Computational Methods in Population Genetics

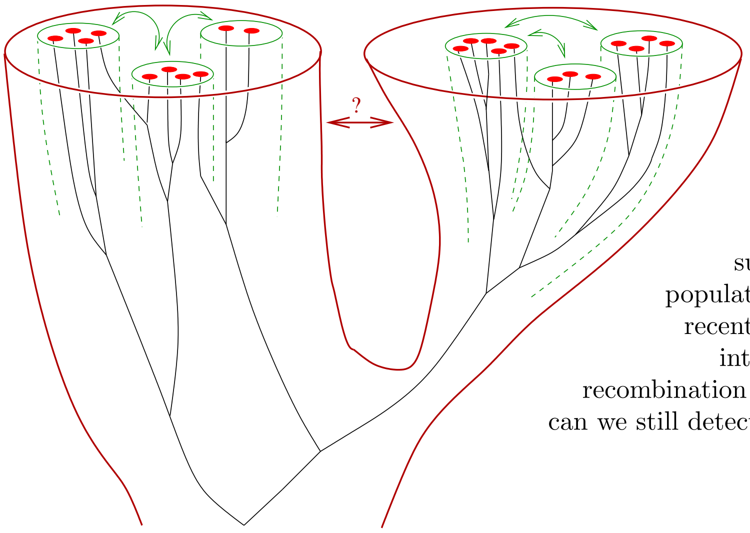
Dirk Metzler

Winter Semester 2011/2012, updated as at Oct 27, 2011

Contents

1 Examples

Complex Demography



substructure
population growth
recent speciation
introgression?
recombination within loci
can we still detect selection?

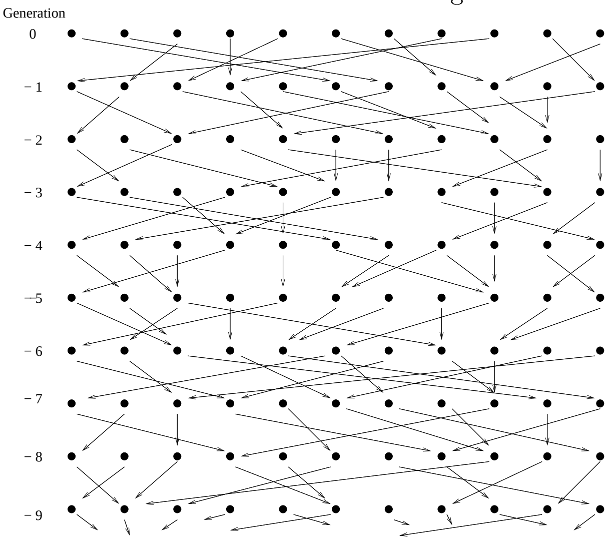
2 Wright Fisher model and Kingman's Coalescent

Basic assumptions of the Wright Fisher model

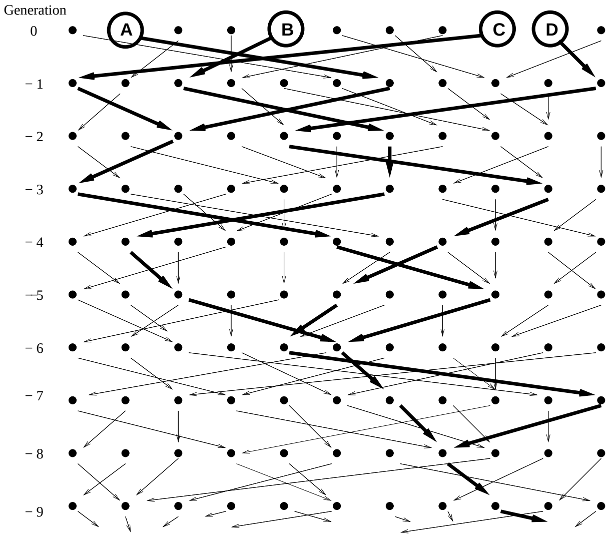
- non-overlapping generations
- constant population size
- panmictic
- neutral (i.e. no selection)
- no recombination
- N diploid individuals $\rightsquigarrow$ population of $2N$ haploid alleles (in case of autosomal DNA)

Wright Fisher model

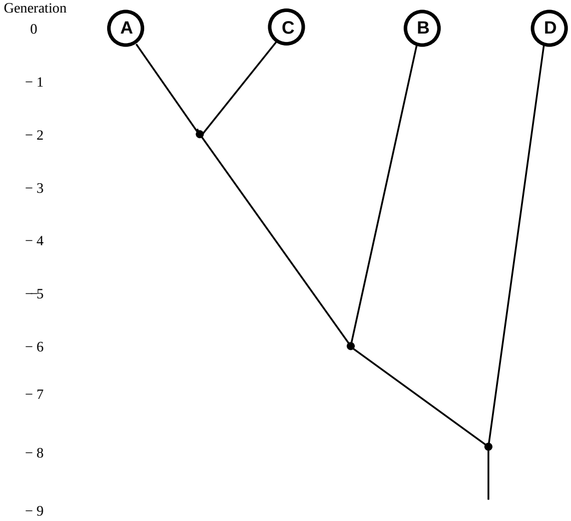
Each allele chooses an ancestor in the generation before.



Samples are assumed to be taken purely randomly from the population.



This induces a specific random distribution for the genealogies of the sampled alleles.

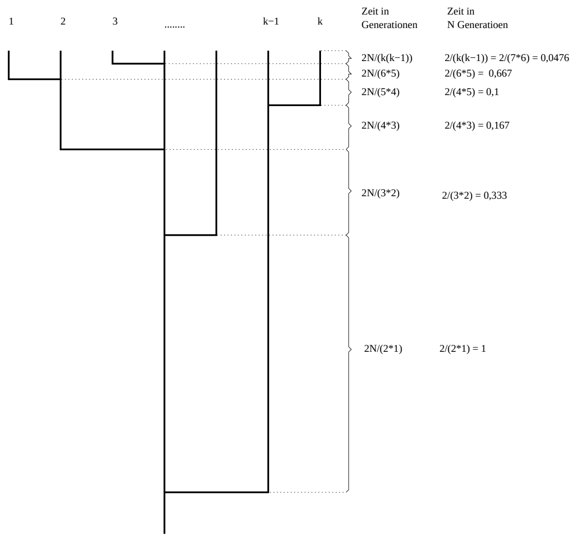


Haploid population of size N_e
Average time until two ancestral lineages coalesce: N_e generations.
Scale time: (1 time unit) = (N_e generations) $\Rightarrow$ pairwise coalescence rate = 1
$\mu :=$ mutation rate per generation

$$\theta := 2N_e \cdot \mu$$

is the expected number of mutations between 2 random individuals
Let $N_e \longrightarrow \infty$

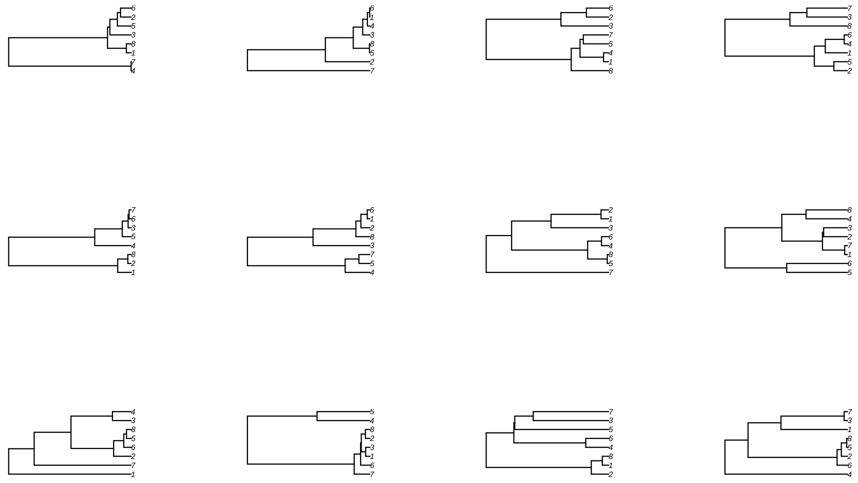
The Kingman Coalescent



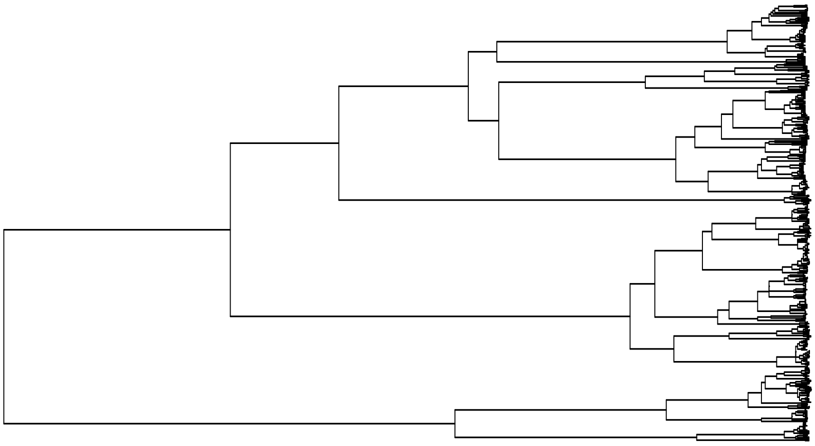
$\mathbb{E}(\text{total length})$

$$= 2 \cdot \sum_{i=1}^{k-1} 1/i$$

typical coalescent trees for $n = 8$:



simulated coalescent tree with $n = 500$:



3 Estimators for θ and Tajima's π

Two estimators of θ

θ_π ("Tajima's π") Average number of pairwise differences.

θ_W ("Watterson's θ") $= \frac{\text{number of mutations}}{\sum_{i=1}^{k-1} 1/i}$

Both are unbiased estimators of θ, i.e. $\mathbb{E}\theta_W = \mathbb{E}\theta_\pi = \theta$.

Example: Ward et al. (1991) sampled 360 bp sequences from mtDNA control region of $n = 63$ Nuu Chah Nulth and observed 26 mutations.

$$\theta_W = \frac{26}{\sum_{i=1}^{63} 1/i} = 5.5123$$

This corresponds to 0.0153 Mutations per base and per $2 \cdot N_e$ generations.Assuming a mutation rate $\widehat{\mu} \approx 6.6 \cdot 10^{-6}$ per generation per site this leads to an effective population size of

$$\widehat{N_e} = \frac{\theta_W/360}{2 \cdot \widehat{\mu}} \approx 1150 \text{ females}$$

How precise is this estimation?

$$\mathsf{var}(\theta_W) = \frac{\theta}{\sum_{i=1}^{n} 1/i} + \theta^2 \cdot \frac{\sum_{i=1}^{n} 1/i^2}{\left(\sum_{i=1}^{n} 1/i\right)^2}$$

Theorem 1 *Any unbiased estimator of θ has variance at least*

$$\frac{\theta}{\sum_{k=1}^{n-1} \frac{1}{k+\theta}}.$$

(Here, we assume that the estimation is based on a single locus without recombination).

For the Nuu Chah Nulth data we get:

$$\theta_W = 5.5123$$

$$\sigma_{\theta_W} = 3.42$$

Confidence range? (2σ-rule would leed to negative values...)

Conclusion: N_e could perhaps also be 200 or 3000 females.

How can we improve this estimate? Sample more individuals? How many individuals n would we need to get $\sigma_{\theta_W} = 0.1 \cdot \theta$? From the formula for $\mathsf{var}\theta_W$ follows that we need $n \approx 2 \cdot e^{100/\theta}$. For $\theta = 5$, this is $n \approx 10^9$. For $\theta = 1$, this is $n \approx 10^{43}$. number of water molecules on earth$\approx 10^{47}$ number of seconds since big bang$\approx 4.3 \cdot 10^{17}$

Solution: sample many loci!

