Learning population structure using Multilocus genotype data

STATS 701 Presentation

Trong Dat Do - 6 Oct 2020

Agenda

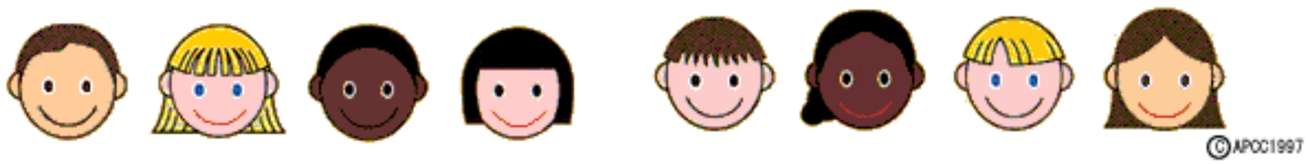
1. Preliminaries
2. Population structure models (Pritchard, Stephens & Donelly 2000)
3. Examples and applications of STRUCTURE model
4. Extensions of the model
5. Discussion

1. Preliminaries

- In researching population genetics, it is useful to assign each individual in the sample to a population, then study some properties of the population or talk about their origins.

- But how to define a population? It can be very subjective: Based on linguistic, culture, physical characters,….

- In this presentation: We are going to cluster individuals into populations using genetics information.



- Let $X_1, X_2, \ldots, X_N$ be all individuals in the sample
- Each X_i has the following form

$$X_i = \begin{pmatrix} X_i^{(1,1)} & X_i^{(1,2)} \\ X_i^{(1,1)} & X_i^{(1,2)} \\ \vdots & \vdots \\ X_i^{(L,1)} & X_i^{(L,2)} \end{pmatrix}$$

where L is number of loci

- So we will try to cluster the population based on multilocus genotype data instead of subjective criteria.

$$\begin{pmatrix} Aa \\ B_1b_2 \\ C_2c_1 \\ DD \end{pmatrix}, \begin{pmatrix} aa \\ B_2b_1 \\ C_1c_1 \\ Dd \end{pmatrix}, \begin{pmatrix} AA \\ B_1b_1 \\ C_2c_1 \\ dd \end{pmatrix}, \begin{pmatrix} Aa \\ B_1b_2 \\ C_1c_2 \\ DD \end{pmatrix}$$

Statistically speaking: Clustering multi-dimensional categorical data.

2. Population structure models

TITLE	CITED BY	YEAR
Inference of population structure using multilocus genotype data JK Pritchard, M Stephens, P Donnelly Genetics 155 (2), 945	29327	2000

- Inference of population structure using multilocus genotype data [1] has been cited almost 30k times in the last 20 years.
- When Pritchard did his postdoc under the supervision of Donelly, they met Stephens and shared the idea in a workshop in Cambridge (1998). All of them had expertise in clustering using Bayesian methods. [2]
- A program software is written based on this: STRUCTURE [3].

Basic model (no admixture)

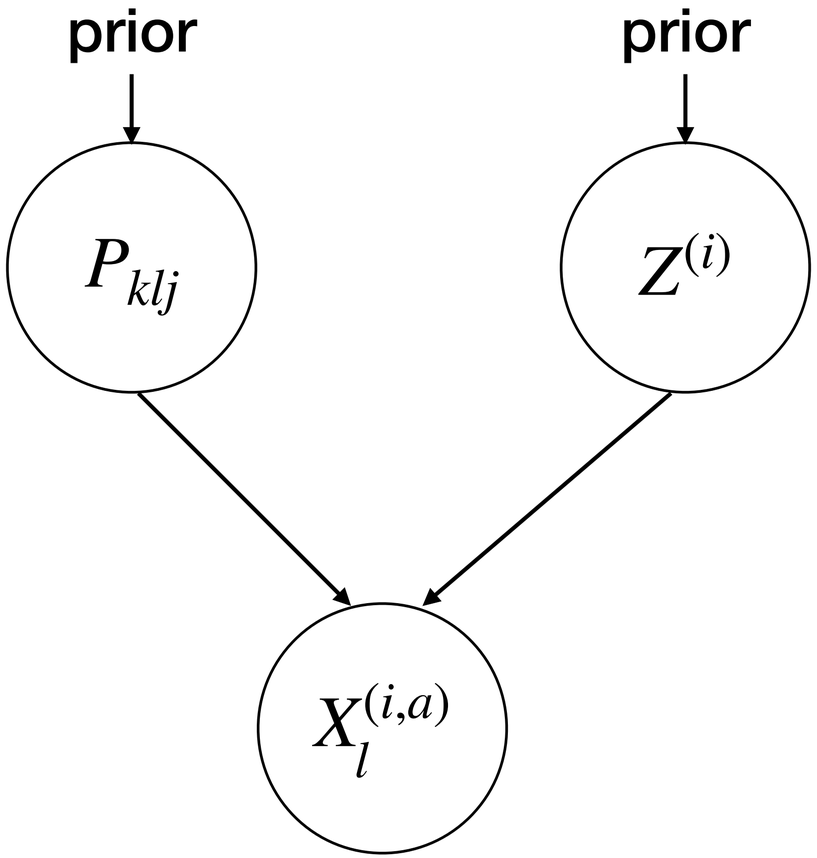
- We have N individuals $X_1, \ldots, X_N$, each has information in L loci

