see in Figure 1, both of these models need the number of ancestral groups (K) specified. These models determine the optimal number by getting Bayesian Information Criterion or a similar kind of evaluation function.

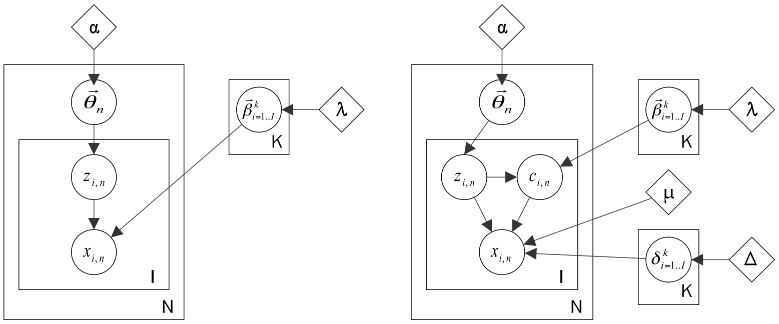

Figure 1: (Left) The model used by Structure. (Right) The model used by mStruct.

3 Proposed Model (*HDP-Structure*)

The new model I suggest is basically a Hierarchical Dirichlet Process mixture model. Given that the previous models are applications of Latent Dirichlet Allocation (LDA), this extension is a natural way of extending the previous models to get the optimal number of ancestral groups. To avoid confusions, I follow the notations used in Teh et al [2]. In the context of population structure inference problem, $x_{j,i}$ represents the observed allele of individual *j* at locus *i*, and $\theta_{j,i}$ represents the multinomial prior for allele at locus *i* of individual *j*.

One noticeable difference is that unlike the previous models, which had different sets of possible alleles for each locus, this model does not differentiate each locus. Instead, it has the entire set of observed alleles as the support of allele distribution. from the whole set of observed alleles

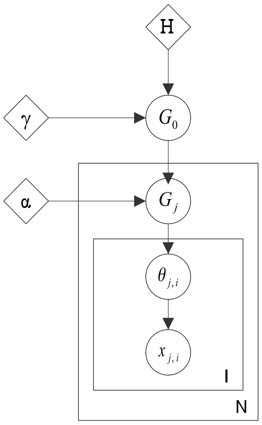

Figure 2: The proposed model

4 Inference

For the inference step, I used a Markov Chain Monte Carlo method derived by Teh et al [2]. Specifically, the method used is 'posterior sampling by direct assignment'. This scheme directly maps allele *i* of an individual *j* to an ancestral group *k* by introducing a variable $z_{ji.}$ The specific sampling steps are the followings:

(1) Initialize z_{ji}'s randomly with uniform probability $1/K_{init}$ for each $k = 1 \ldots K_{init}$. Initialize

(2) Sample Z:

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

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

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

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.

(3) 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.

(4) Update β:

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

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

(5) Update α

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

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

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

(6) 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, I tested *HDP-Structure* against a number of simulated data sets generated by the program used in Shringarpure et al [5]. First, I tested on 4 different simulated datasets and then I tested on a single dataset with different initializations to show that this model achieves the main objective – getting the optimal number without human intervention. For space issue, the tables and figures for this section are attached at the end of the report. 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*, I 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, I 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.*

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 match 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, I 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.

6 Conclusion and Future Works

From the tests on the simulated datasets, I 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 I 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

This is also my senior-thesis research project advised by Professor Eric Xing and Suyash Shringarpure, which is one academic-year long thesis.

References

[1] D. Blei., A. Ng, and M. Jordan, (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research3*: 993–1022.

[2] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M., (2006) Hierarchical Dirichlet Process es. *Journal of the American Statistical Associaton*, 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, S., 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*, ACM Press, 879---886

Dataset	# Individuals	# Loci	Ploidy	$\text{Error}_{\text{pop}(1)}$
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

