ADMIXMAP statistical methods

Clive J. Hoggart

Paul M. McKeigue

Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland. Tel: +353 1 716 6952

paul.mckeigue@ucd.ie

1. Introduction

These notes briefly describe the statistical model and the algorithms used in ADMIXMAP. It is intended for users requiring more detail of the model than is given in the manual and in our published papers, and for developers working on the source code. For documentation on how to use the program, see the user manual.

Note: This document is a rough draft and may not accurately represent the current program. Last Updated 28 Jan 2005.

2. Model for genotypes and haplotypes at a compound locus (implemented in class CompositeLocus)

At a simple locus, there are S possible alleles, numbered from 1 to S. We observe unphased genotypes at some of these loci. Each possible unphased genotype can be represented as a pair of unsigned integers. If the two integers in this pair are different, the individual is heterozygous. Where the genotype is missing, both alleles are missing. Missing alleles can be represented as 0.

A compound locus is a sequence of one or more adjacent simple loci that are separated by zero map distance.

A haplotype is a sequence of alleles at the L simple loci within a compound locus. The alleles that specify each possible haplotype can thus be represented by a vector of unsigned integers.

The number H of possible haplotypes at a compound locus with $s_1, \ldots, s_L$ alleles at the L simple loci is $H = s_1 s_2 \ldots s_L$. Thus each possible haplotype can be represented by an unsigned integer between 1 and H.

For each individual at each compound locus, there is a pair of haplotypes which can be represented by a pair of unsigned integers between 1 and H. If the genotype

at any simple loci are missing, or there is more than one locus in the compound locus and the individual is heterozygous at more than one of these simple loci, the observed genotype data do not uniquely assign a pair of haplotypes.

For a single individual at a compound locus, there are K possible ancestry states on the paternal and maternal gametes: thus K^2 possible ordered ancestry states represented as a pair of unsigned integers between 1 and K.

The class CompositeLocus should have the following methods

1. to return all possible ordered haplotype pairs (as pairs of integers) given the observed genotypes (some of which may be missing) at the L simple loci within the compound locus. This method should be called only once for each locus and each individual, and the results should be stored by the Individual object.

2. to calculate the likelihood of each possible ordered state of locus ancestry, given the observed genotype data and the ancestry-specific haplotype frequencies.

3. to sample an ordered haplotype pair from the list of possible haplotype pairs, given the ordered states of locus ancestry and the ancestry-specific haplotype frequencies.

4. to return the sampled haplotype pair, given the ordered states of locus ancestry, as an array of dimension $H \times K$ giving the realized counts of each haplotype in each subpopulation.

5. to return the vector of alleles that specify the haplotype, given a haplotype coded as an unsigned integer between 1 and H.

*** comment - it's probably possible to speed up computation by writing special algorithms for the simplest situation where the compound locus consists of one diallelic simple locus, and thus each haplotype consists of a single allele that can be stored as a bit.

We observe unphased multilocus genotypes $y_{11}, \ldots, y_{1L}, \ldots, y_{N1}, \ldots, y_{NJ}$ on N unrelated individuals $i = 1, \ldots, N$ typed at $j = 1, \ldots, J$ compound loci,
At the jth compound locus, there are H possible haplotypes. The realized haplotype pair in the ith individual is x_{ij1}, x_{ij2}. We have K ancestral populations. The ancestry at locus j of individual j on the gth gamete is denoted by $\boldsymbol{A}_{ijg}$. The paternal and maternal admixture proportions of individual i are denoted by vectors $\boldsymbol{\theta}_{i1}$ and $\boldsymbol{\theta}_{i2}$ respectively.

The ancestry-specific haplotype frequency is defined as the probability of haplotype h on gth gamete of ith individual at diallelic locus j, given ancestry from kth population

$$p(X_{ijg} = h \mid A_{ijg} = k, \phi_{jk.}) = \phi_{jkh}$$

The probability of the observed multilocus genotype, given the ancestry of the paternal and maternal gametes at that locus, is the sum of the probabilities of all ordered haplotype pairs that are compatible with the observed genotype.

For each compound locus, we calculate a vector Λ that specifies for each possible ordered diploid ancestry state the probabilities of the observed (unphased) multilocus genotype conditional on the realized haplotype frequencies

$$\mathbf{\Lambda}_j = \begin{pmatrix} p(g_j \mid A_j = 1) \\ \vdots \\ p(g_j \mid A_j = K) \end{pmatrix}$$

*** comment - these vectors need only be computed once, for all observed multilocus genotypes, when the haplotype frequencies are updated. Should fix this method in class CompositeLocus

3. Model for locus ancestry, individual admixture and population admixture

Variation of ancestry along chromosomes of a single gamete is modelled as the sum of K independent Poisson arrival processes, with a parameter ρ for the sum of the intensities of these arrival processes. For loci $j-1, j$ separated by distance d_j

$$p(A_{ij} = k \mid A_{i,j-1}, \boldsymbol{\theta}, \rho) = \delta_{A_{i,j-1}k} \exp -\rho d_j + (1 - \exp -\rho d_{ij})\theta_{ik}$$

This in turn specifies the variation of ancestry along chromosomes as a Markov process, for which the transition matrices between two loci separated by a genetic map distance d morgans can be derived.

Thus for a three-state arrival process with intensities α, β and γ of states 1, 2 and 3 respectively, the instantaneous transition matrix (generator matrix) is given by

$$\boldsymbol{G} = \begin{pmatrix} -\beta-\gamma & \beta & \gamma \\ \alpha & -\alpha-\gamma & \gamma \\ \alpha & \beta & -\alpha-\beta \end{pmatrix}$$

From this the matrix of haploid transition probabilities (on a single gamete can be derived as

$$\begin{aligned} &\boldsymbol{P}(x) \\ &\quad = \frac{1}{\rho} \begin{pmatrix} \alpha + (\beta+\gamma)\exp\{-\rho d\} & \beta - \beta\exp\{-\rho d\} & \gamma - \gamma\exp\{-\rho d\} \\ \alpha - \alpha\exp\{-\rho d\} & \beta + (\alpha+\gamma)\exp\{-\rho d\} & \gamma - \gamma\exp\{-\rho d\} \\ \alpha - \alpha\exp\{-\rho d\} & \beta - \beta\exp\{-\rho d\} & \gamma + (\alpha+\beta)\exp\{-\rho d\} \end{pmatrix} \end{aligned}$$

where $\rho = \alpha + \beta + \gamma$ and d is the map distance between the two loci of interest. The transition matrix is thus specified by the sum of intensities parameter ρ, the genetic map distance x, and the gamete admixture proportions $\theta_1, \theta_2, \theta_3$.

$$\boldsymbol{P} = \begin{pmatrix} f + (1-f)\theta_1 & (1-f)\theta_2 & (1-f)\theta_3 \\ (1-f)\theta_1 & f + (1-f)\theta_2 & (1-f)\theta_3 \\ (1-f)\theta_1 & (1-f)\theta_2 & f + (1-f)\theta_3 \end{pmatrix}$$

where $\theta_1 = \frac{\alpha}{\alpha+\beta+\gamma}$ and $f = \exp\{-(\alpha+\beta+\gamma)x\}$.