$$X_i = \begin{pmatrix} X_i^{(1,1)} & X_i^{(1,2)} \\ X_i^{(1,1)} & X_i^{(1,2)} \\ \vdots & \vdots \\ X_i^{(L,1)} & X_i^{(L,2)} \end{pmatrix}$$

- Choose a K: Number of population (clusters), we want to assign each individual to a cluster. Denote $Z^{(i)}$ the population of i-th individual, $i = 1,\ldots,N$
- There is a Hardy-Weinberg equilibrium within each population:

$$P_{klj} = \text{Proportion of allele } j \text{ in locus } l \text{ within population } k$$

Basic model (no admixture)



- $Z^{(i)}$ the population of i-th individual, $i = 1,\ldots,N$
- P_{klj} is the frequency of allele j to appear in locus l in population k

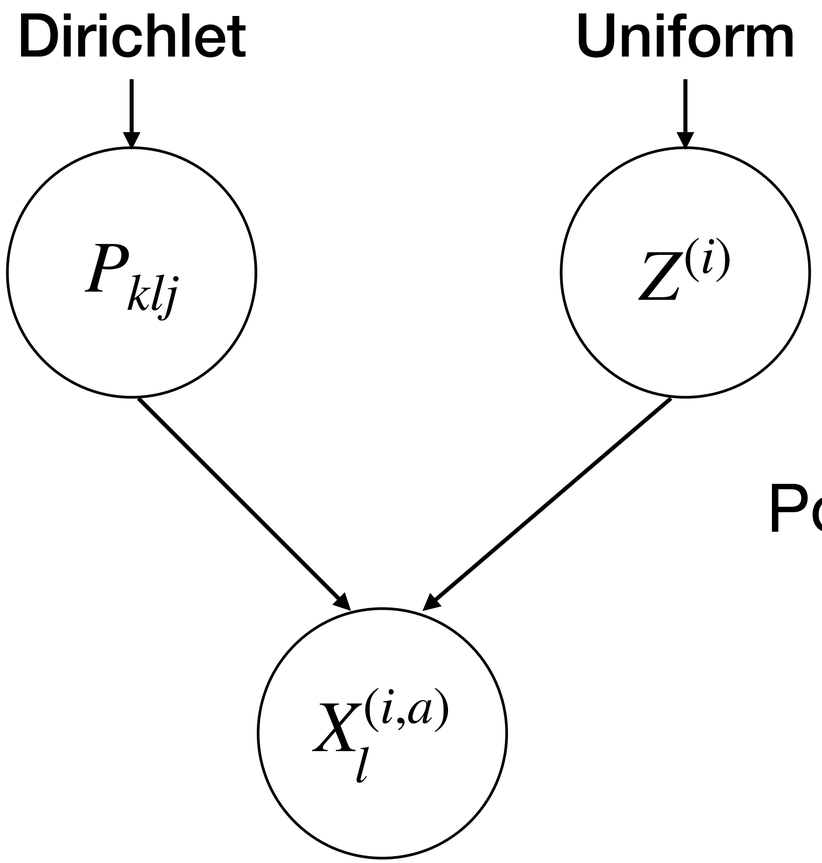
$$Pr(X_l^{(i,a)} = j \mid P, Z) = p_{Z^{(i)}lj}$$

$i = 1,\ldots,N$

$k = 1,\ldots,K$

$j = 1,\ldots,J_l$

Basic model (no admixture)