References

[Fel06] J. Felsenstein (2006) Accuracy of Coalescent Likelihood Estimates: Do We Need More Sites, More Sequences, Or More Loci? *Mol. Biol. Evol.*, **23.3**: 691–700.

How to sample if

- one read is 600 bp long
- costs for developing a new locus is 40$
- costs for collecting a sample is 10 or 0.10$
- costs for a single read is 6$
- you can spend 1000$
- true θ is 1.8 (per locus)

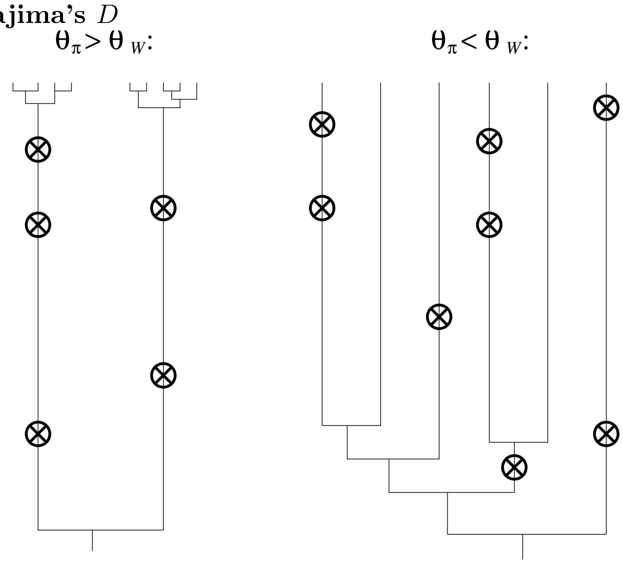
Optimal sampling scheme: $n = 7$ or $n = 8$, respectively, individuals and 11 loci.

With this sampling scheme we get:

$$\sigma_{\theta_W} \approx 0.2 \cdot \theta \text{ and } \sigma_{\theta_\pi} \approx 0.22 \cdot \theta$$

(all this is based on infinte-sites assumptions)

Tajima's D



$D := \frac{\theta_\pi - \theta_W}{\widehat{\sigma}_{\theta_\pi - \theta_W}}$
substructure?
population
growth?
selection?

4 Outline of methods

4.1 ML with Importance Sampling

The Likelihood

$\psi = (\psi_i)_i$ vector of model parameters

D sequence data

$$L_D(\psi) = \Pr_\psi(D) = \int_{\text{all Genealogies } G} \Pr_\psi(D \mid G) \cdot P_\psi(dG).$$

Importance Sampling

Draw $G_1, \ldots, G_k$ (approx.) i.i.d. with density Q and approximate

$$\int \Pr_\psi(D \mid G)\ P_\psi(dG) \approx \frac{1}{k} \sum_{i=1}^{k} \frac{\Pr_\psi(D \mid G_i) \cdot P_\psi(G_i)}{Q(G_i)}.$$

efficient for ψ with

$$\Pr_\psi(D \mid G_i) \cdot P_\psi(G_i) \approx Q(G_i)$$

Methods differ in their choice of Q.

Griffiths & Tavaré (1994)

Q: Generate G backwards in time, greedy proportional to coalescence and mutation probabilities. Choose between all allowed events.

Good for infinite sites models, inefficient if back-mutations are allowed.

4.2 MCMC for frequentists and Bayesians

Felsenstein, Kuhner, Yamato, Beerli,. . .

For some initial ψ_0, sample Genealogies G approx. i.i.d. according to $\Pr_{\psi_0}(G \mid D)$ by Metropolis-Hastings MCMC.

Coalescent is a natural prior for G!

Two flavours:

for frequentists: use $G_1, \ldots, G_k$ for Importance Sampling

Optimize approx. Likelihood → ψ_1

Iterate with ψ_0 replaced by ψ_1

for Baysians: Then sample ψ conditioned on Genealogies and iterate to do Gibbs-sampling from $\Pr(\psi, G \mid D)$.

Problems of full-data methods

- usual runtime for one dataset: several weeks or months
- complex software, development takes years
- most programs not flexible, hard to write extensions

4.3 Approximate Bayesian Computation (ABC)

Pritchard et al. (1999)

Approximate Bayesian Computation

1. Select summary statistics $S = (S_i)_i$ and compute their values $s = (s_i)_i$ for given data set
2. Choose tolerance δ
3. repeat until k accepted ψ':
   - Simulate ψ' from prior distribution of ψ
   - Simulate genealogy G according to $\Pr_{\psi'}(G)$.
   - Simulate data and compute values s' of S
   - accept ψ' if $\|s - s'\| \leq \delta$

Only possible if a few summary statistics suffice. We will later discuss refinements and extensions of this approach.

Beaumont, Zhang, Balding (2002)

> *"[...] the MCMC-based method is consistently superior to the summary-statistics-based methods and highlights that it is well worth making the effort to obtain full-data inferences if possible."*

> *"[...] there are advantages to the use of summary statistics, both in the ease of implementation and in the time to obtain the results [...]"*

> *"Further research is needed to find a more rigorous way for choosing summary statistics, including the use of orthogonalization and 'projection-pursuit' methods"*

5 Importance sampling for genealogies

D: data set of DNA sequences sampled from a population. In case of a structured population sampling locations are known.

Aim: Estimate parameters $\Theta := (\theta_i, M_{ij})_{ij}$.

Maximum-Likelihood (ML) approach: Find the set of parameter values that maximizes the likelihood:

$$\widehat{\Theta} := \arg\max_{\Theta} \mathsf{Pr}_\Theta(D)$$

How to compute the likelihood?

$$L_D(\Theta) = \mathsf{Pr}_\Theta(D) = \sum_G \mathsf{Pr}_\Theta(G) \cdot \mathsf{Pr}_\Theta(D \mid G).$$

More precisely:

$$L_D(\Theta) = \mathsf{Pr}_\Theta(D) = \int_{\text{all genealogies } G} \mathsf{Pr}_\Theta(D \mid G)\ P_\Theta(G) dG$$

where $P_\Theta(G)$ is the density of the (structured) coalscent distribution at the genealogy G.

What does this mean?

And what is dG?

Let's first ask: What is the dx in

$$\int_0^1 x^2 dx \qquad ?$$

dx is used in an ambigous way. This is sloppy but intuitive.

It means "a small environment around x", but also the size of this environment.

To explain this we be a little bit less sloppy for a few minutes and write $\mathrm{d}x$ for the environment and dx for its size.

For some small $n \in \mathbb{N}$ and $x \in \mathbb{R}$ we can define $\mathrm{d}x = [x - \frac{1}{2n}, x + \frac{1}{2n}]$.Then, $dx = 1/n$.

We can approximate $\int_0^1 x^2 dx$ by

$$\sum_{x \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}} x^2 \cdot \frac{1}{n} = \sum_{x \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}} x^2 \cdot dx \overset{n\to\infty}{\to} \int_0^1 x^2 dx$$

dx is always meant to be “infinitesimally small”, i.e. $dx \to 0$

What is a probability density?

$P(x)$ is the probability density of a random variable X in x if

$$\Pr(X \in \mathsf{d}x) \approx P(x) \cdot dx$$

and the “$\approx$” becomes a “=” for “infinitesimally small” dx. This is again sloppy and intuitive.It actually means that

$$\lim_{dx \to 0} \frac{\Pr(X \in \mathsf{d}x)}{dx} = P(x)$$

It then follows that

$$\Pr(X \in [a, b]) = \int_a^b P(x) dx.$$

Examples

The density of the exponential distribution with rate λ at x is

$$\lambda e^{-\lambda x}.$$

The density of the normal distribution with mean value μ and standard deviation σ is

$$\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Now for dG

Let $\mathrm{d}G$ be a small environment around the genealogy G. This means, $\mathrm{d}G$ consists of all genealogies G' that have the same topology as G and if $\tau_1, \ldots, \tau_n$ are the points in time where coalescent events or migrations of lineages or thelike occurr in G, and $\tau_1', \ldots, \tau_n'$ are the corresponding points in time for G', then

$$\forall_{k \le n} |\tau_k - \tau_k'| \le \varepsilon.$$

Thus, the volume dG of $\mathrm{d}G$ can be defined to be $(2\varepsilon)^n$. The density $P_\Theta(G)$ is then defined by

$$\mathsf{Pr}_\Theta(G' \in \mathsf{d}G) \approx P_\Theta(G) \cdot dG$$

where $\Pr_\Theta(G' \in \mathsf{d}G)$ is the probability that a genealogy G' that was generated according to the probability distribution of a structured coalecent with parameter values Θ results to be in the environment $\mathrm{d}G$ of G, or, more precisely:

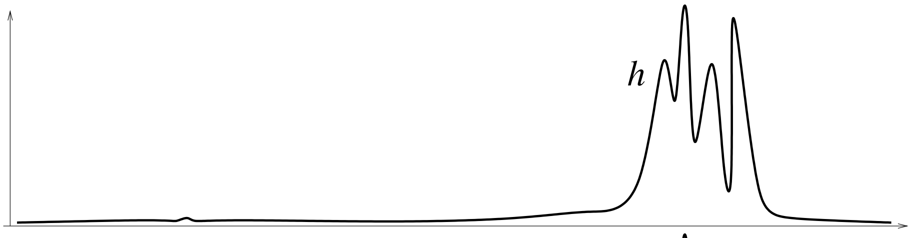
$$\frac{\mathsf{Pr}_\Theta(G' \in \mathsf{d}G)}{dG} \overset{dG\to 0}{\longrightarrow} P_\Theta(G)$$

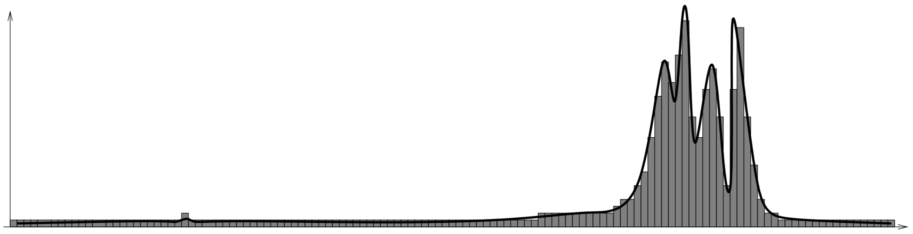
The equation

$$L_D(\Theta) = \mathsf{Pr}_\Theta(D) = \int_{\text{all genealogies}} \mathsf{Pr}_\Theta(D \mid G)\ P_\Theta(G) dG$$

should now make some more sense to us. But how can we compute it? We use Importance Sampling.

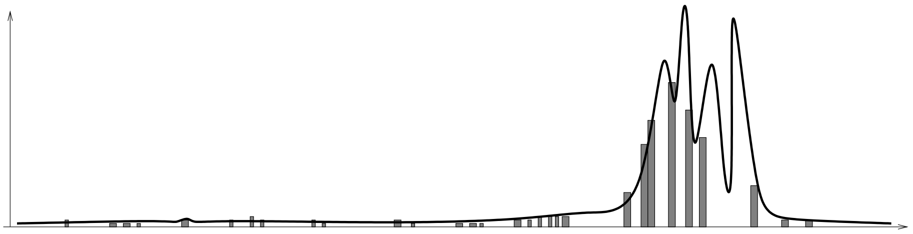
How can we compute the integral $\int_a^b h(x)dx$ of this function h?