In the transition matrix, the columns index the population at locus $j+1$ and the rows index the population at locus j.

From the haploid transition matrices of order K we can calculate the transition matrix for ordered diploid states, of order K^2.

4. Generalising the model for population admixture (not yet implemented)

We should allow for models in which the distribution of admixture in the population is not unimodal, or where the population includes both admixed and unadmixed individuals. We could achieve this using a mixture of Dirichlet distributions.

Priors

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}) &\propto \sum_l \pi_l \mathrm{Di}(\boldsymbol{\alpha}_l) \\ p(\boldsymbol{\pi}) &= \mathrm{Di}(\tau) \end{aligned}$$

Full conditional

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}) = \sum_l \delta_l \mathrm{Di}(\boldsymbol{\alpha}_l)$$

where

$$\delta_l = \pi_l \frac{n!}{\left(\sum_l \alpha_l\right)^n} \prod_l \frac{\alpha_l^{\sum_j A_{jl}}}{\left(\sum_j A_{jl}\right)!} \tag{1}$$

Equation 1 is the multinomial-Dirichlet likelihood for the realized count of ancestry state arrivals $\boldsymbol{A}$, conditional on the prior. These weights are the same as those used in the Dirichlet process.

4.1. Priors on population admixture and sum of intensities parameters

If the option globalrho=1 is specified, the sum of Poisson intensities parameter ρ is assigned the prior

$$p(\rho) = \mathrm{Ga}(\rho \mid \rho_0, \rho_1)$$

Default values are 5 for the the shape parameter ρ_0 and 1 for the location parameter ρ_1.
Alternatively, if the option globalrho=0 is specified, a hierarchical model is specified with a sum of intensities parameter ρ for each gamete.

Parental admixture proportions $\boldsymbol{\theta}_i$ are distributed in the population as $\mathrm{Di}(\boldsymbol{\theta}_i \mid \boldsymbol{\alpha})$. The hyperparameters $\boldsymbol{\alpha}$ of this Dirichlet distribution are specified with independent gamma prior distributions, with parameters ϵ_0, ϵ_1.

$$p(\alpha_k \mid \epsilon_k) = \mathrm{Ga}(\alpha_k \mid \epsilon_0, \epsilon_1), \quad k = 1, \ldots, K$$

The admixture proportions of the two parental gametes can be drawn independently from the Dirichlet distribution (option randommatingmodel=1) or specified to be the same.

5. Model for haplotype frequencies (implemented in class AlleleFrequencies)

If the option priorallelefreqfile is specified, at each locus the ancestry-specific haplotype frequency vector $\phi_{jk\cdot}$ has a Dirichlet prior distribution $\mathrm{Di}\,(\alpha_1, \ldots, \alpha_H)$. The vector $(\alpha_1, \ldots, \alpha_H)$ can be specified by the user. The haplotype frequencies $\phi_{jk\cdot}$ are updated as a conjugate Dirichlet update, using the realized vector of counts of each haplotype on gametes that have ancestry from subpopulation k at locus j.

If no prior on haplotype frequencies are supplied, the haplotype frequencies are given an uninformative prior.

If option fixedallelefreq is specified, the haplotype frequencies are fixed.

5.1. Hierarchical (dispersion) model for ancestry-specific allele frequencies

For a given locus locus and subpopulation, we specify
$\phi^{(1)}$ - ancestry-specific haplotype frequencies within the admixed population
$\phi^{(2)}$ - haplotype frequencies in modern unadmixed descendants

If the locus has H haplotypes, we assume the ϕ's are distributed as

$$\boldsymbol{\phi}^{(i)} \sim \mathrm{Di}(\mu_1, \ldots, \mu_{n-1}, \eta), \quad i = 1, 2,$$

This specifies that $\phi^{(1)}$ and $\phi^{(1)}$ are draws from a Dirichlet prior, with parameter vector of length H, the elements of which sum to η. The dispersion parameter η indexes the dispersion of allele (haplotype) frequencies between modern unadmixed descendants and the corresponding ancestry-specific allele frequencies in the admixed population.

The priors for μ and η are specified as

$$\begin{aligned} \eta &\sim \mathrm{Ga}(\psi, \tau) \\ \frac{\mu_i}{\eta} &\sim \mathrm{Be}(1, 1) \end{aligned}$$

with the constraints that

$$\mu_i \geq 0.5, \quad \eta \geq 0.5 + \sum \mu_i.$$

*** check - are we really using these priors? ?

The prior distributions for μ_l and η were chosen to be uninformative but to give little prior weight at extreme values (0 or 1 for μ_i, 0 or large values for η). This helps to make the computation robust.

6. Regression model for dependence of outcome variable on individual admixture and covariates specified by the user (implemented in class Regression)

The current version of the program allows the user to specify either a linear regression model for a quantitative trait, or a least-squares regression model for a binary trait.

7. Sampling for parental admixture (implemented in class Individual)

To sample individual admixture, we introduce for each individual an array of binary latent variables $\boldsymbol{\xi}$, in which rows index compound loci, and columns index gametes.

$\boldsymbol{\xi} = (\xi_{01}, \xi_{02}, \ldots, \xi_{m-1,1}, \xi_{m-1,2})$

where $\xi_{jg} = 1$ if at least one arrival has occurred between $j-1, k$ and j, k (thus $\xi_{1g} = 1$). We define a vector of distances $\boldsymbol{d} = (d_1, \ldots, d_m)$ where d_j is the distance between $j-1$ and j. The ancestry states A_j at each locus on each gamete are sampled using a hidden Markov model forward-backward algorithm as described later. For each gamete, the jump indicators ξ_j are then sampled conditional on A_j, A_{j-1}.

$$p(\xi_{ij}) = \mathrm{Br}(1 - \exp -\rho d_j)$$

The likelihood can then be written as

$$p(A_{ij} = k \mid A_{i,j-1}, \xi_{ij}, \boldsymbol{\theta}, \rho) = \delta_{A_{i,j-1}k}(1 - \xi_{ij}) + \xi_{ij}\boldsymbol{\theta}_{ik})$$

where d_j is the distance from locus $j-1$ to locus j.

If there is no regression model, the likelihood for parental admixture is then a conjugate Dirichlet likelihood, with parameters calculated by adding the realized counts of ancestry states on the gamete at loci where there has been at least one arrival ($\xi = 1$) to the Dirichlet prior.

*** comment - if not already implemented, probably it's quicker to sample the total number λ of arrivals in each interval directly, then set ξ as an indicator variable for $\lambda > 0$.

If a regression model has been specified, the likelihood for parental admixture is the product of the Dirichlet likelihood and the regression model likelihood for this observation.

