HELSINKI UNIVERSITY OF TECHNOLOGY
NEURAL NETWORKS RESEARCH CENTRE

Inference of Population Structure Using Multilocus Genotype Data.

Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly

*additional material from:***Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies.**

presented by Jarkko Salojärvi, 10.2.2004

Structure of the presentation

- Prerequisites: Gibbs sampling, probability densities concerned.
- Data
- Applied models in increasing complexity:
  - Assumptions of the model, probability density
  - Obtaining samples from the pdf.
  - Results.
- General overview, conclusions.

Gibbs Sampling

Sampling from (multi-dimensional) joint probability distribution is difficult. An easier way to obtain samples is to construct a Markov chain as follows:

1. Select random initial values for parameters $\Theta = (\theta_1, \ldots \theta_r)$.
2. Sample $\theta_1^{(m)}$ from conditional pdf $p(\theta_1 | X, \theta_2^{(m-1)}, \ldots \theta_r^{(m-1)})$.
3. Sample $\theta_2^{(m)}$ from conditional pdf $p(\theta_2 | X, \theta_1^{(m)}, \ldots \theta_r^{(m-1)})$.
4. repeat from (2).

The chain will have a stationary distribution $p(\theta_1, \ldots \theta_r | X)$.
We can define very complex models and still do inference by using the samples from the model posterior.

Probability densities

Observations are discrete. Suitable conjugate exponential model:

- Exponential model: multinomial.

$$p(\mathbf{n}|\mathbf{\Theta}) = \binom{N}{n_1\ n_2\ \ldots n_K} \prod_{k=1}^{K} \theta_k^{n_k}\ ; N = \sum_k n_k$$

- Conjugate prior: Dirichlet.

$$\mathcal{D}(\alpha) = p(\mathbf{\Theta}|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\ ; \alpha_0 = \sum_k \alpha_k$$

$\alpha \to 0$ preference for $\Theta = (0 \ldots 0\ 1\ 0 \ldots)$, each k equally likely;$\alpha_k = 1$ uniform distribution.

Def. **The posterior will be of the same form as prior.**

Data

- X: Alleles of N (diploid) individuals (i) in L loci: $(x_l^{(i,1)}, x_l^{(i,2)})$.
- General assumption 1: Alleles (1,2) in certain loci are independent. (Hardy-Weinberg equilibrium.)
- General assumption 2: The measured loci are far from each other in the genome and can be considered independent (=complete linkage equilibrium).

Allele= any one of a number of alternative forms of the same gene occupying a given locus.

Locus, loci = A certain position in a chromosome, occupied by any of the alleles of the gene

Examples of inference problems

- What is(are) the population(s) of origin of a sample of individuals?
- Evolutionary relationships of populations?
- DNA fingerprinting: what is the probability of a false match?

The pros of generative modelling

- In distance-based clustering methods, the results depend on the metric. Often only visual evaluation of goodness can be made.
- A generative model describes the process which created the data. The differences in the data can be measured in terms of the differences of the parameters of the model.
- Bayesian framework:
  - Other useful information can be incorporated via priors.
  - Uncertainties within the model can be estimated.
  - Model selection criteria.

1. Model without admixture

The genotype $(x_l^{(i,1)}, x_l^{(i,2)})$ of each individual (i) originates from one of K populations.