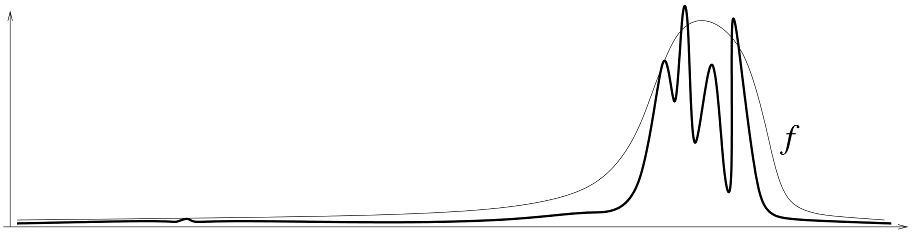
Approximation by a step function: If $x_1, \ldots, x_k$ are the means of the partition intervals and $c = \frac{b-a}{k}$ is their width, then

$$\int_a^b h(x)\ dx \approx \sum_{i=1}^{k} c \cdot h(x_i) = \frac{b-a}{k} \sum_{i=1}^{k} h(x_i).$$

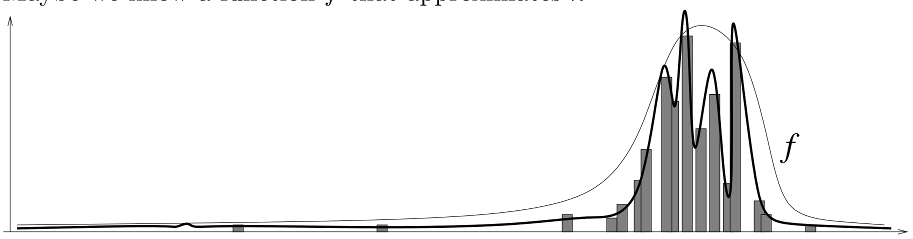


Maybe save some time by just taking a sample of k values $h(x)$.

$$\int_a^b h(x)\ dx \approx \frac{b-a}{k} \sum_{i=1}^{k} h(X_i) = \frac{1}{k} \sum_{i=1}^{k} \frac{h(X_i)}{\frac{1}{b-a}}.$$



Maybe we know a function f that approximates h



We can sample more from the relevant range but we have to correct this by the Importance-Sampling formula:

$$\int h(x)\ dx \approx \frac{1}{k} \sum_{i=1}^{k} \frac{h(X_i)}{q(X_i)}$$

where $X_1, \ldots, X_k$ are independent samples from a distribution whose density q is proportional to f. The closer f is to h, the better the approximation.

Sketch of proof of the IS formula

$$\begin{aligned}\int_a^b h(x)dx &= \int_a^b \frac{h(x)}{q(x)}\cdot q(x)dx \\ &= \mathbb{E}_q \frac{h(X)}{q(X)} \\ &= \frac{1}{k}\cdot\sum_{i=1}^{k}\frac{h(X_i)}{q(X_i)},\end{aligned}$$

where $\mathbb{E}_q$ is the expectation value under the assumption that X has probability density q, and $X_1,\dots,X_k$ are independently sampled with probability density q.

Importance Sampling for computing the likelihood of for a range of parameter values Θ: Generate genealogies $G_1,\dots,G_k$ (more or less) independently according to a probability density $Q(G_i)$. Then,

$$\begin{aligned}L_D(\Theta) &= \int_{\text{all genealogies } G} \underset{\Theta}{\Pr}(D|G)\cdot P_\Theta(G)dG \\ &\approx \frac{1}{k}\sum_{i=1}^{k}\frac{\Pr_\Theta(D|G_i)\cdot P_\Theta(G_i)}{Q(G_i)}.\end{aligned}$$

Method differ in their choice of Q and will be most efficient if

$$Q(G)\approx \underset{\Theta}{\Pr}(D|G)\cdot P_\Theta(G).$$

6 Griffiths und Tavaré

References

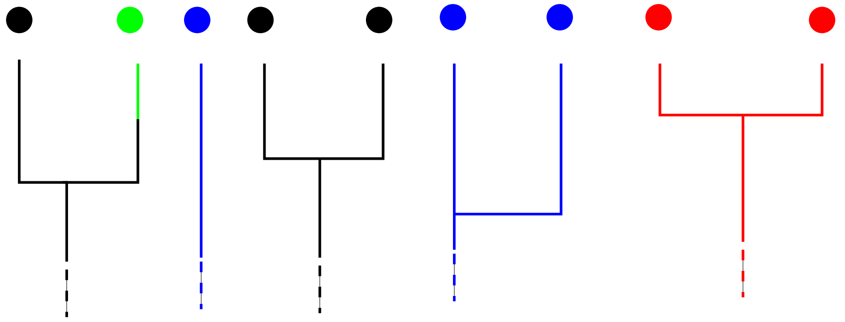
[1] Griffiths und Tavaré (1994) Ancestral Inference in Population Genetics *Statistical Science* 9(3): 307-319. http://www.stats.ox.ac.uk/~griff/software.html

Start with an initial guess Θ_0. Define the history of a sample to be $H=(H_1,H_2,\dots,H_\ell)$, where the historical events H_k can be

1. lineages i and j coalesce
2. mutation on lineage i
3. lineage i from island a traces back to island b

and $H_1,H_2,\dots,H_\ell$ goes from present to past.

For the Importance Sampling procedure, many histories $H^{(1)},H^{(2)},\dots,H^{(M)}$ are generated. For each history $H^{(i)}$ are sampled $H_1^{(i)},H_2^{(i)},\dots$ step by step from the tips to the root of the tree. Given the data, not all events are possible. E.g., lineages cannot coalesce if they are of different allelic type. If the infinite-site mutation model is used (to make the Griffith-Tavaré scheme efficient), not all mutations are allowed.



Let $b_{ij}(\theta_0)$ be the probability of the jth event $h=H_j^{(i)}$ in the ith sampled history $H^{(i)}$ and let $(a_{ijk}(\theta_0))_k$ be the series of rates of all events that would have been allowed for this step. Then, the probability to choose h was $b_{ij}(\theta_0)/\sum_k a_{ijk}(\theta_0)$. Thus, $\prod_j b_{ij}(\theta_0)/\sum_k a_{ijk}(\theta_0)$ is the die importance-sampling probability $Q_{\theta_0}(H^{(i)})$ of the entire history $H^{(i)}$. According to the importance-sampling formula we get for all θ that are not too far from θ_0:

$$L_{(D)}(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} \prod_j \frac{b_{ij}(\theta) \cdot \sum_k a_{ijk}(\theta_0)}{\sum_k a_{ijk}(\theta) \cdot b_{ij}(\theta_0)}$$

- Advantage over MCMC: Histories are sampled really independent of each other.
- Disadvantage: For finite-sites models many different mutation events are allowed in each step, which makes the method very inefficient. Stephens and Donnelly (2000) found a solution for this, which we will discuss later in the semester.

7 Lamarc (and Migrate)

Rate parameters and time scales

For autosomal DNA:

	per generation	per $2N_i$ generations	per $1/\mu$ generations
mutation rate	μ	$\frac{\theta_i}{2} = 2N_i\mu$	1
migration rate of ancestral lineage from i tracing back to j	m_{ij}	$\gamma_{ij} = 2N_i m_{ij}$	$M_{ij} = \frac{m_{ij}}{\mu} = \frac{2\gamma_{ij}}{\theta_i}$
coalescence on island i	$1/(2N_i)$	1	$\frac{1}{2N_i\mu} = \frac{2}{\theta_i}$

Number of alleles on island i that choose their parent allele on island j:

$$2N_i \cdot m_{ij} = \gamma_{ij}$$

Combining IS with MCMC

References

[1] M. Kuhner, J. Yamato, J. Felsenstein (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hasings sampling. *Genetics* **140**: 1421–1430

[2] P. Beerli, J. Felsenstein (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *PNAS* **98.8**: 4563–4568

- MIGRATE-N http://popgen.sc.fsu.edu/Migrate/Migrate-n.html
- LAMARC http://evolution.genetics.washington.edu/lamarc/lamarc.html

LAMARC strategy

Begin with initial parameter guess $\Theta_0 = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, M_{12}^{(0)}, M_{12}^{(0)}, M_{23}^{(0)}, \ldots)$, repeat the following steps for $i = 0, 1, 2, \ldots, m-1$

1. Metropolis-Hastings MCMC sampling of genealogies $G_1, G_2, \ldots, G_k$ (approx.) according to the posterior density $p_{\Theta_i}(G|D)$ given the data D. What is Hetropolis-Hastings MCMC?

2. importance sampling:

$$\frac{L_D(\Theta)}{L_D(\Theta_i)} \approx \frac{1}{k}\sum_{j=1}^{k}\frac{p_\Theta(G_j)}{p_{\Theta_i}(G_j)} =: F_{\Theta_i}(\Theta)$$

Why is this justified as importance sampling?

3. $\Theta_{i+1} := \arg\max_\Theta F_{\Theta_i}(\Theta)$

and hope that $\Theta_m \approx \widehat{\Theta} = \arg\max_\Theta L_D(\Theta)$

Justification of step 2

$$\begin{aligned}
\frac{L_D(\Theta)}{L_D(\Theta_i)} &\approx \frac{\frac{1}{k}\sum_{j=1}^{k}\frac{\Pr_\Theta(D|G_j)\cdot p_\Theta(G_j)}{p_{\Theta_i}(G_j|D)}}{\Pr_{\Theta_j}(D)} && \text{(importance sampling)}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{\Pr_\Theta(D|G_j)\cdot p_\Theta(G_j)}{p_{\Theta_i}(G_j|D)\cdot\Pr_{\Theta_i}(D)}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{\Pr_\Theta(D|G_j)\cdot p_\Theta(G_j)}{p_{\Theta_i}(G_j,D)}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{\Pr_\Theta(D|G_j)\cdot p_\Theta(G_j)}{\Pr_{\Theta_i}(D|G_j)\cdot p_{\Theta_i}(G_j)} \quad = \quad \frac{1}{k}\sum_{j=1}^{k}\frac{p_\Theta(G_j)}{p_{\Theta_i}(G_j)}
\end{aligned}$$

The last equation follows from $\Pr_\Theta(D|G_j) = \Pr_{\Theta_i}(D|G_j)$, which holds since the mutation rate is always 1 and thus the D is independent of Θ when G is given.

Markov-Chain Monte Carlo (MCMC)

MCMC: construct Markov chain $X_0, X_1, X_2, ...$ with stationary distribution $\Pr(G \mid D)$ and let it converge.

Markov property:

$$\forall_{i,x}: \quad \Pr(X_{i+1} = x|X_i) = \Pr(X_{i+1} = x|X_i, X_{i-1}, \ldots, X_0)$$

In words: The probabilty for the next state may depend on the current state but not additionally on the past.

"Equilibrium" or "Stationary distribution" p:

$$\forall_{i,x}: \quad p(x) = \sum_y p(y)\cdot\Pr(X_{i+1} = x|X_i = y)$$

In words: If you choose an element of the state space according to p and go one step, the probability to be in x is $p(x)$ not only in the first step but also in the second step and consequently in any further step.When you are once in equilibrium, you'll be forever.

Theorem 2 *If $X_0, X_1, X_2\ldots$ is a aperiodic, irreducible Markov chain on a finite state space S with equilibrium p, it will converge against the equilibrium p in the following sense:*

$$\forall_{x,y}: \quad \Pr(X_n = x|X_0 = y) \overset{n\to\infty}{\longrightarrow} p(x)$$

Irreducible means:

$$\forall_{x,y}\exists_i\forall_m: \quad \Pr(X_{i+m} = x|X_m = y) > 0$$

Aperiodic means:

$$\forall_{x,y,m}: \quad \mathsf{gcd}\left(\{k\in\mathbb{N}|\Pr(X_{k+m} = x|X_m = y) > 0\}\right) = 1,$$

where gcd means "greatest common divisor".

(let's watch a Tcl/Tk simulation of a Markov chain)
"Equilibrium" or "Stationary distribution" p:

$$\forall_{i,x}: \quad p(x) = \sum_y p(y) \cdot \Pr(X_{i+1} = x | X_i = y)$$

Stronger condition than equilibrium: reversibility (or "detailed balance")

$$p(x) \cdot \Pr(X_{i+1} = y | X_i = x) = p(y) \cdot \Pr(X_{i+1} = x | X_i = y)$$

In words: If you start in equilibrium, and it is reversible, a move from x to y is as probable as a move from y to x.