Likelihood: $Pr(X_l^{(i,a)} = j \mid P, Z) = p_{Z^{(i)}lj}$

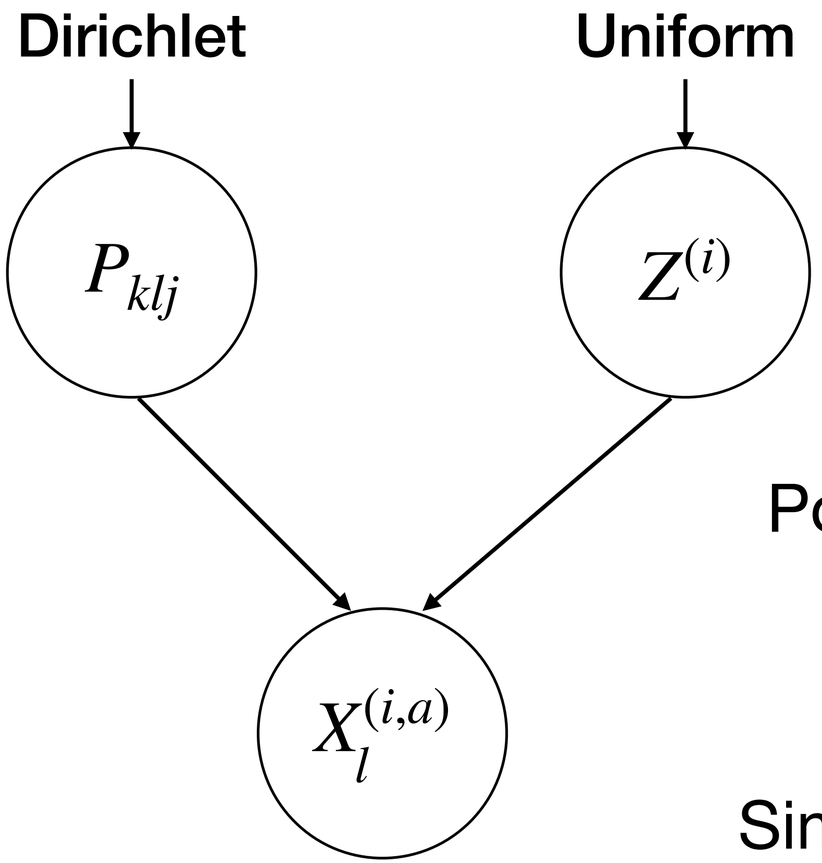
Prior: $P_{kl.} \sim Dir(\lambda_1, \ldots, \lambda_{J_l})$

$Z^{(i)} \sim Unif([1, \ldots, K])$

Posterior: $P_{kl.} \mid X, Z \sim Dir(\lambda_1 + n_1, \ldots, \lambda_{J_l} + n_{J_l})$

$Pr(Z^{(i)} \mid X, P) \propto Pr(X^{(i)} \mid P, Z^{(i)}) Pr(Z^{(i)})$

Basic model (no admixture)



Likelihood: $Pr(X_l^{(i,a)} = j \mid P, Z) = p_{Z^{(i)}lj}$

Prior: $P_{kl.} \sim Dir(\lambda_1, \ldots, \lambda_{J_l})$

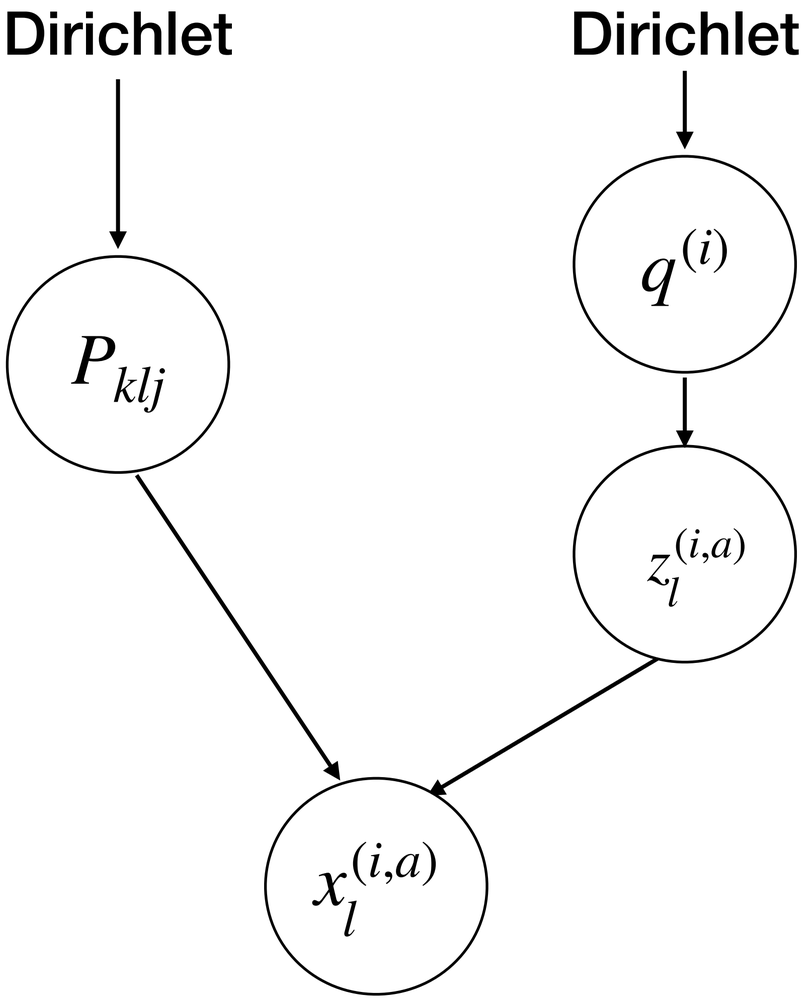
$Z^{(i)} \sim Unif([1, \ldots, K])$

Posterior: $P_{kl.} \mid X, Z \sim Dir(\lambda_1 + n_1, \ldots, \lambda_{J_l} + n_{J_l})$