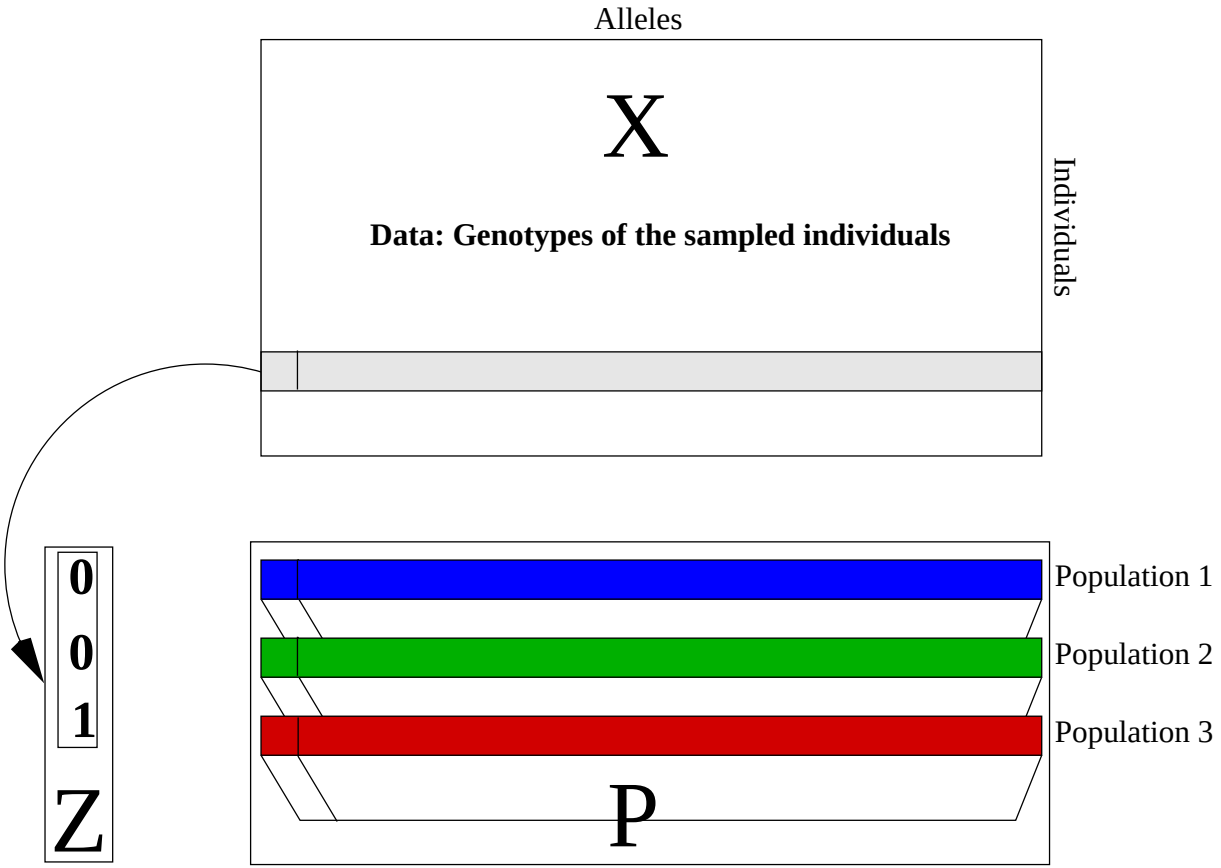
$=$ Hard clustering of samples into K clusters.

Model 1

X genotypes of the sampled individuals

Z (unknown) populations of origin of the individuals

P (unknown) allele frequencies in populations



Model 1 - description

- $p(Z, P|X) \propto p(X|P, Z)p(P)p(Z)$
- $p(z^{(i)} = k) = 1/K$; where $k = 1 \ldots K$.
- $p(p_{kl\cdot}) \sim \mathcal{D}(\lambda_1, \ldots \lambda_{J_l})$; where $l = 1 \ldots L$, and J_l is the number of distinct alleles in locus l.
  - Uniform prior: $\lambda_1 = \lambda_2 = \ldots = \lambda_{J_l} = 1$.
- $p(x_l^{(i,a)} = j|P, Z) = p(p_{z^{(i)}lj})$; $j = 1 \ldots J_l$

Model 1 - Gibbs sampling

1. Sample $P^{(m)}$ from $p(P|X, Z^{(m-1)})$:
   - $p_{kl\cdot}^{(m)} \sim \mathcal{D}(\lambda_1 + n_{kl1}, \ldots, \lambda_{J_l} + n_{klJ_l})$, where $n_{klj} = \#\left\{(i,a) \ : \ x_l^{(i,a)} = j \text{ and } \ z^{(i)} = k\right\}$
2. Simulate $z^{(i)}$ from:
   $p(z^{(i)} = k|X, P) = \frac{p(x^{(i)}|P, z^{(i)}=k)}{\sum_{k'} p(x^{(i)}|P, z^{(i)}=k')}$,
   where $p(x^{(i)}|P, z^{(i)} = k) = \prod_{l=1}^{L} p_{klx^{(i,1)}} p_{klx^{(i,2)}}$
   An equal prior $p(z^{(i)} = k) = 1/K$ is assumed.