Alternative explanation: If you watch a movie of the process starting in a reversible equilibrium, the probability of what you see does not change if you watch the movie backwards.

Given the probability distribution $\Pr(.|D)$, how can we construct a Markov chain that converges against it?

One possibility: **Metropolis-Hastings**
Given current state $X_i = x$ propose y with Prob. $Q(x \to y)$
Accept proposal $X_{i+1} := y$ with probability

$$\min \left\{ 1, \frac{Q(y \to x) \cdot \Pr(y \mid D)}{Q(x \to y) \cdot \Pr(x \mid D)} \right\}$$

otherwise $X_{i+1} := X_i$

(All this also works with continuous state space, with some probabilities replaced by densities.)

Why Metropolis-Hastings works

Let's assume that $\frac{Q(y \to x) \cdot \Pr(y \mid D)}{Q(x \to y) \cdot \Pr(x \mid D)} \leq 1$. (Otherwise swap x and y in the following argument).Then, if we start in x, the probability $\Pr(x \to y)$ to move to y (i.e. first propose and then accept this) is

$$Q(x \to y) \cdot \frac{Q(y \to x) \cdot \Pr(y \mid D)}{Q(x \to y) \cdot \Pr(x \mid D)} = Q(y \to x) \frac{\Pr(y \mid D)}{\Pr(x \mid D)}$$

If we start in y, the probability $\Pr(y \to x)$ to move to x is

$$Q(y \to x) \cdot 1,$$

since our assumption implies $\frac{Q(x \to y) \cdot \Pr(x \mid D)}{Q(y \to x) \cdot \Pr(y \mid D)} \geq 1$.

This implies that the reversibility condition

$$\Pr(x \mid D) \cdot \Pr(x \to y) = \Pr(y \mid D) \cdot \Pr(y \to x)$$

is fulfilled.This implies that $\Pr(. \mid D)$ is an equilibrium of the Markov chain that we have just constructed, and the latter will converge against it.(let's watch a simulation in R)

Applying Metropolis-Hastings

- You are never in equilibrium (your target distribution), but you can get close if you run enough steps.
- You can take more than one sample from the same chain, but you should run enough steps between the sampling steps to make the sampled objects only weakly dependent.
- Your initial state may be "far from equilibrium" (i.e. very improbable). So you should run the chain long enough before you start sampling ("burn-in").

Lamarc's Metropolis-Hastings step

Target distribution density: $p_\Theta(G|D)$, where Θ is the current set of parameter values, G is the genealogy and D is the data.

Proposal chain: Remove a randomly picked branch and let the ancestral lineage of the isolated subtree coalesce with the rest accoring to Θ.
$\Rightarrow$

$$\frac{Q(G' \to G)}{Q(G \to G')} = \frac{p_\Theta(G)}{p_\Theta(G')}$$

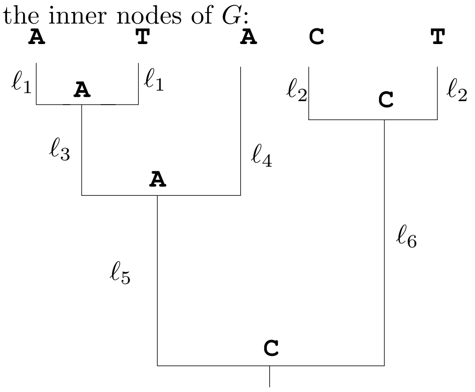
$\Rightarrow$ The MH acceptance probability is:

$$\begin{aligned}\min\left\{1, \frac{Q(G' \to G) \cdot p_\Theta(G'|D)}{Q(G \to G') \cdot p_\Theta(G|D)}\right\} &= \min\left\{1, \frac{p_\Theta(G) \cdot p_\Theta(G', D)/Pr(D)}{p_\Theta(G') \cdot p_\Theta(G, D)/Pr(D)}\right\} \\ &= \min\left\{1, \frac{p_\Theta(G) \cdot \Pr(D|G') \cdot p_\Theta(G')}{p_\Theta(G') \cdot \Pr(D|G) \cdot p_\Theta(G)}\right\} \\ &= \min\left\{1, \frac{\Pr(D|G')}{\Pr(D|G)}\right\}\end{aligned}$$

How to compute $\Pr(D|G)$? Felsenstein's pruning!
We assume that all sites evolve independent of each other. $\Rightarrow$

$$\Pr(D|G) = \prod_i \Pr(D_i|G),$$

where D_i is the i-th column in the alignment.
How to compute $\Pr(D_i|G)$? For any nucleotides (or amino acids) x, y let p_x be the frequency of x and $\Pr_{x\to y}(\ell)$ be the probability that a child node has type y, given that the parent node had type x and the branch between the two nodes has length ℓ. Let's first assume that D_i knows the nucleotides at the inner nodes of G:



$$\begin{aligned}&\Pr(D_i|G) \\ &= p_C \cdot \Pr_{C\to A}(\ell_5) \cdot \Pr_{C\to C}(\ell_6) \cdot \\ &\quad \Pr_{A\to A}(\ell_3) \cdot \Pr_{A\to A}(\ell_4) \cdot \\ &\quad \Pr_{A\to A}(\ell_1) \cdot \Pr_{A\to T}(\ell_1) \cdot \\ &\quad \Pr_{C\to C}(\ell_2) \cdot \Pr_{C\to T}(\ell_2) \cdot\end{aligned}$$

How to compute or define $\Pr_{x\to y}(\ell)$?

Jukes-Cantor model for DNA evolution

- All nucleotide frequencies are $p_A = p_C = p_G = p_T = 0.25$.
- "mutation events" happen at rate λ and let the site forget its current type and select a new one randomly from {A,C,G,T}. (New one can be the same as old one.)

$\Rightarrow$

$$\Pr_{x\to y}(\ell) = \begin{cases} = & \left(1 - e^{-\lambda\ell}\right) \cdot \frac{1}{4} & \text{if } x \neq y \\ = & e^{-\lambda\ell} + \left(1 - e^{-\lambda\ell}\right) \cdot \frac{1}{4} & \text{if } x = y \end{cases}$$

(More sophisticated sequence evolution models in the phylogenetics part of the lecture.)

Felsenstein's pruning algorithm

How to compute $\Pr(D_i|G)$ if (as usual) the data do only contain the nucleotides for the tips of the tree?

For any node k of the genealogy and any nucleotide (or amino acid) x define $w_k(x)$ to be the probability that, given the nucleotide (or a.a.) in k is x, the tipps that stem from k get the nucleotides (or a.a.) given in D_i. Then

$$\Pr(D_i|G) = \sum_{x\in\{A,C,G,T\}} p_x \cdot w_r(x),$$

where r is the root of the genealogy,and for any node k with child nodes i and j and corresponding branch lengths ℓ_i and ℓ_j we get:

$$w_k(x) = \left(\sum_{y\in\{A,C,G,T\}} \Pr_{x\to y}(\ell_i)\cdot w_i(y)\right)\cdot\left(\sum_{z\in\{A,C,G,T\}} \Pr_{x\to y}(\ell_j)\cdot w_j(z)\right)$$

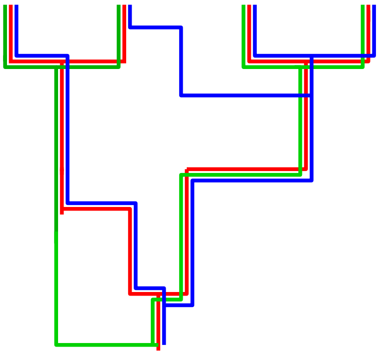
Felsenstein's pruning algorithm

If b is a tip of G, then $w_b(x)$ is 1 if x is the nucleotide at b in D_i, and $w_b(x)$ is 0 otherwise.

With the recursion for$w_k(x)$ given above, we can compute $w_k(x)$ for all x and all k starting with the tips and ending in the root r.

From the $w_r(.)$ we can compute $\Pr(D_i|G)$.

Ancestral Recombination Graph



When recombination occurs, ancestral lineages for the left and the right part of the sequence split up. Each site has a tree-shaped ancestry, and these trees have complex stochastic dependencies.
LAMARC can also sample Ancestral Recombination Graphs instead of trees.

References

[1] I. J. Wilson, D. J. Balding (1998) Genealogical inference from microsatellite data. *Genetics* **150**: 499-510

- assign data to inner nodes
- when choosing new parent node take mutation probs into account
- more intelligent proposals but larger state space
- may be superior for microsatellite data

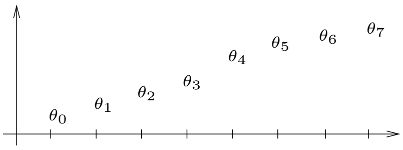
LAMARC Search Strategies

initial chains: several short chains to optimize driving values

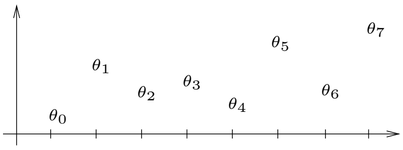
final chain: longer chain to narrow the final interval

burn-in: discard e.g. first 5% of each chain

symptom of too few chains: parameters are still changing directionally



symptom of too short chains: parameters leap wildly from chain to chain



(MC)^3=MCMCMC

=Metropolis-Coupled MCMC= MCMC with "heated chains".

If $\beta_i \in (0,1]$ is heat parameter for chain i, then chain i samples from distribution $p^{\beta_i} : x \mapsto p^{\beta_i}(x)\cdot$const, with $\beta_1 = 1$.

The usual MH acceptance prob. for chain i is

$$\min\left\{1, \frac{p(y)^{\beta_i}}{p(x)^{\beta_i}} \cdot \frac{Q_{y\to x}}{Q_{x\to y}}\right\}.$$

Sometimes a swap between the current state x_i of chain i and the current state x_j of chain j is proposed. The acceptance with probability

$$\min\left\{1, \frac{p(x_i)^{\beta_i}}{p(x_j)^{\beta_i}} \cdot \frac{p(x_j)^{\beta_j}}{p(x_i)^{\beta_j}}\right\}$$

fulfills the requirements of both chaines (check this!).

Bayesian Lamarc

Aim: sample parameter values Θ (and Genealogies) according to the posterior probability distribution $\Pr(\Theta|D)$ (or $\Pr(\Theta, G|D)$) given the data D.

- needs priors $\Pr(\Theta)$ for the parameters
- Gibbs sampling scheme: iterate uptdate of the Θ, given D and G, and update of G, given Θ and D.

Gibbs samping

Assume we want to sample from a joint distribution $\Pr(A = a, B = b)$ of two random variables, and for each pair of possible values (a, b) for (A, B) we have Markov chains with transition probabilities $P^{(A=a)}_{b\to b'}$ and $P^{(B=b)}_{a\to a'}$ that converge against $\Pr(B = b|A = a)$ and $\Pr(A = a|B = b)$.