*** check how this is done, and where it is implemented

7.1. Sampling for population admixture parameters $\boldsymbol{\alpha}$ (implemented in class Latent)

The full conditional densities for the coordinates of the Dirichlet parameter vector $\boldsymbol{\alpha}$ are proportional to the product of the gamma prior density and the Dirichlet likelihood.

$$p(\alpha_k \mid \alpha_{-k}, \epsilon_0, \epsilon_1) \propto \left(\frac{\Gamma\left(\alpha_k + \sum_{l=1, l\neq k}^{d} \alpha_l\right)}{\Gamma(\alpha_k)} \right)^n$$

$$\alpha_k^{\epsilon_0 - 1} \exp\left\{ -\alpha_k \left(\epsilon_1 - \sum_{i=1}^{n} \log \theta_i \right) \right\} p(\boldsymbol{\theta}_i \mid \boldsymbol{\alpha}, \boldsymbol{A}_i) = \mathrm{Di}_K \left(\boldsymbol{\theta}_i \middle| \boldsymbol{\alpha} + \sum_{j=1}^{m} \xi_{ij} \boldsymbol{A}_{ij} \right)$$

We sample from this density using an adaptive rejection sampler for α_k. This requires the log density and its derivative

$$\log f(\alpha_k) = n \log\Gamma\left(\alpha_k + \sum_{l=1,l\neq k}^{d} \alpha_l\right) - n\log\Gamma(\alpha_k) - \alpha_k\left(\epsilon_1 - \sum_{i=1}^{n}\log\theta_i\right) ... + (\epsilon_0 - 1)\log\alpha_k$$

$$\frac{d}{d\alpha_k}\log f(\alpha_k) = n\Psi\left(\alpha_k + \sum_{l=1,l\neq k}^{d} \alpha_l\right) - n\Psi(\alpha_k) - \left(\epsilon_1 - \sum_{i=1}^{n}\log\theta_i\right) + \frac{\epsilon_0 - 1}{\alpha_k}$$

This density is log-concave since

$$\frac{d^2}{d{\alpha_k}^2}\log f(\alpha_k) = n\Psi'\left(\alpha_k + \sum_{l=1,l\neq k}^{d} \alpha_l\right) - n\Psi'(\alpha_k) - \frac{\epsilon_0 - 1}{\alpha_k^2}$$

and Ψ', the trigamma function, is decreasing.

7.2. Sampling for global sum of intensities parameter ρ

In the current version of the program, the global sum of intensities parameter is sampled conditional on the binary latent variables ξ, using an adaptive rejection sampler.

Given the ξ's the conditional distribution of ρ is

$$\begin{aligned} p(\rho \mid \dots) &\propto \pi(\rho)\prod_{i=1}^{n}\prod_{j=1}^{m}(1-\exp\{-\rho d_j\})^{\xi_{ij}}\exp\{-\rho d_j(1-\xi_{ij})\} \\ &= \pi(\rho)\exp\left\{-\rho\sum_{i,j}(1-\xi_{ij})d_j\right\}\prod_{i,j}(1-\exp\{-\rho d_j\})^{\xi_{ij}} \qquad (2) \end{aligned}$$

The log density is

$$\log f(\rho) = -\rho\left(\rho_0 + \sum_{i,j}(1-\xi_{ij})d_j\right) + \sum_j \xi_{.j}\log(1-\exp\{-\rho d_j\}) + (\rho_1 - 1)\log\rho$$

The first derivative of the log density is

$$\frac{d}{d\rho}\log f(\rho) = -\left(\rho_0 + \sum_{i,j}(1-\xi_{ij})d_j\right) + \sum_j \frac{\xi_{.j}d_j\exp\{-\rho d_j\}}{1-\exp\{-\rho d_j\}} + \frac{(\rho_1 - 1)}{\rho}$$

where $\xi_{.j} = \sum_i \xi_{ij}$

The log density is log concave since

$$\frac{d^2}{d\rho^2} f(\rho) = -\sum_j \frac{\xi_{.j} d_j^2 \exp\{\rho d_j\}}{(\exp\{\rho d_j\} - 1)^2} - \frac{(\rho_1 - 1)}{\rho^2}$$

which is negative for all $\rho \geq 0$.

*** Because this sampler conditions on ξ (which is sampled conditional on the realized locus ancestry states), it mixes slowly. Are we still doing this? Sampler should be replaced by a Metropolis step that uses the likelihood calculated by the HMM, conditioning only on the individual admixture parameters $\boldsymbol{\theta}$ and the observed genotype data.

7.3. Sampling for gamete-specific sum of intensities parameter ρ (implemented in class Individual)

To sample ρ for each gamete, we introduce another array of latent variables λ to represent the number of arrivals between each adjacent pair of linked loci. If the two loci are d morgans apart then $\lambda \sim \mathrm{Pn}(d\rho)$

The update of ρ is then a standard conjugate update of a Poisson intensity parameter with a gamma prior, conditional on the sum of observed arrivals between loci and the sum of the lengths of intervals between loci.

For L linked loci

$$p(\rho \mid \dots) \propto \pi(\rho) \prod_{j=2}^{L} \rho^{\lambda_j} \exp\{-d_j \rho\} \sim \mathrm{Ga}\left(\rho_0 + \sum_j \lambda_j, \rho_1 + \sum_j d_j\right)$$

We sample λ conditional on ξ; if $\xi = 1$ $\lambda \geq 1$ and zero otherwise. Given ξ we sample d', the distance from the last arrival in the interval between loci $j-1, j$ to the locus j. If the locus j has ancestry k then d' is distributed exponentially in the region $(0,\ d)$ with parameter $\theta_k \rho$ (since given $\xi = 1$ we know that there is at least one arrival in the region between $0,\ d$). Integrating constant K for distribution of d' is

$$K^{-1} = \int_0^d \rho \theta_k \exp\{-x \rho \theta_k\}\ dx = 1 - \exp\{-d \rho \theta_k\}$$

It follows that the cdf for d' is

$$F(d') = \frac{1 - \exp\{-d'\rho\}}{1 - \exp\{-d\rho\theta_k\}}$$

Thus we can sample d' from

$$d' = -\frac{1}{\rho}\left[1 - u(1 - \exp\{-\rho\theta_k d\}\right]$$

It follows that

$$\lambda - 1 \sim \mathrm{Pn}\left[\rho(d - d')\right]$$

7.4. Incorporating reported ancestry

*** is the code for this still in the model? if so, where?

We can model an individual's reported ancestry proportions as two Dirichlet distributions, one for each parent. Thus if the reported ancestry for individual i is expressed as $\theta \sim \mathrm{Di}_d(\boldsymbol{\pi}_i)$ the full conditional distribution of paternal/maternal admixture is

$$p(\boldsymbol{\theta}_i \mid \boldsymbol{\alpha}, \boldsymbol{\pi}_i) = \mathrm{Di}_d\left(\boldsymbol{\theta}_i \;\middle|\; \boldsymbol{\alpha} + \boldsymbol{\pi}_i + \sum_{j=1}^{m} \xi_{ij} \boldsymbol{A}_{ij} - \mathbf{1}\right).$$

8. Hidden Markov model algorithms (implemented in class HMM)

The transition matrices for probabilities of ancestry at each locus, conditional on the preceding locus, specify a Markov process on each gamete with stationary distribution θ. We combine the two haploid transition matrices of order K, one for each gamete, to a diploid transition matrix of order K^2, for which the state space is the ordered diploid ancestry states. For each pair of chromosomes in each individual, we can specify a hidden Markov model

For each compound locus at which genotypes are observed, we have a vector Λ that specifies for each possible ordered diploid ancestry state the probabilities of the observed (unphased) multilocus genotype conditional on the realized haplotype frequencies. To perform componentwise multiplication as a matrix operation, we convert the vector Λ to a diagonal matrix

$$diag\left(\Lambda_j\right) = \begin{pmatrix} p(g_j \mid A_j = 1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p(g_j \mid A_j = k) \end{pmatrix}$$

*** comment - is it computationally inefficient to use a diagonal matrix? Should we just code the componentwise multiplication directly?