$Pr(Z^{(i)} \mid X, P) \propto Pr(X^{(i)} \mid P, Z^{(i)}) Pr(Z^{(i)})$

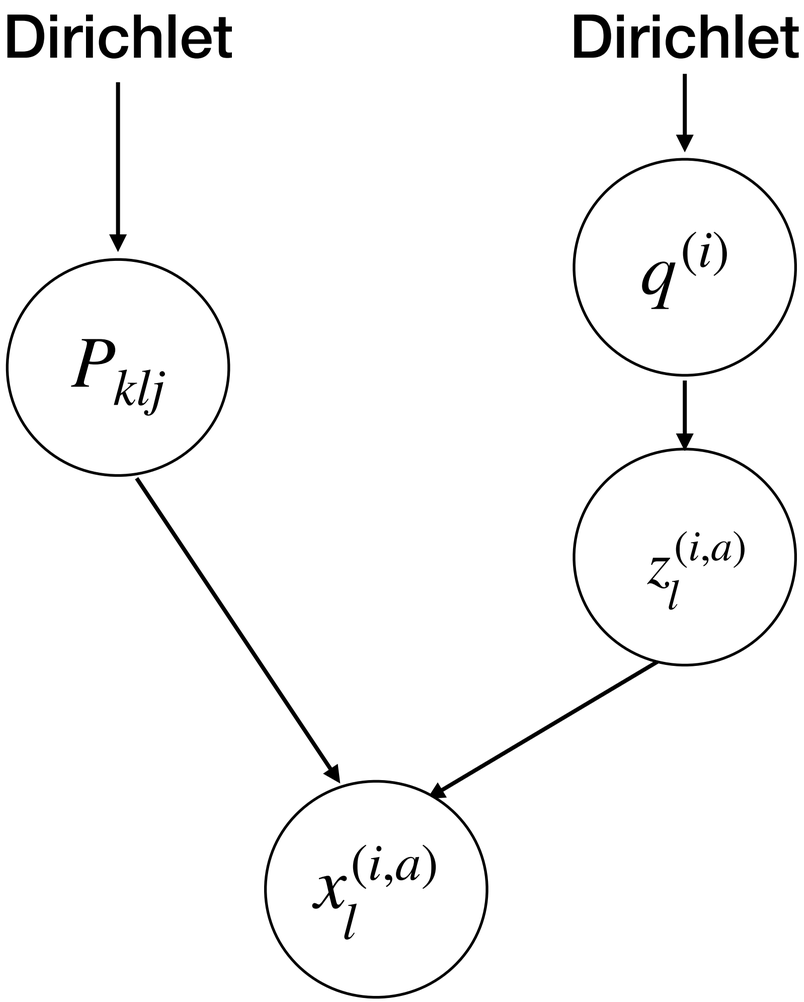
Simple and intuitive: But there may be person having multi-population origin (Admixture).

Admixture model



- $z_l^{(i,a)}$ the population of origin of allele copy $x_l^{(i,a)}$
- $q^{(i)} = (q_k^{(i)})$ is the proportion of gene of the i-th individual belongs to the population k
- P_{klj} is the frequency of allele j to appear in locus l in population k

Admixture model



Likelihood: $Pr(X_l^{(i,a)} = j \mid P, Z_l^{(i,a)}) = p_{Z^{(i,a)}lj}$

$$Pr(z_l^{(i,a)} = k \mid q) = q_k^{(i)}$$

Prior: $P_{kl.} \sim Dir(\lambda_1, \ldots, \lambda_{J_l})$

$$q^{(i)} \sim Dir(\alpha, \ldots, \alpha)$$

Posterior: $P_{kl.} \mid X, Z \sim Dir(\lambda_1 + n_1, \ldots, \lambda_{J_l} + n_{J_l})$

$$q^{(i)} \sim Dir(\alpha + m_1, \ldots, \alpha + m_K)$$

$$Pr(Z_l^{(i,a)} = k \mid X, P) \propto Pr(X^{(i)} \mid P, Z^{(i)}) q_k^{(i)}$$

Fit to model in data

- In the model, we need to choose the number of population K in the beginning -> can do by estimating $Pr(X \mid K)$ for each K and take the largest one. However, choosing different K's can lead to different insights about the sample. We will see it in the next part.
- A natural interest is making inference about $Q = (q_k^{(i)})$, where $q_k^{(i)}$ is the proportion that the i-th individual belongs to the population k.
- We will illustrate the model by applying it into simulated data and real world data.