Then, any Markov chain with transition law

$$P_{(a,b)\to(a',b')} = \begin{cases} \frac{1}{2}P^{(B=b)}_{a\to a} + \frac{1}{2}P^{(A=a)}_{b\to b} & \text{if} \quad a = a' \quad \text{and} \quad b = b' \\ \frac{1}{2}P^{(B=b)}_{a\to a'} & \text{if} \quad a \neq a' \quad \text{and} \quad b = b' \\ \frac{1}{2}P^{(A=a)}_{b\to b'} & \text{if} \quad a = a' \quad \text{and} \quad b \neq b' \\ 0 & \text{else} \end{cases}$$

Priors in Bayesian Lamarc

When new values for Θ are to be proposed,

- e.g. the new values of θ and the recombination rate are chosen according to a exponential prior that is uniform on the log scaled interval $[10^{-5}, 10]$and the
- growth rate g is chosen uniformly from $[-500, 1000]$.
- For the MH acceptance step use a U that is uniform on $[0, 1]$ and accept if

$$U < \frac{\Pr(G|\Theta_{\text{proposal}})}{\Pr(G|\Theta_{\text{old}})}$$

8 IM, IMa, IMa2

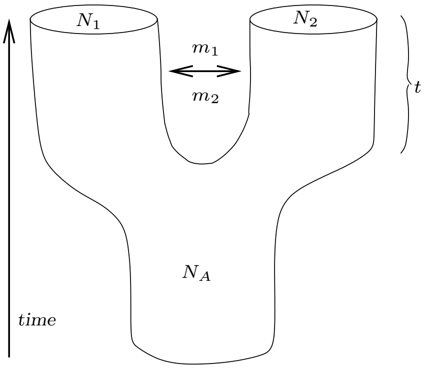
References

[1] Nielsen, R. and J. Wakeley 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**:885-896

[2] Hey, J., and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics* **167**:747-760

[3] Hey, J., and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *PNAS* **104**:27852790.

[4] Hey J. 2010. Isolation with Migration Models for More Than Two Populations. *Mol Biol Evol* **27**: 905-20



m_i**:** proportion of individuals of pop. i that are replaced by immigrants

N_i**:** effective size of pop. i

N_A**:** effective size of ancestral population

t**:** number of generations since the split

μ**:** mutation rate per generation

Asymptotics and rescaled parameters:

$$\begin{array}{rcl@{\qquad\qquad}rcl} N_i & \to & \infty & 2N_i m_i & \to & M_i \\ N_2/N_1 & \to & r & 4N_1\mu & \to & \theta \\ N_A/N_1 & \to & a & t/(2N_1) & \to & \tau \end{array}$$

$$\Theta = (\theta, r, a, \tau, M_1, M_2)$$

IM is an implemetion of a Bayesian sampler with flat priors, e.g.

$$\begin{array}{rcl@{\qquad\qquad}rcl} M_i & \sim & \mathsf{Unif}([0, 10]), & T & \sim & \mathsf{Unif}([0, 10]) \\ \log(r) & \sim & \mathsf{Unif}([-10, 10]), & \log(a) & \sim & \mathsf{Unif}([-10, 10]) \end{array}$$

Proposals G^* for genealogy updates like in Lamarc with MH acceptance probability

$$\min\left\{1, \frac{\Pr(D|\Theta_i, G^*)}{\Pr(D|\Theta_i, G_i)}\right\},$$

where G_i is the current genealogy and Θ_i is the current vector of parameter values in MCMC step i.

Proposals for parameter updates: Given the current value λ of some parameter, the new value is proposed from Unif$[\lambda - \Delta, \lambda + \Delta]$. MH acceptance probability:

$$\min\left\{1, \frac{p(G_i|\Theta^*)}{p(G_i|\Theta_i)}\right\}$$

IM can handle datasets of unlinked loci (but NO intralocus-recombination!).

$D = (D^1, \ldots, D^n)$, D^i: data from locus i. $G = (G^1, \ldots, G^n)$, G^i: genealogy of locus i (including topology, branch lengths, migration times, coalescent times)

$$p(\Theta|D) = \frac{p(\Theta)}{\Pr(D)} \cdot \prod_{i=1}^{n} \int_{G^i} \Pr(D^i|G^i, \Theta) \cdot p(G^i|\Theta) dG^i$$

additional parameters: locus-specific mutation scalars u_i with constraint $\prod_i u_i = 1$.

Updating $(u_1, \ldots, u_n)$: choose i and j and propose

$$u_i^* = x \cdot u_i \text{ and } u_j^* = u_j/x,$$

where $\log(x) \sim \mathsf{Unif}(-\delta, \delta)$.

In IMa, some MCMC steps are replaced by faster numerical computation. We discuss this first in a 1-population model with sample size m.

- Let τ_k be the time while the number of lineages is k, measured in $1/\mu$ generations.
- $\Rightarrow$ coalescence rate is $2/\theta$
- $\Rightarrow p(G|\Theta) = \left(\frac{2}{\theta}\right)^{m-1} \cdot \exp(-2 \cdot f_m/\theta)$,
- where $f_m := \sum_{i=2}^{m} \tau_i \cdot i \cdot (i-1)$

Assume a flat prior $\theta \sim \mathsf{Unif}(0, \theta_{max})$.This implies

$$p(G) = \int_0^{\theta_{\mathsf{max}}} p(\theta) \cdot p(G|\theta) d\theta = \frac{2}{\theta_{\mathsf{max}} f_m^{m-2}} \cdot \Gamma(m-2, 2f_m/\theta_{\mathsf{max}}),$$

where $\Gamma(a, b) = \int_b^\infty x^{a-1} e^{-x} dx$ is the "incomplete Gamma-function".

This implies

$$p(\theta|G) = \frac{p(G|\theta) \cdot p(\theta)}{p(G)} = \frac{(2f_m/\theta)^{m-2} \exp\left(-2f_m/\theta\right)}{\theta \cdot \Gamma(m-2, 2f_m/\theta_{\mathsf{max}})}$$

Hence, given f_m, the posterior probability can be computed and the expression above gives a smooth curve.

- works in a similar way for models with subpopulations with migration
- for the split time τ a standard MH step is required
- population growth not allowed in IMa (other than IM)
- "branch sliding" proposals for G: move randomly chosen branch a random distance. Current migration events are removed and replaced by a Poisson number of migration events conditioned on odd or even.

Likelihood Ratio Testing with IMa

Let

$$\widehat{\Theta}_0 = \mathsf{arg}\max p(\Theta|D) \text{ in the general model}$$

and

$$\widehat{\Theta}_r = \mathsf{arg}\max p(\Theta|D) \text{ in a restricted model, e.g. without migration.}$$

Since we use uniform priors for all parameters (some log-scaled), we get

$$\frac{p(\Theta_0|D)}{p(\Theta_r|D)} = \frac{\Pr(D|\Theta_0)\cdot p(\Theta_0)}{\Pr(D|\Theta_r)\cdot p(\Theta_r)} = \frac{L_D(\Theta_0)}{L_D(\Theta_r)}$$

Hence, $\widehat{\Lambda} = \log\left(\frac{\widehat{p}(\Theta_0|D)}{\widehat{p}(\Theta_r|D)}\right)$ is an approximation of the log likelihood-ratio and thus, $2\widehat{\Lambda}$ is approximately χ^2_d-distributed under the null hypothesis of the restricted model, where d is the number of additional parameters in the general model.However, this approximation is only appropriate for extremely large datasets. IMa assesses the significance of $\widehat{\Lambda}$ by comparing it to values of $\widehat{\Lambda}$ from simulations based on the null hypothesis (restricted model).

Bayes factors

Other authors use so-called Bayes factors to decide between two models M_1 and M_2:

$$B_{M_1,M_2} = \frac{\Pr(D|M_1)}{\Pr(D|M_2)},$$

where

$$\begin{aligned}\Pr(D|M) &= \int p(D,\Theta|M)d\Theta \\ &= \int \Pr(D|M,\Theta)\cdot p(\Theta|M)d\Theta \\ &\approx \left(\frac{1}{m}\sum_{j=1}^{m}\frac{1}{Pr(D|\Theta_j,M)}\right)^{-1},\end{aligned}$$

where $\Theta_1,\ldots,\Theta_m$ are the samples from an MCMC run.

Why harmonic mean estimator for $\Pr(D)$?

Let $\theta_1,\ldots,\theta_m$ be (approx.) independent samples according to $p(\theta|D)$. Then,

$$\begin{aligned}1 = \int p(\theta)d\theta &\approx \frac{1}{m}\sum_{i=1}^{m}\frac{p(\theta_i)}{p(\theta_i|D)} && \text{(importance sampling)} \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{p(\theta_i)}{\frac{\Pr(D|\theta_i)\cdot p(\theta_i)}{\Pr(D)}} && \text{(Bayes formula)} \\ &= \Pr(D)\cdot\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\Pr(D|\theta_i)}.\end{aligned}$$

$\Rightarrow$

$$\Pr(D) \approx \frac{1}{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\Pr(D|\theta_i)}}$$

Advantages of Bayes factors:

- can also support the restriced model while tests can only support the general model by statistically rejecting the restricted one.
- can also compare non-nested models

Problems:

- Prior has influence even for large amount of data
- harmonic mean estimator can have infinite variance (more sophisticated methods exist)
- Tests and Bayesian model selection can lead to opposite results (Lindley's paradox).

9 Approximate Bayesian Computation (ABC)

Problems of full-data methods:

- usual runtime for one dataset: several weeks or months
- complex software, development takes years
- most programs not flexible, hard to write extensions

References

[PSPL+99] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun and M. W. Feldman (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16(12)**:1791–1798

[BZB02] M.A. Beaumont, W. Zhang, D.J. Balding (2002) Approximate Bayesian Computation in Population Genetics. *Genetics* **162**:2025–2035

[MMPT03] P. Marjoram, J. Molitor, V. Plagnol, S. Tavaré (2003) Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100**:15324–15328

[WCE09] D. Wegmann, C. Leuenberger, L. Excoffier (2009) Efficient approximate Bayesian computation coupled Markov chain Monte Carlo without likelihood. *Genetics* **182**:1207

Pritchard et al. (1999)

- Compute MRCA of human Y chromosome in population models with growth.
- Find strong signal of population expansion in all populations.
- Explanations: recent expansion from a small ancestral population in the last 120,000 years or natural selection on the Y chromosome.
- data: 8 microsatellite loci from 445 humans
- Try various microsatellite mutation models
- Use summary statistics:
  1. mean accross loci in the variance of repeat numbers
  2. mean effective heterozygosity
  3. number of distinct haplotypes

Pritchard et al. (1999)

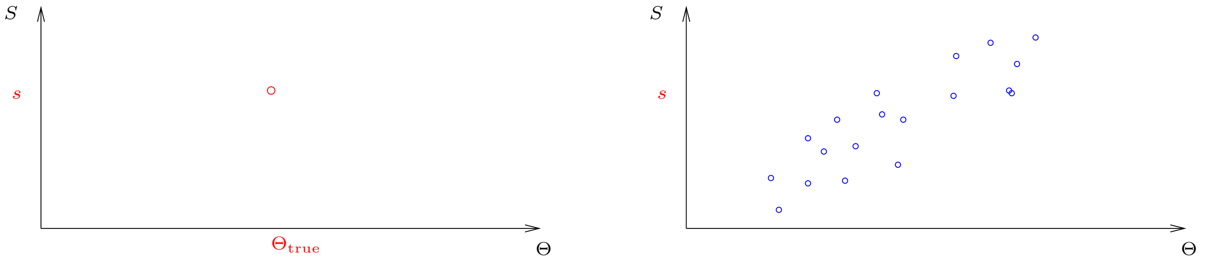
Approximate Bayesian Computation

1. Select summary statistics $S = (S_i)_i$ and compute their values $s = (s_i)_i$ for given data set
2. Choose tolerance δ
3. repeat until k accepted parameter combinations Θ':
   (a) Simulate Θ' from prior distribution of Θ
   (b) Simulate genealogy G according to $\Pr_{\Theta'}(G)$.
   (c) Simulate data and compute values s' of S
   (d) accept Θ' if $\|s - s'\| \leq \delta$

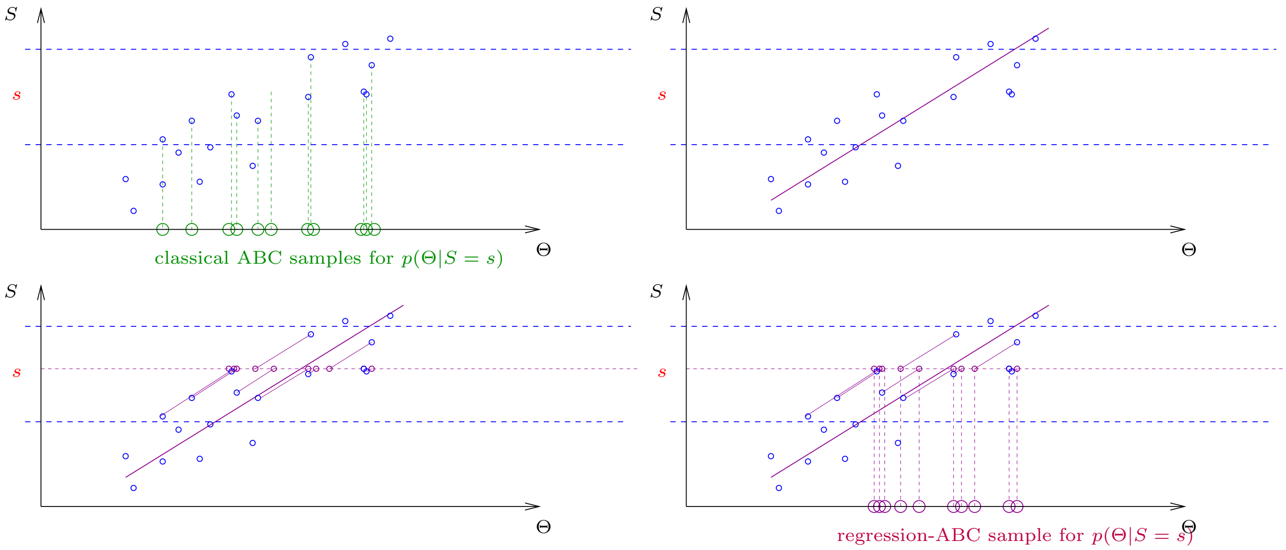
Only possible if a few summary statistics suffice. Otherwise acceptance will be rare.