We use standard HMM algorithms to calculate the likelihood of the observed genotype data at all loci on each chromosome in each individual, to sample the hidden states (locus ancestry), and to calculate the marginal conditional distribution of ancestry at each locus. The first step is to compute the forward and backward

probability vectors at each locus.

8.1. Forward-backward algorithm

For each locus we can calculate a vector $\boldsymbol{\alpha}$ and a vector $\boldsymbol{\beta}$ known as the forward probabilites and backward probabilities. The forward probabilities are the conditional probabilities $\alpha_1, \ldots, \alpha_m$: the probabilities of each possible hidden state (ancestry) at locus j given the observed data (genotypes) at loci $1, \ldots, j$. These probabilities are given by

$$\alpha^{(1)} = \boldsymbol{\theta}\mathrm{diag}\left(\Lambda_1\right), \quad \text{for } j = 2, \ldots, m$$
$$\alpha^{(j)} = \alpha^{(j-1)} P_{j-1}\left(\Lambda_j\right).$$

The backward probabilities are given by

$$\beta^{(m)} = 1, \quad \text{for } j = m-1, \ldots, 1$$
$$\beta^{(j)} = P_j\left(\Lambda_{j+1}\right)\beta^{(j+1)}$$

With K subpopulations, there are K^2 possible ordered diploid states of ancestry. Thus $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of length K^2.

8.2. Marginal distribution of locus ancestry and likelihood

The marginal distribution of ancestry at each locus is sampled as

$$\alpha^{(j)}_{1,k}\beta^{(j)}_{k,1} = p(A_j = k, g_1, \ldots, g_m) \propto p(A_j = k \mid g_1, \ldots, g_m)$$

The likelihood of the model parameters with the observed genotype data is calculated as

$$p(g_1, \ldots, g_m) = \sum_{k=1}^{K} p(A_j = k, g_1, \ldots, g_m) = \alpha^{(j)}\beta^{(j)}.$$

This expression for the likelihood can be calculated at any j.

*** comment - add a method to calculate the log-likelihood of each genotype at each compound locus, as a check on genotyping errors.

8.3. Sampling locus ancestry

Posterior samples of locus ancestry are required to update individual admixture, and to update the haplotype frequencies. Tests for linkage (association with locus

ancestry) can be calculated more efficiently from the marginal distribution of locus ancestry conditional on the model parameters, without sampling locus ancestry.

The locus ancestry $A_1, \ldots A_m$ are sampled in sequence, starting at the right-hand end of the chromosome and proceeding backwards. State A_m is sampled from $A_m \sim \mathrm{Mu}(\alpha^{(m)})$. Ancestry states $A_{m-1}, \ldots, A_1$ are then sampled conditional on the previously sampled states. Since the columns of P_j represent the distribution of ancestry at locus j given ancestry at locus $j+1$ we sample $A_j \sim \mathrm{Mu}(V)$ where V is the component wise product of $\alpha^{(j)}$ and the A_{j+1}th column of P_j.

9. Sampling regression parameters (implemented in class RegressionModel

To reduce posterior covariance between the regression parameters, the covariates X_2 should be standardized about their sample mean. Where covariates X_2 are not observed directly (as for individual admixture), their sample mean is estimated during the burn-in period.

9.1. Linear regression

The model is specified as

$$y \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \lambda)$$

where y is the response, $\boldsymbol{X}$ are the independent variables (covariates and individual admixture proportions), $\boldsymbol{\beta}$ is the vector of regression parameters and λ is the precision.

With the reference prior $\pi(\boldsymbol{\beta}, \lambda) = \lambda^{-1}$, the marginal posterior distribution of $(\boldsymbol{\beta})$ is

$$p(\boldsymbol{\beta}, \beta \mid \boldsymbol{z}) = \mathrm{St}\left(\boldsymbol{\beta} \middle| \boldsymbol{\beta}_n, \frac{1}{2}\boldsymbol{X}^t\boldsymbol{X}(n-k)\beta_n^{-1}, n-k\right) \approx \mathrm{N}\left(\beta \middle| \boldsymbol{\beta}_n, \frac{1}{2}\boldsymbol{X}^t\boldsymbol{X}(n-k-2)\beta_n^{-1}\right)$$

where

$$\boldsymbol{\beta}_n = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} \qquad \beta_n = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_n)^t\boldsymbol{y}$$

The marginal density of any subvector of a multivariate Student distribution $\mathrm{St}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha)$ is Student, with mean vector and inverse of the precision matrix given by the corresponding subvector of $\boldsymbol{\mu}$ and submatrix $\boldsymbol{\lambda}$.

*** comment - algorithm for logistic regression should be documented, preferably in terms of a generalized linear model.

10. Sampling haplotype pairs and haplotype frequencies (implemented in classes Composite Locus and AlleleFrequencies)

Under a model with no dispersion, it is straightforward to sample the ordered haplotype pairs once we have sampled the ordered states of locus ancestry on each gamete.

The simplest way to sample ordered haplotype pairs at each locus in each individual is to condition on the ancestry-specific haplotype frequencies, as well as the observed genotypes, and the ordered state of locus ancestry.

Alternatively, we can integrate out the ancestry-specific haplotype frequencies and update each individual's haplotypes conditional on the Dirichlet prior $\boldsymbol{\beta}$ and the realized haplotype counts in all other individuals.

The haplotype frequencies $\boldsymbol{\phi}$ can then be sampled conditional on the prior and the realized counts, as a conjugate Dirichlet distribution.

The full conditional for updating a pair x_i of haplotypes in the ith individual, conditional on the realized haplotype counts in all other individuals, is

$$P\left[x_i = (g,h) \mid X_{-i}, Y\right] \propto (n_g + \beta_g)\,(n_h + \beta_h)$$

where n_g and n_h are the realized counts of haplotypes g and h in all other individuals in the subpopulations of ancestry of paternal and maternal gametes at this locus, β_g and β_h are the corresponding elements of the Dirichlet parameter vectors, and X_i is the vector of realized haplotype pairs in all other individuals.

Niu T et al (AJHG, 2002) call this algorithm predictive updating.

*** problem with this algorithm - sampling of individuals are not conditionally independent given the population parameters. This won't be easy to parallelize - for parallel version, we maybe should use the simpler algorithm conditioning on ancestry-specific haplotype frequencies instead.