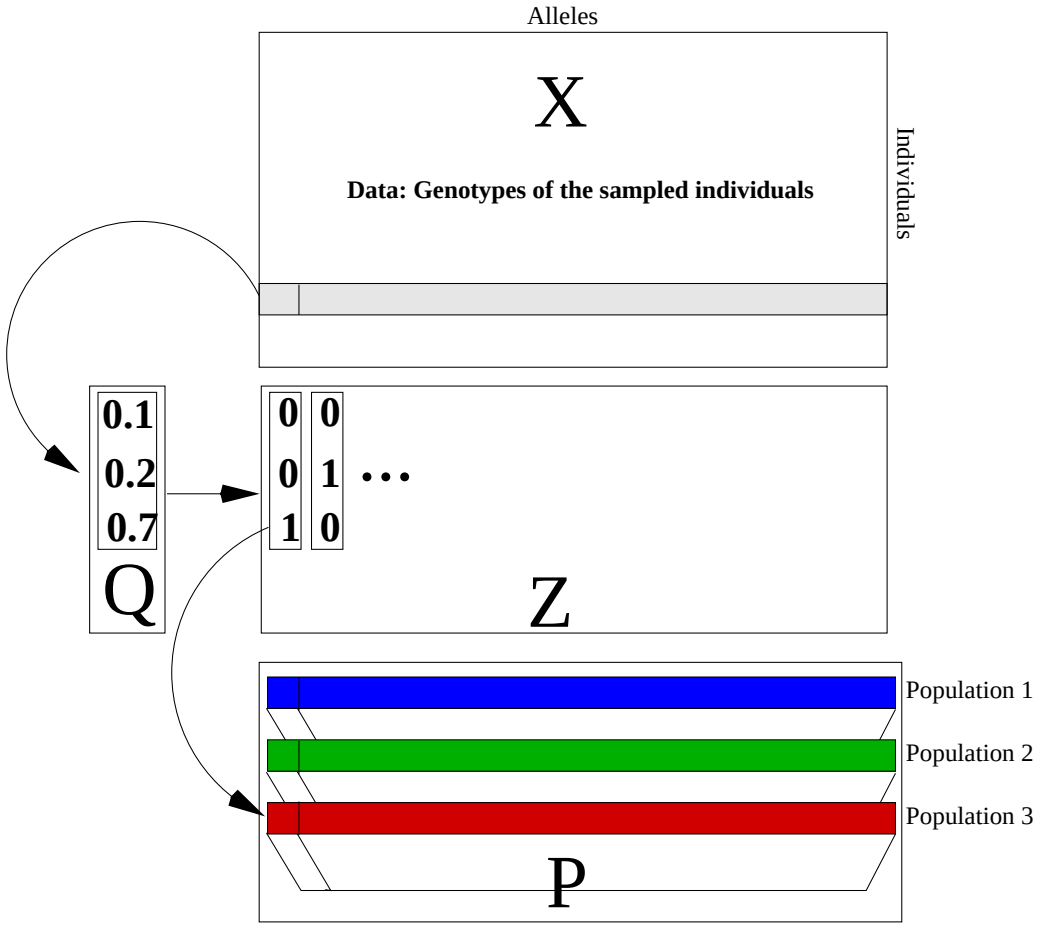
2. Model with admixture

The genotype of each individual is a mixture from populations.

= Probabilistic soft clustering of samples into K clusters.

- The original population of each loci l is defined individually.

