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
Daniel Falush, Matthew Stephens and Jonathan K. Pritchard

Summary by: Jarkko Salojärvi

11th February 2004

1 Background

We will begin this summary article by a brief review of Gibbs sampling and a conjugate-exponential model which is suitable for the data.

1.1 Gibbs Sampling

Sampling from (multi-dimensional) joint probability distribution is often very 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 has a stationary distribution $p(\theta_1, \ldots \theta_r|X)$.

Usually the initial guess is far from the area where the posterior mass is concentrated. Therefore the sampling procedure is run for several thousands of iterations (so called *burn-in* stage) in order to drive the chain to the correct area. The length of the burn-in stage naturally depends on the complexity of the model. Due to the Markov chain property of the sampling method, consecutive samples are correlated. Therefore samples from the chain are taken at some constant interval (i.e. every cth sample). The method is known as *thinning*.

The benefit of Gibbs sampling (as well as other Markov Chain Monte Carlo methods) is that we can define very complex models and still do inference by using the samples from the model posterior. The difficult part is to evaluate

when the sampling chain has reached the area of the posterior mass, and when the correlations between samples comes from the true distribution instead of the Markov Chain property (what is the minimal value of c).

1.2 Probability densities concerned

In population genetics, the data consist of a matrix of categorical values; the types of genes the individuals have at certain loci. For modeling this kind of data, a suitable distribution within the exponential family is the multinomial,

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

with a conjugate Dirichlet prior,

$$\mathcal{D}(\alpha) = p(\mathbf{\Theta}|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\ ; \alpha_0 = \sum_k \alpha_k. \tag{2}$$

The Dirichlet distribution gives preference for $\Theta = (0 \ldots 0\ 1\ 0 \ldots)$, where each k is equally likely, for values of $\alpha \to 0$. When $\alpha_k = 1$, we get a *noninformative prior*, uniform distribution. Increasing the value of α places more weight (compared to likelihood) on the prior, introducing more *pseudo data* to the model. The hyperparameter values α_k are usually equal. This way we do not prefer any particular k, which is why the prior is then said to be *symmetric*.

By definition[4] the posterior of a conjugate model is of the same functional form as prior.

2 Data

In population genetics the design matrix X consists of alleles[1] of N (diploid) individuals (i) in L loci[2]: $(x_l^{(i,1)}, x_l^{(i,2)})$. It is generally assumed that alleles $x_l^{(i,a)}$ in certain loci are independent (in population genetics, Hardy-Weinberg equilibrium). Another assumption is that the measured loci are far from each other in the genome and can be considered independently inherited (complete linkage equilibrium). In modeling this means that alleles in loci are independent of each other. In more advanced models (discussed in Section 5) these assumptions are relaxed somewhat.

Possible inferences from the data:

- 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?

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

[2] Locus, loci(pl.) = A certain position in a chromosome, occupied by any of the alleles of the gene.

3 Models

3.1 Pros of generative models

Different clustering methods can be roughly divided into two categories: distance-based methods and model-based methods.

A distance-based clustering method begins by computing a pairwise distance matrix from the data. After obtaining the distance matrix, the data may be clustered for example by forming a neighbor-joining tree. However, the clustering results are arbitrary, depending on the chosen metric. Often only visual evaluation of goodness of clustering can be made.

A generative model on the other hand describes the process which created the data. Therefore, differences in the distribution of data can be measured in terms of the differences in the parameter values that produced the corresponding distributions. The model therefore provides the correct metric for measuring distances, assuming of course that we have the correct model.

The use of generative models makes it possible to apply the Bayesian framework. This enables us to incorporate useful expert knowledge into the model via priors, estimate the uncertainties within the model, and use (somewhat) objective criteria for selecting the optimal complexity of our model.

3.2 Model without admixture

In a simple model, the genotype $(x_l^{(i,1)}, x_l^{(i,2)})$ of each individual (i) originates from one of K populations (model setup shown in Fig. 1).

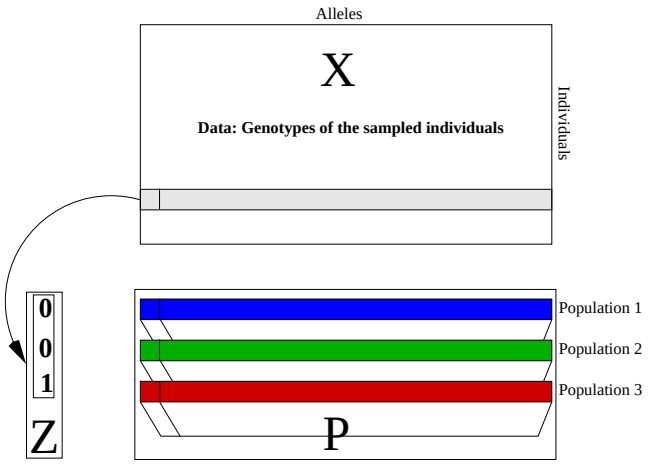

Figure 1: Model without admixture. The X matrix consists of individuals (rows) genotyped at L loci (columns). Each individual is originated from one population, indicated by Z. Each population is described by multinomial distributions at L loci. The parameters of the distributions are collected in matrix P.