3. Examples and Applications of STRUCTURE

- Simulated data
- Taita Thrush data
- The genetic structure of human populations. N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky and M.W. Feldman, 2002. Science, 298: 2381-2385. (and technical comment, 2003)

Example

Simulated data set

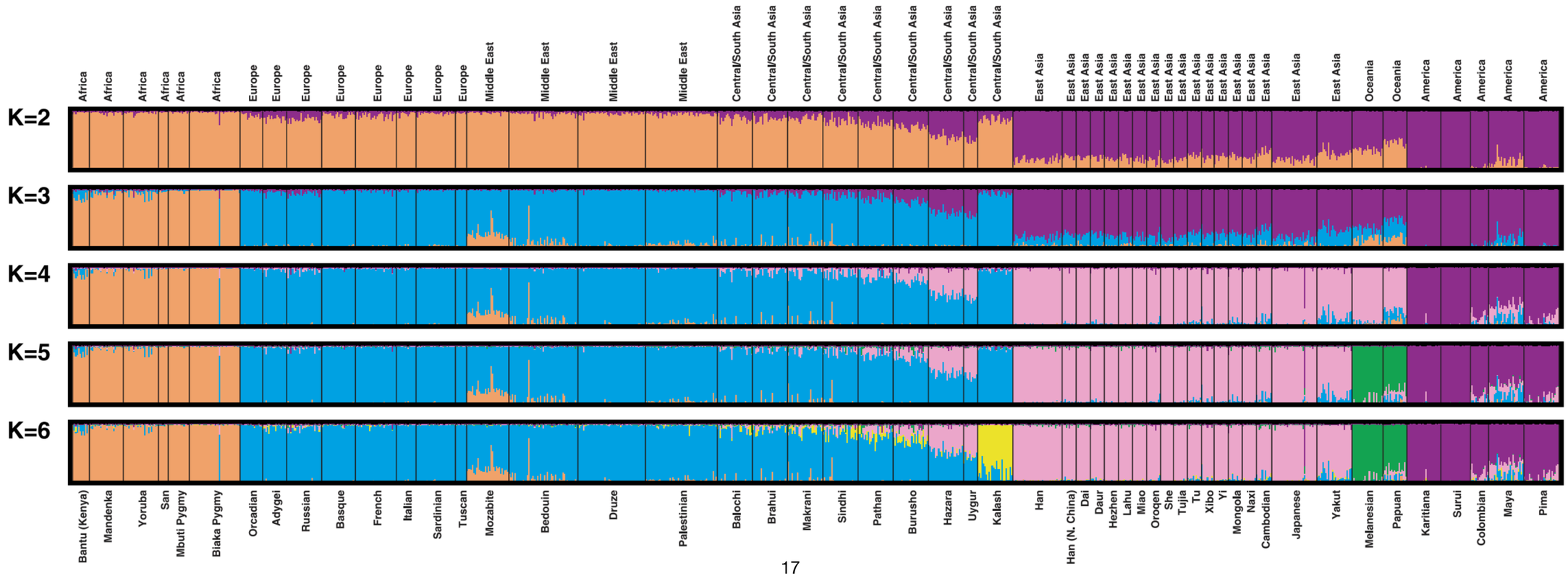
Example

Thrush data set

Example

The genetic structure of human populations

- Study the population structure using genotype at 377 microsatellite loci in 1056 individual from 52 populations.



4. Extensions of the model

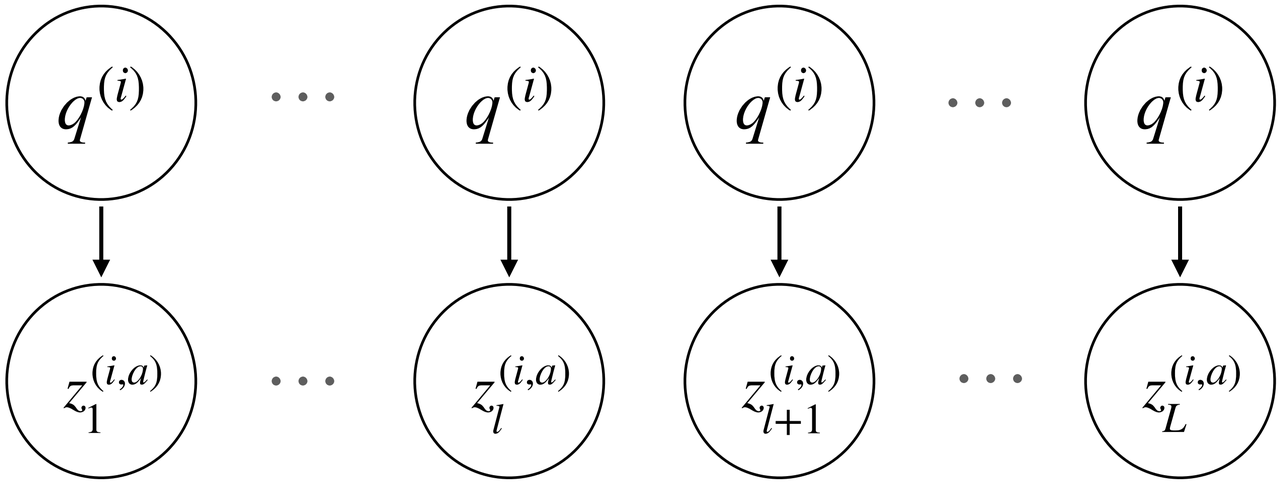
Linkage Disequilibrium:

Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies — 7725 — 2003
D Falush, M Stephens, JK Pritchard
Genetics 164 (4), 1567

Using the VIB for faster calculation:

fastSTRUCTURE: variational inference of population structure in large SNP data sets — 754 — 2014
A Raj, M Stephens, JK Pritchard
Genetics 197 (2), 573-589

Include Linkage Disequilibrium in model



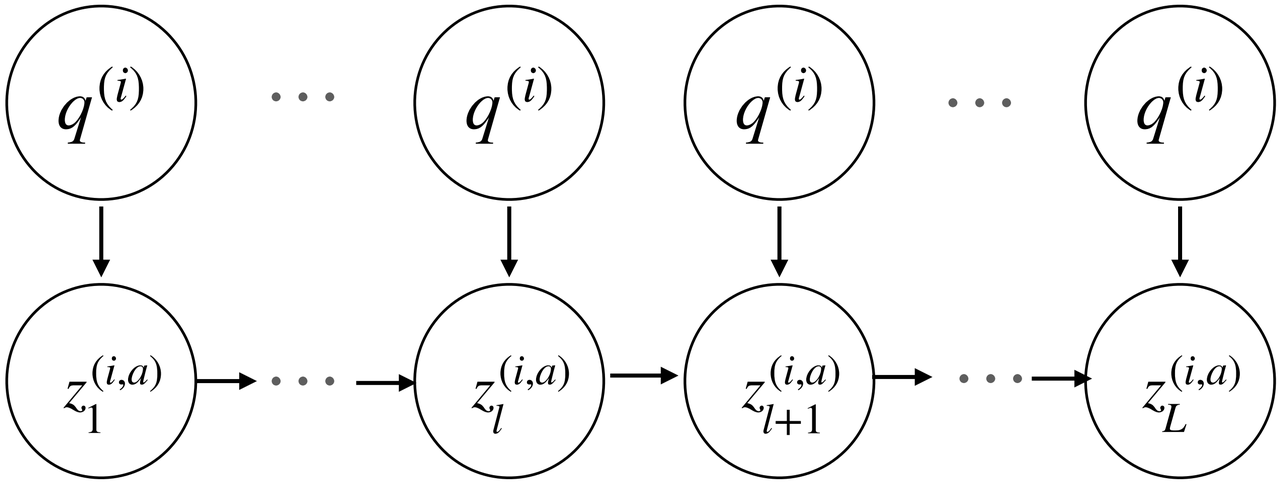
Iid for every locus $l = 1,\ldots,L$

In the admixture model, we assume

$$Pr(z_l^{(i,a)} = k \mid q) = q_k^{(i)} \quad \text{(iid)}$$

But realistically, because each chromosome is a set of “chunks” that are derived from ancestral population. It may create dependency between loci [4].

Include Linkage Disequilibrium in model



Assume Markov chain instead of iid -> Hidden Markov model

$$Pr(z_{l+1}^{(i,a)} = k' \mid z_{l+1}^{(i,a)} = k, r, Q) =$$

$$\begin{cases} \exp(-d_l r) + (1 - \exp(-d_l r)) q_{k'}^{(i)} \\ \qquad \text{if } k = k', \\ (1 - \exp(-d_l r)) q_k^{(i)} \\ \qquad \text{otherwise} \end{cases}$$

Where d_l denotes the genetic distance from locus l to locus $l + 1$.

We still can run MCMC to get the posterior.

Variational Bayes Inference

- MCMC can take a lot of time to run: Running Thrush data in a 4 core i5 machine (10^5 iterations) takes ~ 3 hours. Last week's seminar, Laura Kubatko said she ran MCMC for 2 weeks in a supercomputer to make posterior inference.