Ideas of Beaumont, Zhang, Balding (2002):

- combine ABC with local regression:



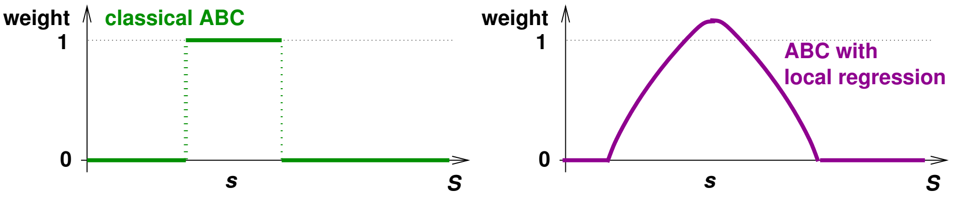
Simulate data for some parameter combinations Θ and compute corresponding s.



classical ABC samples for $p(\Theta|S = s)$

regression-ABC sample for $p(\Theta|S = s)$

- Accept in a wider range but put a smaller weight on s' if $|s - s'|$ is large.

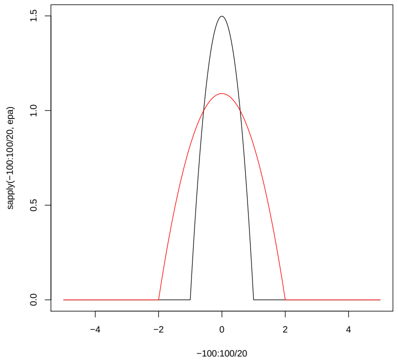


Epanechnikov-Kernel

$$K_\delta(t) = \begin{cases} c \cdot \left(1 - \left(\frac{t}{\delta}\right)^2\right) / \delta & \text{for} \quad t \leq \delta \\ 0 & \text{for} \quad t > \delta \end{cases}$$

where c is a the normalizing constant:

$$c = 1 \Bigg/ \int_{-\delta}^{\delta} \left(1 - \left(\frac{x}{\delta}\right)^2\right) / \delta \; dx$$



Epanechnikov-Kernels with

$$\delta = 1$$

and $\delta = 2$

Beaumont, Zhang, Balding (2002)

Simulate pairs $(\Theta^{(i)}, s^{(i)})$ and fit local regression model, i.e. find α and β to minimize

$$\sum_i \left(\Theta^{(i)} - \alpha - (s^{(i)} - s)^T \beta\right)^2 \cdot K_\delta(\|s^{(i)} - s\|),$$

where $||v|| = \sqrt{\sum_i v_i^2}$ (or some other vector norm).

Consider

$$\Theta_*^{(i)} = \Theta^{(i)} - (s^{(i)} - s)^T \widehat{\beta}$$

as random sample from $\Pr(\Theta \mid S = s)$.

Posterior density estimation:

$$\widehat{p}(\Theta_0 \mid S = s) = \frac{\sum_i K_\Delta(\Theta_*^{(i)} - \Theta_0) \cdot K_\delta(\|s - s^{(i)}\|)}{\sum_j K_\delta(\|s - s^{(j)}\|)}$$

where Δ = density estimation bandwidth.

Solution of the local regression problem

Solution for j-th parameter: $(\widehat{\alpha}, \widehat{\beta_1}, \dots, \widehat{\beta_k}) = \left(X^T W X\right)^{-1} X^T W \Theta^{(j)}$, where

$\Theta^{(j)} = \begin{pmatrix} \Theta_1^{(j)} \\ \Theta_2^{(j)} \\ \vdots \\ \Theta_m^{(j)} \end{pmatrix}$: Values of the j-th parameter from m simulations,

$s = (s^{(1)}, \dots, s^{(k)})$: Vector of summary statistics for observed data,

$s_i = (s_i^{(1)}, \dots, s_i^{(k)})$: Vector of summary statistics from i-th simulation,

$X = \begin{pmatrix} 1 & s_1^{(1)} - s^{(1)} & \cdots & s_1^{(k)} - s^{(k)} \\ 1 & s_2^{(1)} - s^{(1)} & \cdots & s_2^{(k)} - s^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_m^{(1)} - s^{(1)} & \cdots & s_m^{(k)} - s^{(k)} \end{pmatrix}$ and

W is diagonal matrix with diagonal entries $K_\delta(||s_1 - s||), \dots, K_\delta(||s_m - s||)$.

Beaumont, Zhang, Balding (2002)

ABC with local regression

1. Select summary statistics $S = (S_i)_i$ and compute their values $s = (s_i)_i$ for given data set
2. Choose tolerance δ and bandwidth Δ
3. repeat for $i = 1, \ldots, m$:
   (a) Simulate $\Theta^{(i)}$ from prior distribution of Θ
   (b) Simulate genealogy G according to $\Pr_{\Theta^{(i)}}(G)$.
   (c) Simulate data and compute values $s^{(i)}$ of S
4. $(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{m} \left(\Theta_i - \alpha - (s^i - s)^T \beta\right)^2 \cdot K_\delta(||s^i - s||)$
5. $$\Theta_*^{(i)} := \Theta^{(i)} - (s^{(i)} - s)^T \widehat{\beta}$$
6. Approximate $p(\Theta|S = s)$ by
$$\frac{\sum_i K_\Delta(\Theta_*^{(i)} - \Theta) \cdot K_\delta(\|s - s^{(i)}\|)}{\sum_j K_\delta(\|s - s^{(j)}\|)}$$

Summary statistics used by Beaumont et al. (2002) for microsatellite data:

1. mean accross loci in the variance of repeat numbers
2. mean effective heterozygosity
3. number of distinct haplotypes
4. mean accross loci of kurtosis of repeat numbers
5. variance accross loci of variance of repeat numbers
6. mean accross loci of maximum allele-frequency
7. multivariate kurtosis
8. linkage disequilibrium (LD) measured with Hudson's Δ^2

Marjoram et al. (2003) MCMC without likelihoods

Aim: For given data D with summary statistics $S = s$ sample paramter vectors according to $p(\Theta \mid ||S - s|| \leq \varepsilon)$.

1. If current parameter estimation is Θ', propose Θ^* with probability $Q_{\Theta' \to \Theta^*}$
2. Simulate data D^* according to Θ^* and compute their summary statistics s^*.
3. If $||s^* - s|| > \varepsilon$ reject proposal, else accept with probability
$$\min\left\{1, \frac{p(\Theta^*) \cdot Q_{\Theta^* \to \Theta'}}{p(\Theta') \cdot Q_{\Theta' \to \Theta^*}}\right\}.$$
4. repeat steps 1 to 4.

Application example: Nuu Chah Nulth data, n=63 samples of HVR-I.

Estimate θ and time to the MRCA based on F84 substitution model.

Summary statistics: number of variable sites and number of haplotypes.

Simple approach: when updating parameters, generate entirely new tree.(will usually be rejected $\rightsquigarrow$ inefficient.)

Compromise: keep some information about the tree an modify it slightly for next step:

1. tree topology
2. times of coalescence events
3. number of mutations between two coalescents events

Beaumont, Zhang, Balding (2002)

> *"[...] the MCMC-based method is consistently superior to the summary-statistics-based methods and highlights that it is well worth making the effort to obtain full-data inferences if possible."*

> *"[...] there are advantages to the use of summary statistics, both in the ease of implementation and in the time to obtain the results [...]"*

> *"Further research is needed to find a more rigorous way for choosing summary statistics, including the use of orthogonalization and 'projection-pursuit' methods"*

Wegmann et al. (2009)

- combine MCMC-ABC with Beaumont et al.'s regression approach to sample from $p(\Theta|||S-s|| \leq \varepsilon)$.
- apply Box-Cox transformation to each summary statistic with respect to the parameter of interest, based on simulated data
- apply partial least squares (PLS) to find combinations of summary statistics that are informative wrt the parameter of interest
- leave-one-out cross validation to optimize number of PLS components used

Simulation studies show improvements compared to other ABC methods but IMa is still better.
Wegmann et al. "[..] would not recommend using an ABC approach if a full-likelihood method exists [..]".

Box-Cox transformation

$$X^{(\lambda)} = \begin{cases} \frac{(X+c)^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(X+c) & \text{for } \lambda = 0 \end{cases}$$

Idea: fit λ and c such that the residuals of the regression model $Y = \alpha + \beta X$ look as normally distributed as possible.

partial least squares (PLS)

Aim: find combinations of explanatory variables $x_1, \ldots, x_m$ that have highest correlation with variable y.

let y be centered and x_j be normalized, i.e. $\mu_y = 0$, $\mu_{x_j} = 0$, $\sigma_{x_j} = 1$.

1. (n-fold of) univariate regression coefficient: $\varphi_j := \langle x_j, y\rangle := \sum_i x_{ji} y_i$
   $\Rightarrow y \approx \varphi_j \cdot x_j$

2. first partial least squares direction: $z_1 := \sum_j \varphi_j \cdot x_j$

3. first regression coefficient: $\delta := \frac{\langle z_1, y \rangle}{\langle z_1, z_1 \rangle}$
   $\Rightarrow y \approx \delta \cdot z_1$

4. now orthogonalize $x_1, x_2, \ldots, x_m$ with respect to z_1: $x_j^{(2)} := x_j - \frac{\langle z_1, x_j \rangle}{\langle z_1, z_1 \rangle} \cdot z_1$

5. and compute the residuals: $y^{(2)} := y - \delta \cdot z_1$

repeat 1-5 with x_j and y replaced by $x_j^{(2)}$ and $y^{(2)}$. $\rightsquigarrow z_2, x_j^{(3)}, y^{(3)}$

iterate to get $z_1, z_2, \ldots, z_m$.