3.2.1 Model description

Model 1 - description

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

In the basic version we have a uniform prior: $\lambda_1 = \lambda_2 = \ldots = \lambda_{J_l} = 1$.

3.2.2 Sampling

Model 1 - Gibbs sampling

1. Sample $P^{(m)}$ from $p(P|X, Z^{(m-1)})$:
   - $p_{kl\cdot}^{(m)} \sim \mathcal{D}(\lambda_1 + n_{kl1}, \dots, \lambda_{J_l} + n_{klJ_l})$, 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.

3.3 Model with admixture

In a more advanced model, the genotype of each individual is a mixture of populations. The model performs a probabilistic soft clustering of individuals (i) into K clusters, and defines also the population of origin of each locus l of individuals. The model setup is shown in Fig. 2.

3.3.1 Model description

Model 2 - description

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

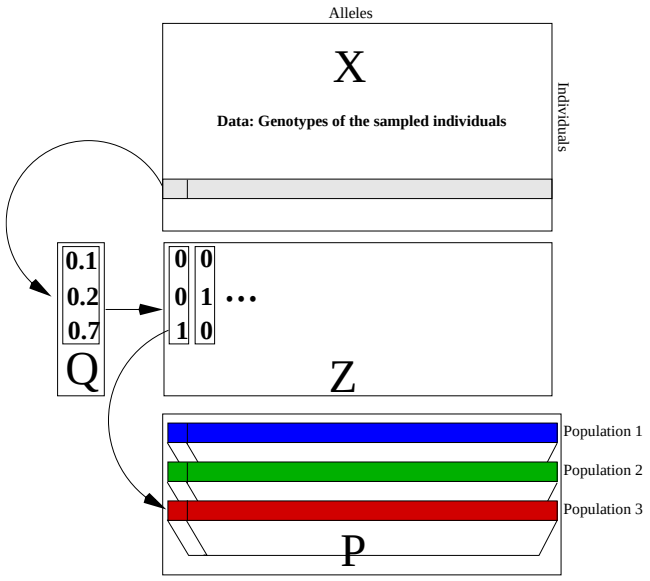

Figure 2: Model with admixture. The X matrix consists of individuals (rows) genotyped at L loci (columns). Each individual is originated from the populations with the proportions indicated by Q. The population of origin for each loci is sampled from a multinomial distribution with parameter values Q. As in the simpler model, each population is again described by multinomial distributions at L loci. The parameters of the distributions are collected in matrix P.

3.3.2 Sampling

Model 2 - Gibbs sampling

1. Sample $P^{(m)}$: $p(p^{(m)}_{kl\cdot}|X, Z^{(m-1)}) \sim \mathcal{D}(\lambda_1 + n_{kl1}, \ldots, \lambda_{J_l} + n_{klJ_l})$,
   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.

3.4 Practical issues

Due to label switching, there are $K!$ different modes in the posterior. However, the authors argue that the MCMC methods often do not switch between modes,

and therefore we obtain an estimate of a posterior mode. In clustering this is exactly what we want.

The number of clusters K was selected using a model selection criterion based on Deviance Information Criterion, DIC [7, 8]. The variant used in the paper assumes that the posterior is Gaussian. The criterion seems to work well, as it gives correct results with simulated data.

4 Applications to data

In [5] the model was first tested with simulated data. The simulations considered three scenarios:

- 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, samples collected after two generations of random mating.

The model selection criterion assigns highest probability to a model with the correct amount of clusters, and the model succeeds in assigning individuals to correct clusters. After validation, the model was applied to two real life data sets, Taita thrush and humans from Africa and Europe (results not discussed here). The first data set consists of 155 individuals of taita thrush (*turdus hellerii*), sampled from four locations in southeast Kenya. Each individual was genotyped at seven microsatellite loci. The estimated mean proportions of populations of individuals is shown in Figure 3.

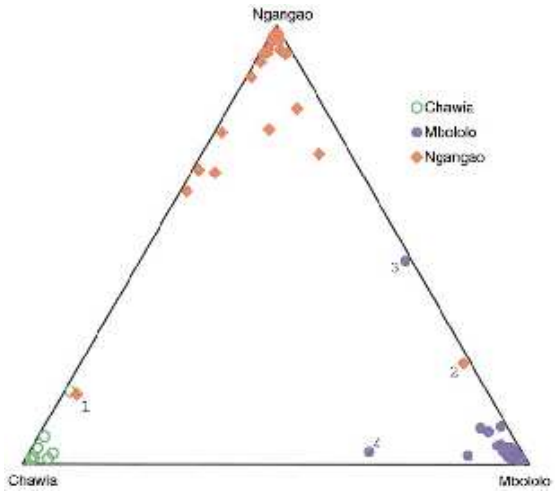

Figure 3: Each point shows the mean estimated ancestry (vector $q^{(i)}$) for an individual. Shown as distances from the corners. Individuals 1-4 appear to be outliers (possible immigrants).