- Instead of using MCMC to estimate the posterior $Pr(Z, P, Q \mid X)$, we try to find a distribution $q(Z, P, Q)$ minimizing the KL distance to it [5]

$$
\begin{aligned}
q^* &= \min_{q \in \mathcal{Q}} KL(q(P, Q, Z), p(P, Q, Z \mid X)) \\
&= \min_{q \in \mathcal{Q}} (\log p(X) - \mathscr{E}_X(q(P, Q, Z)))
\end{aligned}
$$

Which is equivalent to maximize the ELBO $\mathscr{E}_X(q(P, Q, Z))$.

Variational Bayes Inference

- The Evidence Lower Bound (ELBO) has the form

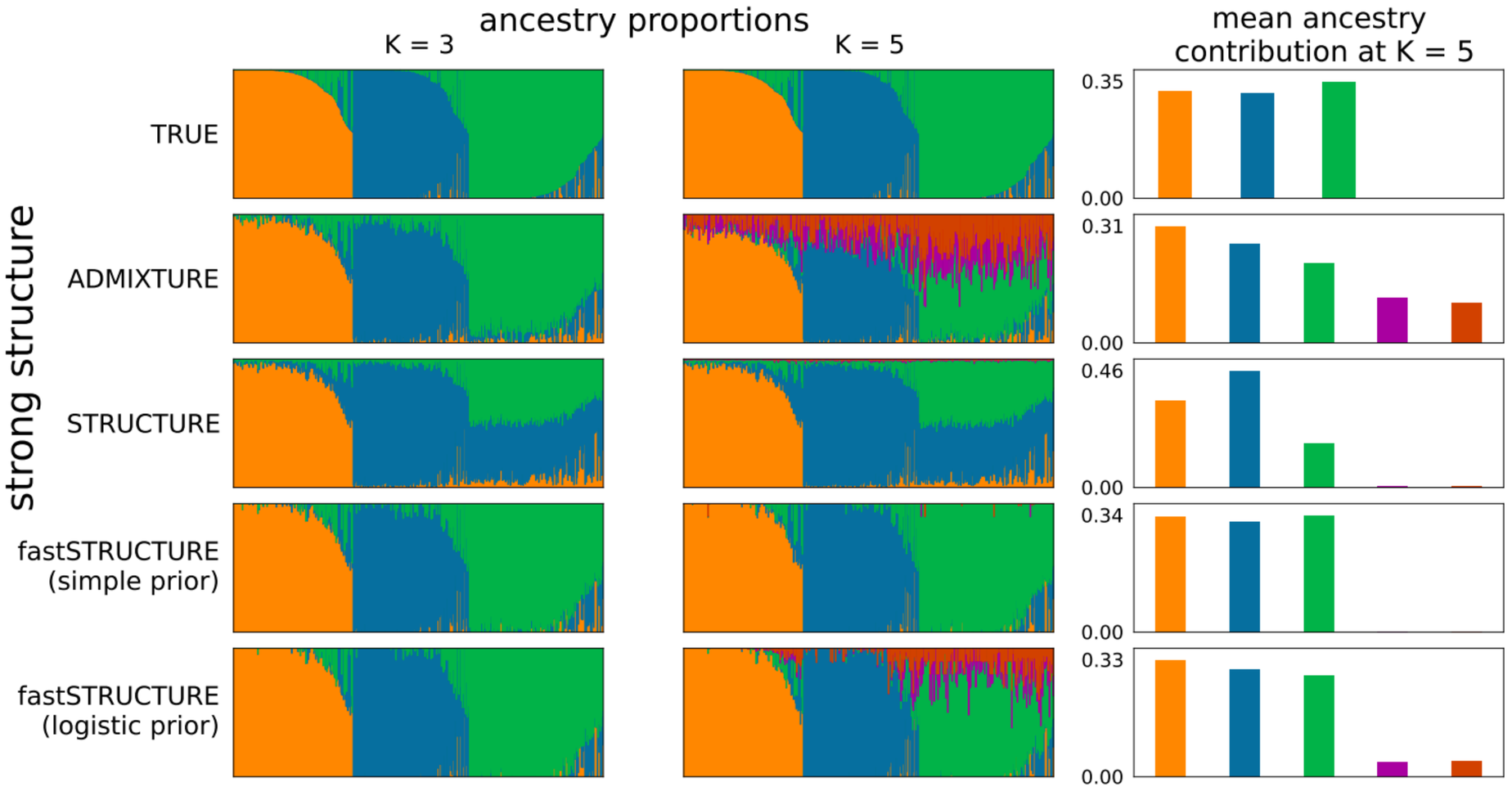
$$\mathscr{E}_X(q(P,Q,Z)) = \int \frac{p(P,Q,Z,X)}{q(P,Q,Z)} q(P,Q,Z) d(P,Q,Z)$$

And the set $\mathscr{Q}$ is set of all distribution such that Z and (P,Q) are independent.