*** check the code - is this really what we're doing?

*** to speed up computation for large haplotypes, we may have to use partition-ligation (Niu 2002)

10.1. Sampling the Dirichlet parameters for ancestry-specific haplotype frequencies under a dispersion model

Under a dispersion model, the Dirichlet parameters for ancestry-specific haplotype frequencies are not specified as constants but have a stochastic dependence on the frequencies in a hypothetical ancestral population from which both modern unadmixed descendants and the admixed population under study are derived.

The Dirichlet parameter vector is reparameterized as a vector of proportions $\boldsymbol{\mu}$ and sum of Dirichlet parameters η.

The joint density for the ancestry-specific haplotype frequencies $\boldsymbol{\phi}$ at a compound locus with H haplotypes, conditional on $\boldsymbol{\mu}$, η is given by the Dirichlet density

$$\pi(\boldsymbol{\phi} \mid \eta, \boldsymbol{\mu}) = \frac{\Gamma(\eta)}{\Gamma(\eta - \sum \mu_i) \prod \Gamma(\mu_h)} \phi_n)^{\eta - \sum \mu_h - 1} \prod_{h=1}^{H-1} \phi_i^{\mu_h - 1}$$

This is also the likelihood function for the dispersion parameter η and the Dirichlet parameters $\boldsymbol{\mu}$, given realized haplotype frequencies $\boldsymbol{\phi}$.

The problem is to sample the Dirichlet proportion vector μ, conditional on the realized haplotype counts. If μ is univariate (in other words, if the locus is diallelic) we can sample μ directly from the product of two beta-binomial likelihoods

$$p(\mu \mid \eta, n_i, r^{(i)}) \propto \frac{1}{\Gamma((\eta - \mu)\Gamma(\mu))^2} \prod_{i=1}^{2} \Gamma\left(\eta - \mu + n_i - r^{(i)}\right) \Gamma\left(\mu + r^{(i)}\right)$$

where $i = 1$ for unadmixed modern descendants and $i = 2$ for the admixed population and n_i and $r^{(i)}$ are the realized haplotype counts in the sampled individuals. r_1 of n_1 gametes in

For multivariate $\boldsymbol{\mu}$ we have to use a Metropolis update. We propose $\boldsymbol{\mu}'$ from

$$\begin{aligned} \frac{\boldsymbol{\mu}'}{\eta} &\sim \text{Di}\left(\frac{\boldsymbol{\mu}}{\eta}\right) \\ q(\boldsymbol{\mu}') &= \prod_{h=1}^{n} \{\Gamma(\mu_h)\}^{-1} \left(\frac{\mu_h'}{\eta}\right)^{\mu_h} \\ p(\boldsymbol{\mu} \mid r) &= \frac{1}{\prod_{h=1}^{H} \Gamma(\eta\mu_h)^2} \prod_{i=1}^{2} \prod_{h=1}^{H} \Gamma\left(\eta\mu_h + r_h^{(i)}\right) \end{aligned}$$

*** comment - maybe a better proposal density would be to draw elements μ_h in sequence, at each step subtracting the realized count for haplotype h from the total haplotype count, and subtracting the drawn element μ_h from the sum η.

η is sampled by a random walk. To maximize step size we propose η', μ', where $\mu' = \frac{\eta'\mu}{\eta}$. Thus a new η' is only a change in dispersion.

$$p(\eta \mid \eta, n_i, r^{(i)}) \propto \frac{1}{\eta} \exp\left\{-\frac{\tau}{2}(\log\eta - \psi)^2\right\} \{\Gamma(\eta)\}^{2m} \prod_{j=1}^{m} \prod_{l=1}^{n} \frac{\left(\phi_{lj}^{(1)} \phi_{lj}^{(2)}\right)^{\mu_{lj}-1}}{\Gamma(\mu_{lj})}$$

*** comment - check this please.

We can then sample the haplotype frequencies ϕ conditional on the Dirichlet parameters $\boldsymbol{\mu}, \eta$ and the realized haplotype counts.

11. Construction of score tests based on the missing-data likelihood

We write $U(\theta; Y)$ for the observed-data score $\frac{d}{d\theta} log f(Y \mid \theta)$, and $I(\theta; Y)$ for the observed information $-\frac{d^2}{d\theta^2} log f(Y \mid \theta)$.

The complete-data log-likelihood can be partitioned into the observed data log-likelihood and the missing-data log-likelihood (Dempster, Laird and Rubin 1977).

$$\log f(Y, X \mid \theta) = \log f(Y \mid \theta) + \log f(X \mid Y, \theta) \tag{3}$$

Differentiating with respect to θ and taking expectations over the posterior distribution of the missing data X yields

$$E_{X|Y,\theta}\left[\frac{d}{d\theta} \log f(Y, X \mid \theta)\right] = U(\theta; Y)$$

as $E_{X|Y,\theta}\left[\frac{d}{d\theta} \log f(X \mid Y)\right]$, the expectation of a score over the probability of the data, is zero. We can thus evaluate the score $U(\theta; Y)$ as the posterior expectation of the complete-data score $\frac{d}{d\theta} \log f(Y, X \mid \theta)$

Differentiatng again, and taking expectations over the posterior distribution leads to the result

$$-E_{X|Y,\theta}\left[\frac{d^2}{d\theta^2} \log f(Y, X \mid \theta)\right] = I(\theta; Y) + \mathrm{Var}_{X|Y,\theta}\left[\frac{d}{d\theta} \log f(Y, X \mid Y, \theta)\right]$$

This result can be interpreted as
Complete information = observed information + missing information
The algorithm for the score test, applied to a Bayesian full probability model in which the parameter θ is fixed at its null value θ_0, with samples from the posterior distribution of the missing data X given the observed data Y generated by MCMC simulation, is as follows:-

1. At each realization of the complete data, compute the realized score vector and information matrix based on the complete-data likelihood at θ_0, and accumulate the results to evaluate:-

    - the score U as the posterior mean of the realized score.

- the complete information as the posterior mean of the realized information
- the missing information as the posterior variance of the realized score.

2. At the end of the run, calculate the observed information V as the complete information minus the missing information
3. Calculate the score test statistic as $UV^{-1/2}$ for scalar U, or $U'V^{-1}U$ where U is a vector.

11.1. Rao-Blackwellization

In tests for the effect of locus ancestry, the computational efficiency is improved by using a Rao-Blackwellized estimator of the score as described below. Suppose that we are sampling the posterior distribution of an unobserved variable X, which has a likelihood in the exponential family for the parameter of interest θ. For an exponential likelihood, the score for a single observation x_i has the form observed minus expected, weighted by a function a_i of a dispersion parameter ϕ.

$$\frac{x_i - \mu_i}{a_i(\phi)}$$

If we can compute the marginal distribution of x_i at each iteration, conditional on the model parameter μ and observed data Y, we can can evaluate the posterior expectation of the score as

$$U(\theta; Y) = E_{\mu|X,Y}\left[E_{X|Y,\mu}\left(\frac{x_i - \mu}{a_i(\phi)}\right)\right]$$

$$= E_{\mu|X,Y}\left[\frac{E_{X|Y,\mu}(x_i) - \mu}{a_i(\phi)}\right]$$

The missing information is then the sum of two components: the posterior variance of the conditional expectation of the complete-data score, and the posterior expectation of the conditional variance of the complete-data score.

$$\mathrm{Var}_{X|Y,\theta}\left[U(\theta; Y, X)\right] = \mathrm{Var}_{\mu|Y}\left(E_{X|\mu,Y}\left[U(\theta; Y, X)\right]\right) + \mathrm{E}_{Y|X}\left[\mathrm{Var}_{X|Y,\mu}\left(U(\theta; Y, X)\right)\right]$$

$$= \mathrm{Var}_{\mu|Y}\left(\frac{E_{X|Y,\mu}(x_i) - \mu}{a_i(\phi)}\right) + \mathrm{E}_{\mu|Y}\left[\frac{\mathrm{Var}_{X|Y,\mu}(x_i) - \mu}{a_i(\phi)}\right]$$

The first term measures the information that is missing because of uncertainty about the model parameter μ, and the second term measures the information that is missing because of uncertainty about the latent variable X.