Four of the samples showed considerable difference from the genotype of the main population in the area (shown in Fig. 3). In order to estimate whether they or their parents are immigrants from other populations, further refinements with the model were carried out (discussed in Section 5.1).

5 Beyond Basic Model

Modifications to the basic model can be made by constructing more informative priors (hierarchical priors). In this section we will discuss three variants. The first variant, introduced in [5], takes into account the geographic sampling location of the individuals. The next two variants, introduced in [3], relax the independence assumptions of the models.

5.1 Modelling geographic location

In a further refinement, the geographic sampling location $g^{(i)} \in [1 \dots K_g]$ of the individuals is taken into account. For this means we may place a hierarchic prior on population proportions:

$$q^{(i)}_{g^{(i)}} = 1,\ q^{(i)}_k = 0;\ (k \neq g^{(i)})$$

with probability $1 - \nu$, and

$$q^{(i)}_{g^{(i)}} = 1 - 2^{-t},\ q^{(i)}_j = 2^{-t};\ q^{(i)}_k = 0;\ (k \neq g^{(i)}, j)$$

for each $j \neq g^{(i)}$with probab. $\frac{2^t \nu}{(K_g - 1)\sum_{T=0}^{G} 2^T}$, where $t \in [0 \dots G]$, and G is the number of generations. The value of ν is an informed guess (a small value).

5.2 Linked markers.

DNA is actually inherited in large chunks. Therefore nearby markers are usually from the same parent. We may incorporate this to the model by constructing a Markov Chain of the linked loci:

- $p(z^{(i)}_1 = k | r, Q) = q^{(i)}_k$
- $p(z^{(i)}_{l+1} = k' | z^{(i)}_l = k, r, Q) = \begin{cases} \exp(-d_l r) + (1 - \exp(-d_l r)) q^{(i)}_{k'} & \text{if } k' = k \\ (1 - \exp(-d_l r)) q^{(i)}_k & \text{otherwise,} \end{cases}$

where d_l is the (known) distance between markers, and r is the rate of mixing ($\log r \sim Unif$).

Notice that the model has similarities to hidden Markov models (HMMs). The $z^{(i)}$ (cf. Fig. 2) can be seen as the realization of the (hidden) state sequence each for individual (i). The transition probabilities in the model are defined by $q^{(i)}_k$, r and distances between loci, d_l. The P matrix is a multinomial emission distribution. The difference to HMMs is that the emission distribution is not stationary, but varies along the state sequence. However, the probabilities of each state can still be computed using the conventional forward-backward - algorithm [6]. These probabilities can then be used to obtain samples of state sequences.

5.3 Correlated allele frequencies.

Allele frequencies in closely related populations are often similar. This may be modelled by constructing a hierarchical prior:

- $p_{Al\cdot} \sim \mathcal{D}(\lambda_1, \ldots, \lambda_{J_l})$.
- $p_{kl\cdot} \sim \mathcal{D}\left(p_{Al1}\frac{1-F_k}{F_k}, \ldots, p_{AlJ_l}\frac{1-F_k}{F_k}\right)$.

F_k can be interpreted as effective size of population k during the time since divergence from ancestral population. It has a prior $\sim \Gamma$, truncated at 1. The parameter F_k defines our distance from ancestor population A. For small F_k we are close to ancestor population A, and have a strong prior, whereas the closer to 1 the F_k gets, the further we are from A, and have a weaker prior strength. The use of this kind of model enables phylogenetic inference.

6 Conclusions

The model introduced in [5, 3] is the same as Latent Dirichlet Allocation (LDA) [1] and multinomial Principal Component Analysis (mPCA) [2]. The difference is how the models are optimized. In [5] MCMC sampling was implemented, instead of using variational approximations as in [1, 2]. As a benefit of Gibbs sampling, the posterior estimate is more accurate than in mPCA, resulting in better performance.

The model is also extended in several ways, such as model order selection (number of populations K) and introducing several informative priors. The extensions are clearly motivated by the nature of the data. For example, the data in population genetics usually consists of a smaller amount of clusters K than applications of mPCA (such as text clustering), making model selection computationally more feasible.

References

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

[2] W. Buntine. Variational extensions to EM and multinomial PCA. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the ECML'02, 13th European Conference on Machine Learning*, Lecture Notes in Artificial Intelligence 2430, pages 23–34. Springer, Berlin, 2002.

[3] D. Falush, M. Stephens, and J. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

[4] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, FL, 1995.

[5] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[6] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[7] D. Spiegelhalter, N. Best, B. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002.

[8] D. J. Spiegelhalter, N. G. Best, and B. P. Carlin. Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical Report 98–009, Division of Biostatistics, University of Minnesota, 1998.