Model 2


Alleles
X
Data: Genotypes of the sampled individuals
Individuals
0.1
0.2
0.7
Q
0 0
0 1
1 0
...
Z
Population 1
Population 2
Population 3
P

Model 2 - description

- $p(Z, P, Q|X) \propto p(X|P, Z, Q)p(Z|P, Q)p(P)p(Q)$.
- $p(x_l^{(i,a)} = j|Z, P, Q) = p(p_{z_l^{(i,a)}lj})$.
- $p(z_l^{(i,a)} = k|P, Q) = q_k^{(i)}$.
- $p(p_{kl\cdot}) \sim \mathcal{D}(\lambda_1, \ldots \lambda_{J_l})$; $\lambda_1 = \lambda_2 = \ldots = \lambda_{J_l} = 1$.
- $p(q^{(i)}) \sim \mathcal{D}(\alpha, \ldots, \alpha)$; $\alpha \sim Unif[0, 10]$.

Model 2 - Gibbs sampling

1. Sample $P^{(m)}$: $p(p_{kl\cdot}^{(m)}|X, Z^{(m-1)}) \sim \mathcal{D}(\lambda_1 + n_{kl1}, \ldots, \lambda_{J_l} + n_{klJ_l})$, where $n_{klj} = \#\left\{(i,a) \ : \ x_l^{(i,a)} = j \text{ and } \ z_l^{(i,a)} = k\right\}$.

2. Sample $Q^{(m)}$: $p(q^{(i)}|X, Z^{(m-1)}) \sim \mathcal{D}(\alpha + m_1^{(i)}, \ldots, \alpha + m_K^{(i)})$, where $m_k^{(i)} = \#\left\{(l,a) \ : \ z_l^{(i,a)} = k\right\}$

3. Sample $Z^{(m)}$:
   $p(z_l^{(i,a)} = k|X, P^{(m)}, Q^{(m)}) = \frac{q_k^{(i)} p(x_l^{(i,a)}|P, z_l^{(i,a)}=k)}{\sum_{k'} q_{k'}^{(i)} p(x_l^{(i,a)}|P, z_l^{(i,a)}=k')}$,
   where $p(x_l^{(i,a)}|P, z_l^{(i,a)} = k) = p_{klx_l^{(i,a)}}$.

4. Simulate proposal α' from $\mathcal{N}(\alpha, \sigma_\alpha^2)$. Reject if $\alpha' \leq 0$; otherwise accept with the appropriate Metropolis-Hastings probability.

Practical issues

- Due to label switching, there are $K!$ different modes in the posterior.
- MCMC methods often do not switch between modes $\Rightarrow$ we obtain an estimate of the posterior mode (usually undesirable; in clustering application this is what we want).
- Number of clusters K was selected using a model selection criterion based on DIC [Spiegelhalter et al. 99] (using quite heavy assumptions on the form of the posterior). Seems to work well, however.

Applications to data

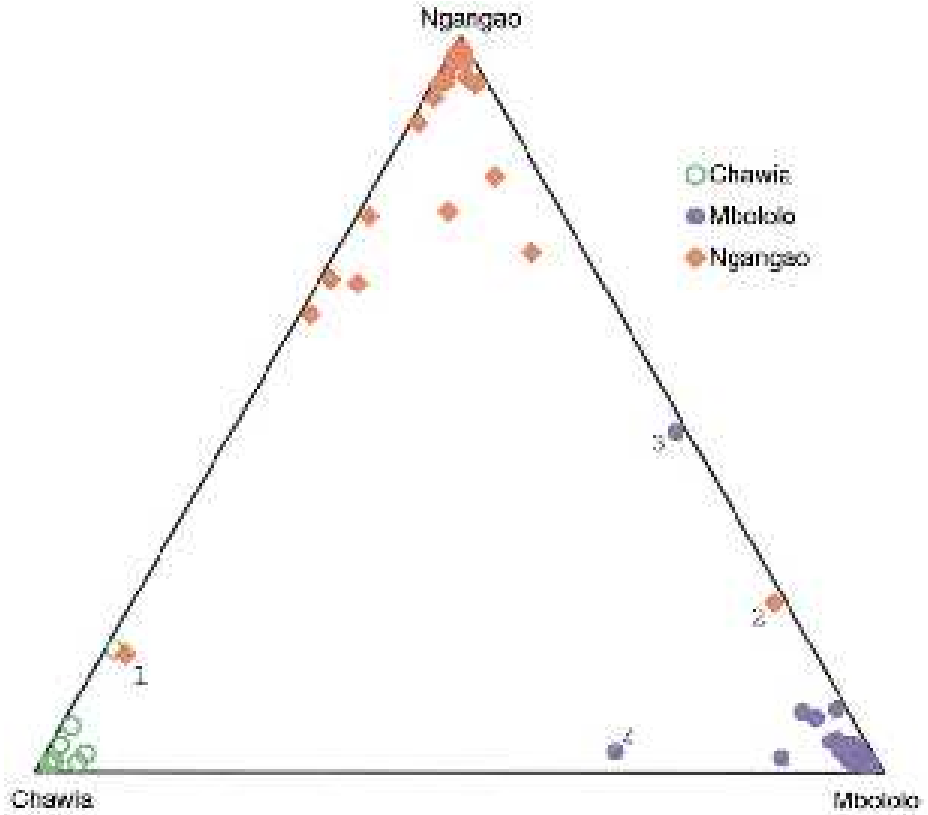
Simulated data – three cases:

- A single random-mating population of size N
- Two random-mating populations of size 2N, split from a single ancestral population. No migration.
- Admixture of populations. Two populations joined, sampled collected after two generations of random mating.

Gives highest probability to correct amount of clusters and assigns individuals to correct clusters.

Taita thrush (Turdus helleri)

- Each point shows the mean estimated ancestry (vector $q^{(i)}$) for an individual.
- Shown as distances from the corners.
- Individuals 1-4 appear to be outliers (immigrants?)



Beyond Basic Model

Modifications can be made by constructing more informative priors.(hierarchical priors).

Model 3. Geographic location.

Variant for estimation of immigrants.

The geographic location $g^{(i)} \in [1 \ldots K_g]$ of the individuals is taken into account.

- Place a hierarchic prior to population proportions:
  $q^{(i)}_{g^{(i)}} = 1,\ q^{(i)}_k = 0;\ (k \neq g^{(i)})$ with probability $1 - \nu$;
  $q^{(i)}_{g^{(i)}} = 1 - 2^{-t},\ q^{(i)}_j = 2^{-t};\ q^{(i)}_k = 0;\ (k \neq g^{(i)}, j)$ for each $j \neq g^{(i)}$ with probab. $\frac{2^t \nu}{(K_g - 1) \sum_{T=0}^{G} 2^T}$,

where $t \in [0 \ldots G]$, and G is the number of generations. The value of ν is an informed guess (very small).

Model 4. Correlated markers.

DNA is inherited in large chunks. Therefore nearby markers are usually from the same parent.

- $p(z_1^{(i)} = k|r, Q) = q_k^{(i)}$
- $p(z_{l+1}^{(i)} = k'|z_l^{(i)} = k, r, Q) =$

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

z's along choromosome form a Markov chain. d_l is the (known) distance between markers, r the rate of mixing ($\log r \sim Unif$).

Model 5. Correlated allele frequencies.

Allele frequencies in closely related populations are often similar.

- $p_{Al\cdot} \sim \mathcal{D}(\lambda_1, \ldots, \lambda_{J_l})$.
- $p_{kl\cdot} \sim \mathcal{D}\left(p_{Al1}\frac{1-F_k}{F_k}, \ldots, p_{AlJ_l}\frac{1-F_k}{F_k}\right)$
- F_k effective size of population k during the time since divergence from ancestral population. Prior $\sim \Gamma$, truncated at 1.
- For small F_k, we are close to ancestor population A. The closer to 1 the F_k is, the further we are from A. $\Rightarrow$ phylogenetic inference.

Conclusions

- Model is the same as LDA and mPCA (next lecture's topic).
- MCMC sampling instead of variational approximations.
- More extensive: model order selection (number of populations K) plus many variations.
- Smaller amount of clusters K than applications of LDA.
- Better than mPCA (Buntine, private communication). (4x) Slower but more accurate.