11.2. Score covariance with other model parameters

We test each locus for association of a latent variable X_1 at the locus with the outcome Y under study, while adjusting for covariates X_2, including genetic background, in a generalized linear model of the form $g\left(E\left[Y\right]\right) = \left[X_1, X_2\right]'\left[\beta_1, \beta_2\right]$, where $g()$ is a link function. As X_1 is not observed directly, we sample from its posterior distribution given the observed genotype data. For a likelihood in the exponential family with canonical link function, the score vector for the ith observation at $\beta_1 = 0$ has the form

$$\mathbf{U}(\beta; Y, X) = \left[\begin{array}{c} \mathbf{U}(\beta_1; Y, X) \\ \mathbf{U}(\beta_2; Y, X) \end{array} \right] = \frac{p_i}{\phi} \left[\begin{array}{c} \mathbf{x_{i1}} \left(y_i - g^{-1}[\eta_i]\right) \\ \mathbf{x_{i2}} \left(y_i - g^{-1}[\eta_i]\right) \end{array} \right]$$

where where ϕ is the dispersion parameter, p_i is a known prior weight, and $\eta_i = \mathbf{x'_{i2}}\beta_2$.

The information matrix has the form

$$\mathbf{V}(\beta; Y, X) = \left[\begin{array}{cc} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{12} & \mathbf{V}_{22} \end{array} \right] = \frac{p_i}{\phi} \left[\begin{array}{cc} \mathbf{x'_{i1}x_{i1}} & \mathbf{x'_{i1}x_{i2}} \\ \mathbf{x'_{i2}x_{i1}} & \mathbf{x'_{i2}x_{i2}} \end{array} \right] \frac{d}{d\eta_i} g^{-1}[\eta_i]$$

If there is covariance between $U(\beta_1)$ and $U(\beta_2)$, the posterior variance of $U(\beta_1; Y, X)$ will include a component attributable to uncertainty in the model parameters β_2. This is undesirable for two reasons. First, it increases the computational workload, as longer runs are required to evaluate the difference between the complete and missing information. Second, it complicates the interpretation of the proportion of information extracted as a measure of the efficiency of the design, in relation to an ideal experiment in which the variables Y and X are observed directly.

To eliminate this covariance, we can standardize the complete-data score and information algebraically. Over the probability distribution of Y in hypothetical repetitions of the experiment, the asymptotic distribution of the score vector $U(\beta_1, Y, X)$ given $U(\beta_2; Y, X)$ is multivariate normal with mean $V_{12}V_{22}^{-1}U(\beta_2; Y, X)$ and variance $V_{11} - V_{12}V_{22}^{-1}V_{21}$

For each realization of the complete data, we calculate the covariance matrix V by summing over all observations, and use this to calculate the conditional expectation

$$E_{Y|X,\beta_1,\beta_2}\left[U(\beta_1; Y, X) \;\; | \; U(\beta_2; Y, X)\right]$$

and its conditional variance over the probability distribution of Y.

The score can now be calculated as

$$E_{X,\beta_2|Y,\beta_1}\left[U\left(\beta_1;Y,X\right)-E_{Y|X,\beta_1,\beta_2}\left(U(\beta_1;Y,X)\mid U\left[\beta_2;Y,X\right]\right)\right]$$

$$=E_{X,\beta_2|Y,\beta_1}\left[U\left(\beta_1;Y,X\right)\right]-E_{X,\beta_2|Y,\beta_1}\left[E_{Y|X,\beta_1,\beta_2}\left(U(\beta_1;Y,X)\mid U\left[\beta_2;Y,X\right]\right)\right]$$

The second term on the right is zero as it is the expectation of a score over the probability distribution of the data.

$$=E_{X,\beta_2|Y,\beta_1}\left[U(\beta_1;Y,X)\right]=U(\beta_1;Y)$$

The complete information is evaluated as the posterior expectation over X of $V_{11}-V_{12}V_{22}^{-1}V_{21}$, and the missing information is evaluated as the posterior variance of

$$U(\beta_1;Y,X)\ -E_{Y|X,\beta_1,\beta_2}\left[U(\beta_1;Y,X)\ \mid U(\beta_2;Y,X)\right]$$

This algorithm can be extended to exploit the Rao-Blackwellization described above, if we are able to evaluate the marginal conditional expectation and variance of the latent variable x_{i1}, given a model parameter μ and the observed data Y.

*** this Rao-Blackwellization is not yet implemented for regression score tests, but should be soon

To calculate the posterior expectation of the complete-data score U_1, we replace x_{i1} in the complete-data score and x_{i1}^2 in the complete-data information matrix by their posterior expectations. The missing information, standardized for U_2, is

$$\mathrm{Var}_{X|Y,\beta_1}\left[U_1-E_{Y|X,\beta_2}\left(U_1\ \mid U_2\right)\right]$$

which can be partitioned into two components

$$=\mathrm{Var}_{\mu,\beta_2|Y}\left(E_{X|\mu,Y}\left[U_1-E_{Y|X,\beta_2}\left(U_1\ \mid U_2\right)\right]\right)+\mathrm{E}_{\mu,\beta_2|Y}\left[\mathrm{Var}_{X|Y,\mu,\beta_2}\left(U_1-E_{Y|X,\beta_2}\left[U_1\ \mid U_2\right]\right)\right]$$

$$=\mathrm{Var}_{\mu,\beta_2|Y}\left(E_{X|\mu,Y}\left[U_1-E_{Y|X,\beta_2}\left(U_1\ \mid U_2\right)\right]\right)\ +\mathrm{E}_{Y|X,\beta_2}\left[\mathrm{Var}_{X|Y,\beta_2}\left(U_1\right)\right]$$

as the variance of $E_{Y|X,\beta_2}\left[U_1\ \mid U_2\right]$ over the distribution of X given Y,μ,β_2 is zero.

$$=\mathrm{Var}_{\mu,\beta_2|Y}\left(E_{X|\mu,Y}\left[U_1-E_{Y|X,\beta_2}\left(U_1\ \mid U_2\right)\right]\right)+\mathrm{E}_{\mu,\beta_2|Y}\left[\mathrm{Var}_{X|Y,\mu,\beta_2}\left(\mathbf{x_{i1}}\right)\left(y_i-g^{-1}[\eta_i]\right)\right]$$

Again, these two terms can be interpreted in terms of the sources of uncertainty that contribute to the missing information.