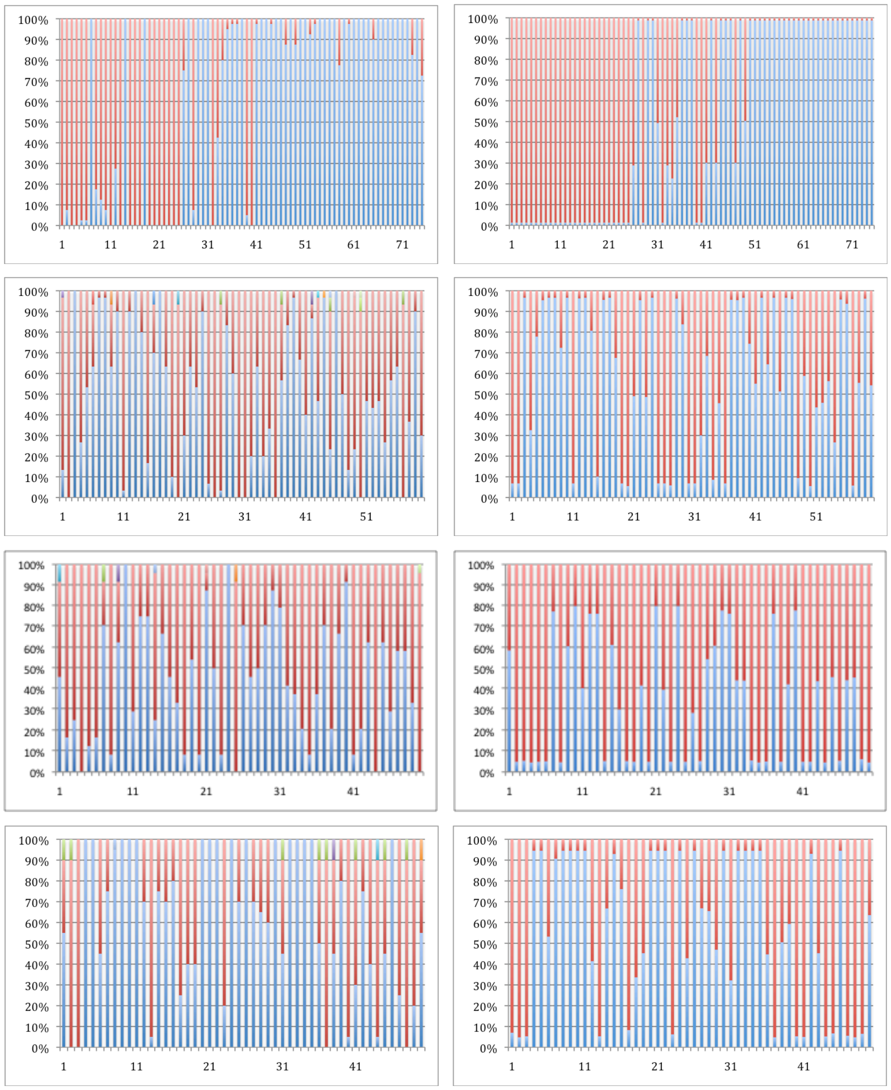

Figure 3: Inference results of *HDP-Structure* (left) and *mStruct* (right) against four datasets (each row)

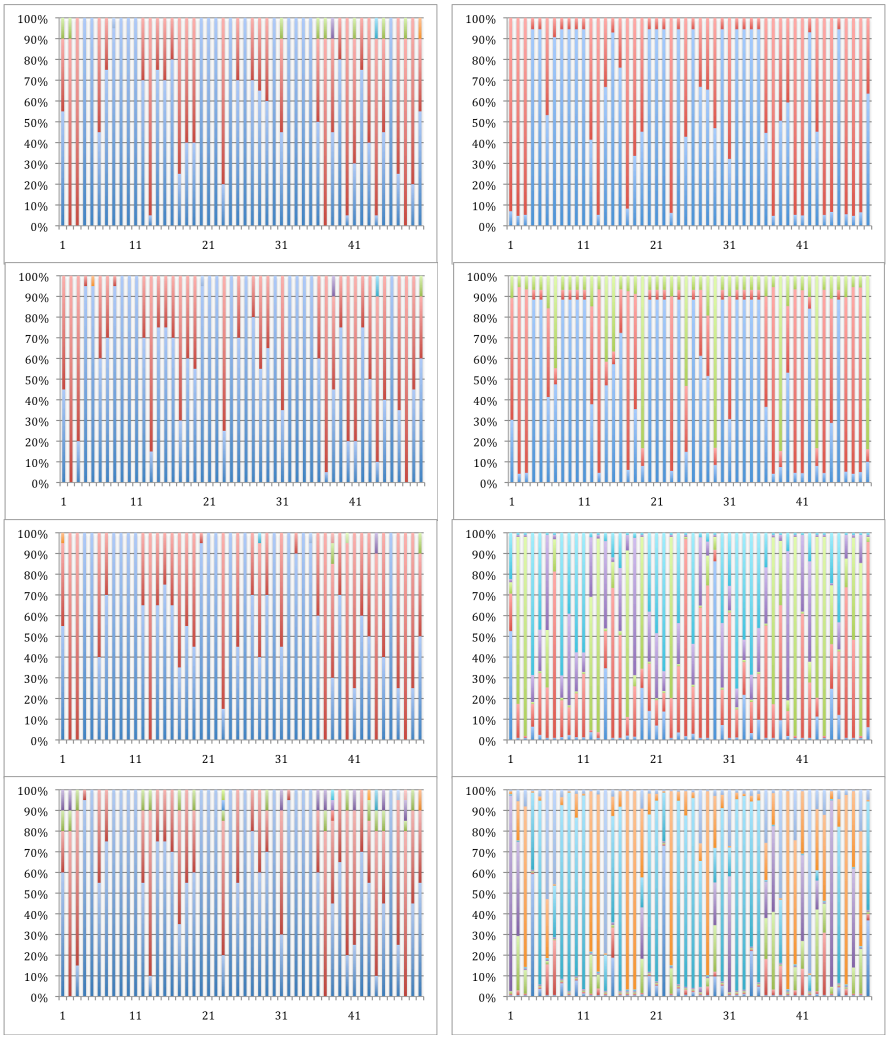

Figure 4: Inference results of *HDP-Structure* (left) and *mStruct* (right) with (initial) number of populations set to 2, 3, 5, 7 respectively in each row