Comparison PCA vs. PLS

Let S be the covariance matrix of the vectors x_i. Then, the principal component directions $v_1, \ldots, v_m$ satisfy:

$$v_j = \operatorname*{arg\,max}_\alpha \left\{ \operatorname{Var}\left(\sum_i x_i \alpha_i\right) \,\middle|\, ||\alpha|| = 1, \forall_{\ell<j} v_\ell^T S\alpha = 0 \right\}$$

The PLS directions $\varphi_1, \ldots, \varphi_m$ satisfy:

$\varphi_j = \operatorname{arg\,max}_\alpha \left\{ \operatorname{Corr}^2\left(y, \sum_i x_i\alpha_i\right) \operatorname{Var}\left(\sum_i x_i\alpha_i\right) \,\middle|\, ||\alpha|| = 1, \forall_{\ell<j} \varphi_\ell^T S\alpha = 0 \right\}$

10 The program STRUCTURE

References

[PSD00] Pritchard, Stephens, Donnelly (2000) Inference of Population Structure Using Multilocus Genotype Data *Genetics* **155**: 945–959

[FSP03] Falush, Stephens, Pritchard (2003) Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* **164**: 1567–1587

[FSP07] Falush, Stephens, Pritchard (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes*

[HFSP09] Hubisz, Falush, Stephens, Pritchard (2009) Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resources* **9**: 1322–1332

10.1 no admixture, no sampling locations

Structure: A program for model-based clustering of genotypes (Microsatellites, SNPS, AFLPs, ...)

N diploid individuals, L loci, K (sub)populations

unknown which individuals belong to which population, even if sampling locations are known, i.e. subpopulations may not correspond to sampling locations.

known is the genotype of individual each i at locus ℓ:

$$X = (x_\ell^{(i,1)}, x_\ell^{(i,2)})_{i \le N, \ell \le L}$$

unknown are the populations from which individual i originates:

$$Z = (z^{(i)})_{i \le N}$$

and the frequencies of allele j at locus ℓ in population k:

$$P = (p_{k\ell j})_{k \le K, \ell \le L, j \le J_\ell}$$

Assumption 1: each population is in Hardy-Weinberg equilibrium

Assumption 2: linkage equilibrium between loci

Bayesian approach: approximate sample from

$$\Pr(Z,P \mid X) \propto \Pr(Z) \cdot \Pr(P) \cdot \Pr(X \mid Z,P)$$

Priors for origin population of individual i:

$$\Pr(z^{(i)} = k) = 1/K$$

Dirichlet prior for allele frequencies in each population:

$$p_{k\ell} \sim \mathcal{D}(\lambda_1, \lambda_2, \dots, \lambda_{J_\ell}) \text{ with } \lambda_1 = \lambda_2 = \ldots = \lambda_{J_\ell} = 1$$

(uniform distribution on all distributions)

$\Pr(X|Z,P)$:

$$\Pr(x_\ell^{(i,a)} = j) = p_{z^{(i)}\ell j}$$

Dirichlet distribution

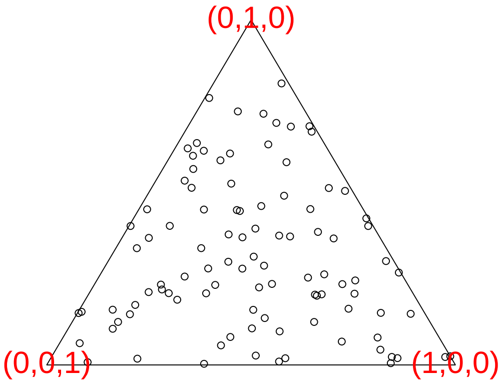
If $Y \sim \mathcal{D}(\alpha_1, \dots, \alpha_k)$ then

$$\Pr(Y = (y_1, \dots, y_k)) = c(\alpha) \cdot \prod_{i=1}^{k} y_i^{\alpha_i - 1}$$

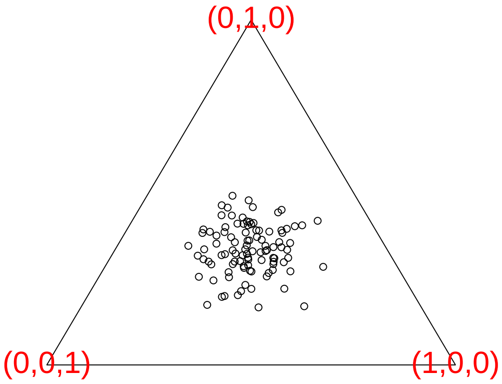
if all $y_i \geq 0$ and $\sum_i y_i = 1$, else 0.

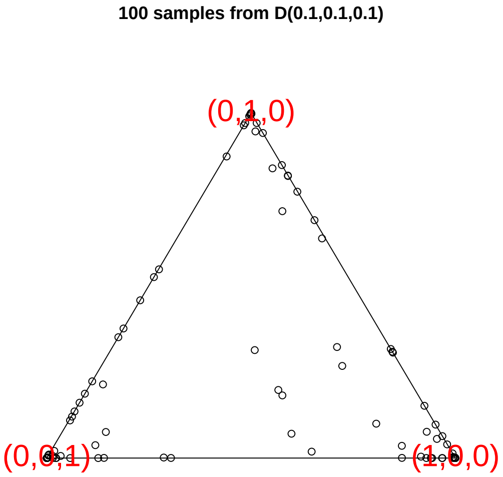
$$\mathbb{E}(Y) = \frac{(\alpha_1, \dots, \alpha_k)}{\sum_i \alpha_i}$$

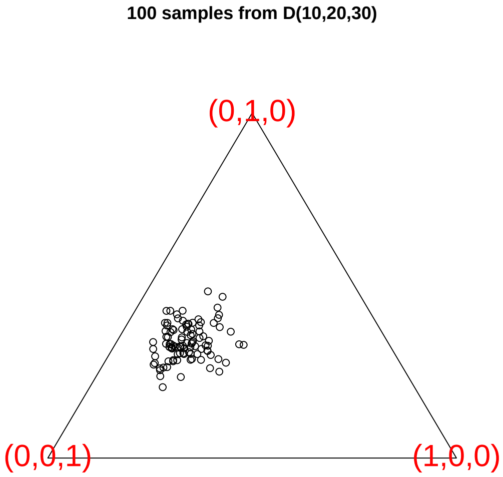
100 samples from D(1,1,1)



100 samples from D(10,10,10)







Important property of Dirichlet distributions

Let $N = (n_1, \ldots, n_K)$ multinomially distributed with (unknown) probabilities $P = (p_1, \ldots, p_K)$, i.e.

$$\Pr(N = (n_1, \ldots, n_m)) = \frac{(n_1 + n_2 + \cdots + n_k)!}{n_1! \cdot n_2! \cdots n_k!} \prod_{i=1}^{k} p_i^{n_i}.$$

If the prior distribution of P is $\mathcal{D}(\lambda_1, \ldots, \lambda_k)$, then the posterior distribution of P given $N = (n_1, \ldots, n_k)$ is

$$\mathcal{D}(\lambda_1 + n_1, \ldots, \lambda_k + n_k).$$

(Exercise!)

MCMC method for sampling from $\Pr(Z, P|X)$: Start with $Z^{(0)}$ (e.g. sampled from prior) and iterate 2 steps for $m = 1, 2, 3, \ldots$:

1. Sample $P^{(m)}$ from $\Pr(P|X, Z^{(m-1)})$

   $$p_{k\ell\cdot}|X, Z \sim \mathcal{D}(\lambda_1 + n_{k/\ell}, \ldots, \lambda_{J_\ell} + n_{k/J_\ell}),$$

   where $n_{k/j} = \#\left\{(i,a)|x_\ell^{(i,a)} = j \text{ and } z^{(j)} = k\right\}$. (using the important property of the Dirichlet distribution.)

2. Sample $Z^{(m)}$ from $\Pr(Z|X, Z^{(m-1)}, P^{(m)})$

   $$\Pr(z^{(j)} = k|X, P) = \frac{\Pr(x^{(j)}|P, z^{(j)} = k)}{\sum_{k'=1}^{K} \Pr(x^{(j)}|P, z^{(j)} = k'},$$

   using $\Pr(x^{(j)}|P, z^{(j)} = k) = \prod_{\ell=1}^{L} p_{k\ell x_\ell^{(j,1)}} \cdot p_{k\ell x_\ell^{(j,2)}}$.

10.2 with admixture

admixture: present individuals stem from k populations that were admixed recently.

$Q: \left(q_k^{(j)}\right)_{j \le N, k \le K}$ = proportion of individual j's genome that origins from population k

$Z: \left(z_\ell^{(i,a)}\right)$ = population of origin of allele copy $x_\ell^{(i,a)}$

$$\Pr\left(x_\ell^{(i,a)} = j \,\middle|\, Z, P, Q\right) = p^{(i,a)}_{z_\ell, l_j}, \qquad \Pr\left(z_\ell^{(i,a)} = k \,\middle|\, P, Q\right) = q_k^{(i)}$$

Prior on Q:

$$q^{(i)} = \left(q_1^{(i)}, \ldots, q_k^{(i)}\right) \sim \mathcal{D}(\alpha, \ldots, \alpha),$$

where α is also random with prior $\alpha \sim \text{unif}([0, \alpha_{\max}])$.

Note:

$$\begin{aligned} \alpha = 0 \quad &\Leftrightarrow \quad \text{no admixture} \\ \alpha \to \infty \quad &\Leftrightarrow \quad \text{all completely admixed} \end{aligned}$$

MCMC for case of admixture

Start with initial $P^{(0)}$, $Q^{(0)}$, $Z^{(0)}$ and $\alpha^{(0)}$ and iterate for $m = 1, 2, \ldots$:

1. Sample $P^{(m)}$ and $Q^{(m)}$ from $\Pr(P, Q|X, Z^{(m-1)})$:

   update $p^{(}_{z_\ell, \ell_j} i, a)$ based on the number of ℓ copies of type j that come from population k

   $$n_{klj} = \left\{(i, a)|x_\ell^{(i,a)} = j \text{ and } z_\ell^{(i,a)} = k\right\}$$

   and sample $q^{(i)}|X, Z$ according to

   $$\mathcal{D}\left(\alpha + \#\left\{(\ell, a) : z_\ell^{(i,a)} = 1\right\}, \ldots, \alpha + \#\left\{(\ell, a) : z_\ell^{(i,a)} = K\right\}\right)$$

2. Sample $Z^{(m)}$ from $\Pr(Z|X, P^{(m)}, Q^{(m)})$ according to:

   $$\Pr\left(z_\ell^{(i,a)} = k \,\middle|\, X, P\right) = \frac{q_k^{(i)} \cdot p_{k\ell x_\ell^{(i,a)}}}{\sum_{h=1}^K q_h^{(i)} \cdot p_{h\ell x_\ell^{(i,a)}}}$$

3. Metroplis Hastings step $\alpha^{(m-1)} \rightsquigarrow \alpha^{(m)}$:

   propose $\alpha' \sim \mathcal{N}(\alpha, \text{some } \sigma^2)$, reject immediately if $\alpha' < 0$, else perform MH step.

Inference for Z, P, Q from MCMC samples

for example for Q it seems obvious to estimate

$$\mathbb{E}(q_i|X) \approx \frac{1}{M} \sum_{m=1}^M q_i^{(m)},$$

but the theoretical posterior mean is

$$\mathbb{E}(q_i|X) = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$$

due to symmetries in the model (numbering of populations exchangeable).

$\rightsquigarrow$ use modes of $\left(q_i^{(1)}, \ldots, q_i^{(M)}\right)_i$ instead of means or use Noah Rosenberg's software CLUMPP to evaluate STRUCTURE output.

Inference for the number K of populations

$$\Pr(K|X) \propto \Pr(X|K) \cdot \Pr(K)$$

can be approximated using the harmonic mean estimator

$$\Pr(X|K) \approx M \Bigg/ \sum_{i=1}^{M} \frac{1}{\Pr\left(X \left| K, Z^{(i)}, P^{(i)}, Q^{(i)}, \alpha^{(i)} \right.\right)},$$

but the harmonic mean estimator is know to be imprecise.