12. Score tests for linkage with locus ancestry

12.1. Affecteds-only score test for linkage

We define A_i as the number of gene copies at the locus under study that have ancestry from the high-risk population. θ_{i1} and θ_{i2} are the paternal and maternal admixture proportions.

The parameter under test is the ancestry risk ratio r. The likelihood function is a trinomial (sum of two bernoulli variates).

The score U with respect to $\log r$ at $\log r = 0$ is given by

$$U = \frac{1}{2}\sum_{i=1}^{n}(A_i - \theta_{i1} - \theta_{i2})$$

The information I is given by

$$I = \frac{1}{4}\sum_{i=1}^{n}\left[\theta_{i1}\left(1-\theta_{i1}\right) + \theta_{i2}\left(1-\theta_{i2}\right)\right)]$$

*** code in program may compute score and info for each gamete separately, then compute info for 1 individual by adding variances corrected for covariance. Simpler to calculate it as above.

Rao-Blackwellised estimate of $\mathrm{E}(U)$ can be obtained by replacing A_i by its conditional expectation, obtained from the HMM marginal conditional distribution. Conditional variance of U is obtained similarly from this marginal distribution.

12.2. Regression-based score tests (implemented in class ScoreTests)

These score tests test loci one at a time for association of a latent variable X_1 (ancestry or haplotype count) at the locus with the outcome Y under study, while adjusting for covariates X_2, including genetic background, in a generalized linear model of the form $g\left(E\left[Y\right]\right) = [X_1, X_2]'[\beta_1, \beta_2]$, where $g()$ is a link function.

The score and information are computed from the expressions given earlier for a generalized linear model with canonical link function.

12.3. Regression-based score test for linkage with binary outcome (implemented in class ScoreTests)

For a binary outcome, the link function is logistic, the dispersion parameter is 1, and the derivative of the inverse link function is

$f(1-f)$ where f is the expected value of Y given X, calculated by applying the the inverse link function to $\boldsymbol{X}'\beta$.

12.4. Score test for linkage with quantitative trait (linear regression model)

For a quantitative trait, the link function is the identity function, the dispersion parameter is the precision, and the derivative of the inverse link function is 1.

13. Score tests for association with alleles or haplotypes

13.1. Score tests for allelic association conditional on locus ancestry (not yet implemented)

For a locus with H haplotypes with ancestry specific haplotype frequencies $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{s-1})$, the likelihood of observing $\boldsymbol{r} = (r_1, \ldots, r_H)$ where r_h is the realized count of haplotype h. The complete-data likelihood is;

$$\begin{aligned}
L &= (1 - \sum \phi_k)^{r_s} \prod_{k=1}^{s-1} \phi_k^{r_k} \\
\log L &= r_s \log(1 - \sum \phi_k) + \sum_{k=1}^{s-1} r_k \log \phi_k \\
\frac{d}{d\phi_i} \log L &= \frac{r_i}{\phi_i} - \frac{r_n}{(1 - \sum \phi_k)} \\
\frac{d^2}{d\phi_i^2} \log L &= -\frac{r_i}{\phi_i^2} - \frac{r_n}{(1 - \sum \phi_k)^2} \\
\frac{d^2}{d\phi_i \phi_j} \log L &= -\frac{r_n}{(1 - \sum \phi_k)^2}
\end{aligned}$$

14. Computing the log-likelihood function (not yet implemented)

The p-value computed from the score test is useful when screening many null hypotheses to decide which ones to investigate further. For quantitative inference, we require the log-likelihood for the scalar parameter θ. For admixture mapping, inference is based on ancestry X at the locus under study, and the probability of the observed genotype data Y at all loci, given X and model parameters which include allele frequencies and parental admixture, does not depend upon the parameter under test (the ancestry risk ratio θ). Although $P(Y \mid X, \lambda, \theta)$ is an integral over states at other loci $X^\star$ of the form $\sum P(Y, X^\star \mid X, \lambda, \theta)$, the model specifies that

the outcome is dependent only upon ancestry at the locus under test and μ, so that $P(Y, X^\star \mid X, \lambda, \theta)$ does not depend on θ. It follows that

$$P(Y \mid X, \lambda, \theta) = P(Y \mid X, \lambda, \theta_0)$$

Similarly, in a genetic association study of the effect of unobserved haplotypes, the probability of the unphased genotype data given the pair of haplotypes does not depend upon the parameter under test (the effect of the haplotype upon the outcome. In this situation it is possible to compute the log-likelihood function directly as a function of θ, if we can evaluate the marginal conditional distribution $P(X \mid Y, \lambda, \theta_0)$.
Thompson and Guo (1991) show that the marginal log-likelihood ratio can be evaluated as the expectation of the complete-data log-likelihood ratio over the posterior distribution of the missing data under the null hypothesis.

$$L(\theta; Y) = \frac{P(Y \mid \theta)}{P(Y \mid \theta_0)} = E_{X|Y,\theta_0}\left[\frac{P(Y, X \mid \theta)}{P(Y, X \mid \theta_0)}\right]$$

We can rewrite this expectation in Rao-Blackwellized form as the expectation of a conditional expectation

$$L(\theta; Y) = E_{\lambda|Y,\theta_0}\left(E_{X|Y,\lambda,\theta_0}\left[\frac{P(X \mid \lambda, \theta)\, P(Y \mid X, \lambda, \theta)}{P(X \mid \lambda, \theta_0)\, P(Y \mid X, \lambda, \theta_0)}\right]\right)$$

If $P(Y \mid X, \lambda, \theta) = P(Y \mid X, \lambda, \theta_0)$

$$L(\theta; Y) = E_{\lambda|Y,\theta_0}\left(E_{X|Y,\lambda,\theta_0}\left[\frac{P(X \mid \lambda, \theta)}{P(X \mid \lambda, \theta_0)}\right]\right)$$

For discrete X with finite range of values, the inner expectation can be evaluated exactly, giving

$$L(\theta; Y) = E_{\lambda|Y,\theta_0}\left(\sum_X \frac{P(X \mid \lambda, \theta)}{P(X \mid \lambda, \theta_0)} P(X \mid Y, \lambda, \theta_0)\right)$$

To evaluate the log-likelihood function, we can compute the log of this likelihood ratio at multiple values of θ in a single run of the MCMC sampler.

Examples of this algorithm: Patterson 2004, Holmans 2002 (ascribed to Rice)

15. Model choice and model diagnostics

To choose between alternative models, we should compute the marginal likelihood of the model (evidence). This is currently implemented only for analyses of a single individual.

We also provide model diagnostics, based on score tests or on the posterior predictive check probability, to allow the user to identify specific ways in which the model fits the data poorly.