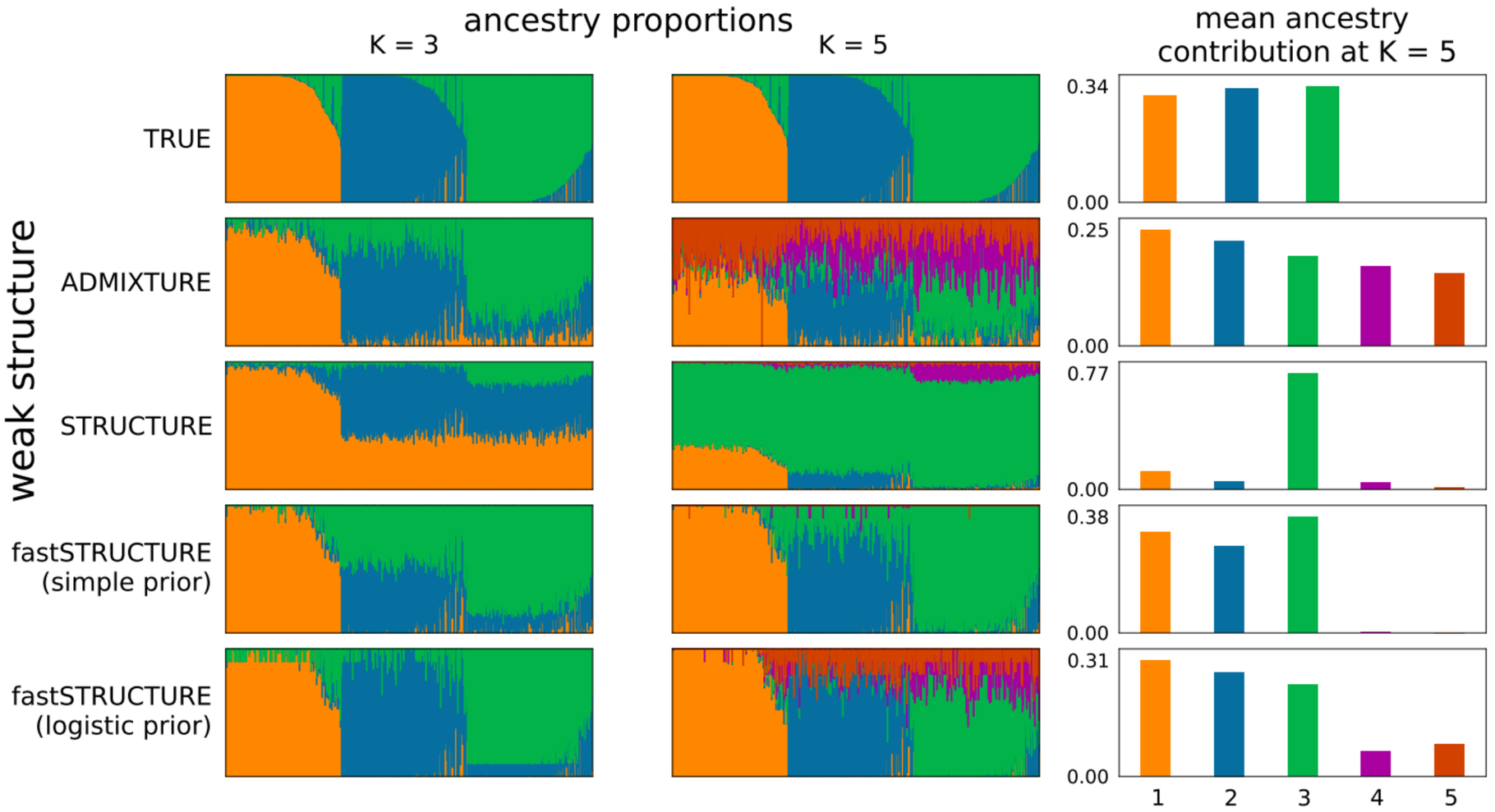
- This is not true in biology. But it keeps the computation tractable. To maximize the ELBO, they use Cauchy-Barzilai-Borwein method.

- We trade the complexity of the posterior for the computation time. Let's see a simulated example to compare STRUCTURE and this method (fastSTRUCTURE)

Variational Bayes vs. other model: An example



Variational Bayes vs. other model: An example



5. Discussion, Comments, questions

Reference

[1] Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly GENETICS June 1, 2000 vol. 155 no. 2 945-959

[2] John Novembre, GENETICS October 1, 2016 vol. 204 no. 2 391-393; https://doi.org/10.1534/genetics.116.195164

[3] https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/structure.html

[4] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003 Aug;164(4):1567-87. PMID: 12930761; PMCID: PMC1462648.

[5] Anil Raj, Matthew Stephens, Jonathan K. Pritchar, Variational Inference of Population Structure in Large SNP Datasets