Instead, we hope that $-2\log L(Z, \widehat{P, Q}, \alpha|X)$ is approximately normally distributed and estimate

$$\Pr(X|K) \approx e^{-\widehat{\mu}/2 - \widehat{\sigma}^2/8}$$

with $\widehat{\mu} = \frac{1}{M}\sum_{i=1}^{M} -2\log\Pr\left(X|Z^{(i)}, P^{(i)}, Q^{(i)}, \alpha^{(i)}\right)$

and $\widehat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}\left(-2\log\Pr\left(X|Z^{(i)}, P^{(i)}, Q^{(i)}, \alpha^{(i)}\right) - \widehat{\mu}\right)^2$

Pritchard et al. write about this approximation:

> "*In fact the assumption underlying* [this] *are dubious at best, and we do not claim (or believe) that our procedure provides a quantitatively accurate estimate of the posterior distribution of K. We see it merely as an ad hoc guide to which models are most consistent to the data, with the main justification being that it seems to give reasonable answers in practice.*"

and:

> "*The inferred value of K may not always have a clear biological interpretation.*"

and about the multiple-modes problem:

> "[The] *Gibbs-sampler did not manage to move between two modes in any of the runs*"

Data examples

Bird example: Without using informations on sampling locations, STRUCTURE gave clear clusters corresponding to sampling locations, up to a few exceptions. Neighbor-Joining results did not show clear clusters when labels were removed.

Human data: Found $K \geq 2$ corresponding to African and European oringin of samples. Evidence for $K > 2$ may indicate substructure.

10.3 taking sampling locations into account

First attempt: populations correspond to sampling locations with a few migrants in the last few generations.

$g(i)$: sampling location of individual i

ν: probability that i is immigrant or offspring of an immigrant in the last G generations, where G is not too large.

$\Rightarrow q^{(i)}_{g(i)} = 1$ with probability $1 - \nu$ and for $t \leq G$:
$q^{(i)}_{g(i)} = 1 - 2^{-t}$ and $q^{(i)}_{j} = 2^{-t}$ with probability $\frac{2^t \nu}{(k-1)\sum_{T=0}^{G} 2^T}$ (neglecting the possibility of more than one migranting ancestor in the last G generations.)

in MCMC: sampling of $q^{(i)}$ is conditioned on X and P, and not on X and Z.
Falush et al. (2003) allow for LD between loci. Advantages:

1. detection of admixture further back into past

2. inference of population of origin of chromosomal regions

3. more accurate estimate od statistical incertainty when linked loci are used

Sources of LD:

mixture LD: variation in ancstry among sampled individuals (Prichard et al.)

admixture LD: correlation of ancestry along each chromosome causes additional LD between linked markers (Falush et al.)

background LD: within population decaying on a much shorter scale, e.g. tens of kb in humans. (not yet in STRUCTURE)

Approach of Falush et al. (2003):

- breakpoints occur as Poisson process at rate r
- uniform prior on $\log(r)$
- use HMM to sample from conditional distribution of Z
- data allowed to be unphased

more options: corelated allele frequencies between populations accordingt to star-shaped phylogeny of populations with drift rates $F_1, \dots, F_K$ and ancestral allele frequency distribution $p_A \sim \mathcal{D}(\lambda_1, \dots, \lambda_{J_\ell})$.

$$p_{k\ell.}|p_A \sim \mathcal{D}\left(p_{A\ell 1}\frac{1-F_1}{F_1}, \dots, p_{A\ell K}\frac{1-F_K}{F_K}\right)$$

(be careful with this model!)

Approach of Hubisz et al. (2009): Allow uncertainty in the information about sampling location

$$\begin{aligned}
r &\sim \text{unif}([0, r_{\max}]) \qquad \text{(informativeness of sampling location)} \\
q^{(i)} &\sim \mathcal{D}(\alpha_{h_1}, \dots, \alpha_{h_K}), \quad \text{if individual } i \text{ comes from location } h \\
\alpha_{h_k} &\sim \Gamma\left(r \cdot \alpha_k^{\text{glob}}, 1/r\right), \quad \text{(which entails that the mean is } \alpha_k^{\text{glob}}) \\
\alpha_k^{\text{glob}} &\sim \text{unif}(0, \alpha_{\max})
\end{aligned}$$

Hubisz et al.: "*However, we would still encourage users to run the original models as well, and to check that substantial differences between the results from the new and the old models seem biologically sensible.*"

When STRUCTURE has problems

- number of clusters not well-defined when allele frequencies vary slowly accross the landscape
- inbreeding or relatedness between individuals
  In this case, the software INSTRUCT may help, cf.

References

[GWB07] H. Gao, S. Williamson, S.D. Bustamante (2007) An MCMC Approach for Joint Inference of Population Structure and Inbreeding Rates from Multi-Locus Genotype Data. *Genetics (online)*

11 The PAC method

11.1 LD and recombination hotspots

Problems of models to estimate local recombination rates:

LAMARC etc. (ARG-based): not feasible for larger parts of the genome

Summary-statistics-based: lose too much information

some composite-likelihood methods: Hudson (2001), Fearnhead, Donnelly (2002), McVean (2002) assume fixed recombination rate along the genome

Li & Stephens' approach to analyze patterns of LD

References

[LS03] Na Li, Matthew Stephens (2003) Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data *Genetics* **165**

ideas:

- relate LD directly to underlying recombination process
- Sometimes, block-like LD structure is reported. True or artifact of LD mapping? Allow for both.
- consider all loci simultaneously, not pairwise
- should be compuationally tractable even for complete chromosomes

Li& Stephens' PAC approach

$h_1, h_2, \ldots, h_n$**:** haplotypes sampled from panmictic population with constant size and random mating

ρ**:** recombination parameter (may be a vector if recombination rate varies within the region of interest)

Product of Approximate Conditionals (PAC)

$$\Pr(h_1, \ldots, h_n|\rho) = \Pr(h_1) \cdot \Pr(h_2|h_1, \rho) \cdot \ldots \cdot \Pr(h_n|h_1, \ldots, h_{n-1}, \rho)$$

approximate $\Pr(h_k|h_1, \ldots, h_{k-1}, \rho)$ by simpler $q(h_k|h_1, \ldots, h_{k-1}, \rho)$.
Properties of $\Pr(h_k|h_1, \ldots, h_{k-1}, \rho)$

1. h_k is more likely to match another haplotype if the latter is frequent among $h_1, h_2, \ldots, h_{k-1}$
2. the probability of seeing a novel haplotype decreases as k increases
3. the probability of seeing a novel haplotype increases with $\theta = 4N_e\mu$.
4. if a new haplotype does not exactly match any previous one, it will differ from one of those only by a small number of mutations.
5. effect of recombination: the next haplotype will be composed by segments which are similar to segments in previously sampled haplotypes. These segments tend to be longer if recombination rates are low.

Assume the sampled haplotypes $h_1, h_2, \ldots, h_n$ are typed at S biallelic loci (e.g. SNPs).

$$q(h_1) = \left(\frac{1}{2}\right)^S$$

For the definition of $q(h_{k+1}|h_1, h_2, \ldots, h_k)$ let $X_i := j$ if at the i-th locus, the closest relative of h_{k+1} among $h_1, \ldots, h_k$ is h_j.

d_i distance between loci i and $i+1$

c_i recombination rate between loci i and $i+1$ per site an per generation

ρ_i $= 4N_e c_i$

The simplifying assuption is then that $X_1, X_s, \dots, X_S$ is a Markov chain on $\{1,\dots,k\}$ with $Pr(X_1 = j) = 1/k$ and

$$\Pr(X_{i+1} = j | X_i = \ell) = \begin{cases} (1-e^{-\rho_i d_i/k})/k & \text{if} \quad j \neq \ell \\ e^{-\rho_i d_i/k} + (1-e^{-\rho_i d_i/k})/k & \text{if} \quad j = \ell \end{cases}$$

Mutations

For SNP data we assume that each locus is hit by one mutation, such that

$$\widetilde{\theta} := 1 \Bigg/ \sum_{m=1}^{n-1} \frac{1}{m}$$

is assumed to be the corrected rate of mutations per SNP site. Note that this does not exclude double hits (just some bias if double hits are frequent.)

Then, with probability $\frac{k}{k+\widetilde{\theta}} + \frac{\widetilde{\theta}}{2(k+\widetilde{\theta})}$ the copy has the same type as the original

and with probability $\frac{\widetilde{\theta}}{2(k+\widetilde{\theta})}$ the haplotype has the other of the two possible alleles.

Compute $q(h_{k+1}|h_1,\dots,h_k)$ by HMM forward algo:

$h_{k+1,\leq j}$ $:= (h_{k+1,1},\dots,h_{k+1,j}) :=$ types of the first j sites in h_{k+1}

$\alpha_j(x)$ $:= \Pr(h_{k+1,lej}, X_j = x | h_1,\dots,h_k)$

(note that with mutations any $X_1,\dots,X_S$ can emit h_k.)
Then,

$$q(h_{k+1}|h_1,\dots,h_k) = \sum_{x=1}^{k} \alpha_S(x).$$

"dynamic programming": we can compute all $\alpha_j(x)$ by the recursion

$$\begin{aligned} \alpha_{j+1}(x) &= \Pr(h_{k+1,j+1}|X_{j+1} = x, h_1,\dots,h_k) \cdot \sum_{x'=1}^{k} \alpha_j(x') \cdot \\ &\qquad \Pr(X_{j+1} = x | X_j = x') \\ &= \Pr(h_{k+1,j+1}|X_{j+1} = x, h_1,\dots,h_k) \cdot \\ &\qquad \left(e^{-\rho_j d_j/k} \cdot \alpha_j(x) + \left(1 - e^{-\rho_j d_j/k}\right) \cdot \frac{1}{k} \sum_{x'=1}^{k} \alpha_j(x') \right) \end{aligned}$$

Bias correction

Simulations show that estimations of ρ based on q are biased.

For bias-correction replace ρ_j in the computation of $\Pr(X_{j+1} = x'|X_j = x)$ by

$$\rho_j \cdot e^{a + b\log_{10}\rho_j},$$

where a and b are fitted to simulated data, taking the numbers of haplotypes and segregating sites into account.

Models for ρ considered by Li and Stephens

1. constant ρ
2. single-hotspot model
3. all recombination rates $\rho_1, \rho_2, \ldots, \rho_{S-1}$ may differ

Software by Matthew Stephens using PAC: Hotspotter, PHASE

11.2 Population splitting and recombination

References

[DPC09] D. Davison, J.K. Pritchard, G. Coop (2009) An approximate likelihood for genetic data under a model with recombination and population splitting. *Theoretical Population Biology* **75**:331-345

- two populations split G generations ago
- no ongoing geneflow
- for simplicity: assume that both populations and the ancestral population have size N
- Copying occurs in daughter population ($S = d$) and in ancestral population ($S = a$)

to be specified:

1. prob of hidden copying states (S_ℓ, X_ℓ) at a single site ℓ.
   unlinked case:

$$\begin{aligned}
\Pr(X_\ell = i | S_\ell = d) &= \begin{cases} \frac{1}{k_{z_*}} & \text{if } z_* = z_i \\ 0 & \text{else} \end{cases} \\
& \quad \text{where } k_{z_*} \text{ is the no. of lineages sampled} \\
& \qquad\quad \text{from pop. } z_* \text{ so far} \\
\Pr(X_\ell = i | S_\ell = a) &= \mathbb{E}\left(\frac{J_{z_i}}{J_1 + J_2}\right) \cdot \frac{1}{k_{z_i}},
\end{aligned}$$

   Where J_{z_i} is the number of ancestral lineages that enter the ancestral pop. from pop. z_i

2. probability of new allelic state conditioned on the state of the copied allele and the level S_ℓ.
3. Transition probabilities between the hidden copying state at adjacent states

In case of loosely linked data: combine with HMM methods.

11.3 Diversifying selection and recombination

References

[WM06] Wilson, McVean (2006) Estimating diversifying selection and functional constraints in the presence of recombination *Genetics* **172**:1411–1425