15.1. Computation of marginal likelihood, implemented in class Chib

Chib (1995) suggests calculating the marginal likelihood $p(y)$ as

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y} \mid \boldsymbol{\pi}^*) + \log p(\boldsymbol{\pi}^*) - \log p(\boldsymbol{\pi}^* \mid \boldsymbol{y})$$

This follows from Bayes theorem and holds for any $\boldsymbol{\pi}^*$. Typically, it is straightforward to evaluate $p(\boldsymbol{y} \mid \boldsymbol{\pi}^*)$ and $\log p(\boldsymbol{\pi}^*)$, thus we just need an estimate of the posterior ordinate $p(\boldsymbol{\pi}^* \mid \boldsymbol{y})$. This is most efficiently estimated if $\boldsymbol{\pi}^*$ is at or near its posterior mode.

If we specify a non-hierarchical model (independent priors on the admixture proportions of each individual, and a global sum of intensities parameter), the model parameters $\boldsymbol{\pi}$ are $\boldsymbol{\theta}, \boldsymbol{\phi}$ and ρ: respectively individual admixture, ancestry-specific allele frequencies and sum of intensities.

The posterior density of $\boldsymbol{\pi}$ can be written as

$$p(\boldsymbol{\pi} \mid \boldsymbol{y}) = \int p(\boldsymbol{\pi} \mid \boldsymbol{y}, \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{y}) d\boldsymbol{z}$$

Conditional on the latent variables $\boldsymbol{z}$ - ancestry states at each locus and the number of arrivals in each interval between loci - the full conditional density of $\boldsymbol{\pi}$ is the product of full conditional densities of $\boldsymbol{\theta}, \boldsymbol{\phi}$ and ρ, which are conjugate and can be evaluated directly at each realization of z.

We can thus evaluate the posterior ordinate $p(\boldsymbol{\pi}^* \mid \boldsymbol{y})$ with the following Monte Carlo estimate.

$$\hat{p}(\boldsymbol{\pi}^* \mid \boldsymbol{y}) = \frac{1}{M} \sum_{i=1}^{M} p(\boldsymbol{\pi}^* \mid \boldsymbol{y}, \boldsymbol{z}^{(i)}).$$

The posterior modes of $\boldsymbol{\theta}, \boldsymbol{\phi}$ and ρ are estimated during the burn-in period.

In the current version of the program, this method is implemented only for data on a single individual.

*** for given allele frequencies, a more efficient algorithm would be to calculate the likelihood at $\boldsymbol{\pi}^*$ using the HMM, without conditioning on locus ancestry states. Then use Metropolis proposal steps to estimate the posterior ordinate. With a prior on allele frequencies, this would be more complicated.

*** can this be extended to a hierarchical model for admixture and sum intensities? We would have to group the parameters into blocks (one block for each individual, one block for each population-level parameter).

**** should test other algorithms for calculation of evidence: annealing, or nested sampling

15.2. Score test for mis-specification of ancestry-specific allele frequencies at a diallelic locus

With k populations, the probabilities of observing 0, 1 or 2 copies of allele 1, conditional on parental admixture, are given by;

$$\begin{aligned}
P(X=0) &= \Pi^{(0)} = \sum_i^k \sum_j^k q_i q_j \theta_i^{(P)} \theta_j^{(M)} \\
P(X=1) &= \Pi^{(1)} = \sum_i^k \sum_j^k q_i p_j \theta_i^{(P)} \theta_j^{(M)} + \sum_i^k \sum_j^k p_i q_j \theta_i^{(P)} \theta_j^{(M)} \\
P(X=2) &= \Pi^{(2)} = \sum_i^k \sum_j^k p_i p_j \theta_i^{(P)} \theta_j^{(M)}.
\end{aligned}$$

where p_i is the frequency of allele 1 in population i and $q_i = 1 - p_i$.

The score is obtained by differentiating the logarithms of these expressions with respect to $p_i,\ i = 1, \ldots, k$.

$$\begin{aligned}
\boldsymbol{U}^{(0)} &= \left(U_1^{(0)}, \ldots, U_k^{(0)} \right) \\
U_i^{(0)} &= \frac{1}{\Pi^{(0)}} \left(-2 q_i \phi_{ii} - \sum_{j \neq i}^k q_j (\phi_{ij} + \phi_{ji}) \right) \\
\boldsymbol{U}^{(1)} &= \left(U_1^{(1)}, \ldots, U_k^{(1)} \right) \\
U_i^{(1)} &= \frac{1}{\Pi^{(1)}} \left(2(q_i - p_i)\phi_{ii} + \sum_{j \neq i}^k (q_j - p_j)(\phi_{ij} + \phi_{ji}) \right) \\
\boldsymbol{U}^{(2)} &= \left(U_1^{(2)}, \ldots, U_k^{(2)} \right) \\
U_i^{(2)} &= \frac{1}{\Pi^{(2)}} \left(2 p_i \phi_{ii} + \sum_{j \neq i}^k p_j (\phi_{ij} + \phi_{ji}) \right)
\end{aligned}$$

Where $\phi_{ij} = \theta_i^{(P)} \theta_j^{(M)}$.

$$
\begin{aligned}
I_{ii}^{(0)} &= \left(U_i^{(0)}\right)^2 - \frac{2\phi_{ii}}{\Pi^{(0)}} \\
I_{ij}^{(0)} &= U_i^{(0)} U_j^{(0)} - \frac{(\phi_{ij} + \phi_{ji})}{\Pi^{(0)}} \\
I_{ii}^{(1)} &= \left(U_i^{(1)}\right)^2 + \frac{4\phi_{ii}}{\Pi^{(0)}} \\
I_{ij}^{(1)} &= U_i^{(1)} U_j^{(1)} + \frac{2(\phi_{ij} + \phi_{ji})}{\Pi^{(1)}} \\
I_{ii}^{(2)} &= \left(U_i^{(2)}\right)^2 - \frac{2\phi_{ii}}{\Pi^{(2)}} \\
I_{ij}^{(2)} &= U_i^{(2)} U_j^{(2)} - \frac{(\phi_{ij} + \phi_{ji})}{\Pi^{(2)}}
\end{aligned}
$$

With this test conditional on gamete admixture, we are inferring mis-specified allele frequencies by comparing the observed and expected allele counts given each individual's admixture proportions. The scores for misspecification of allele frequencies in different subpopulations at the same locus are correlated with each other.

By conditioning on locus ancestry we can derive tests for mis-specified allele frequencies that are independent across subpopulations.

*** is this now implemented in the code?

15.3. Posterior predictive check test for residual population stratification

*** the current version of the program uses for this test all loci that are at least x cM apart (what is x?). This is not a pure test for stratification

*** the code for this test should be fixed to use only unlinked loci

* we might have a more efficient test if we use a weighted sum of allele scores at all loci on each chromosome, obtained as the first principal component.

This test exploits the argument that if all population stratification has been accounted for, there should be no residual allelic association between unlinked loci. We construct test statistics and calculate them for the observed and replicate data sets.

We choose a single simple locus from each chromosome, dichotomize the alleles into two bins at any locus where there are more than two alleles, and compute for each individual the observed minus expected count of allele 1, where the expected count is a weighted average of the ancestry-specific allele frequencies with weights given by the admixture proportions of the two gametes. We then compute the covariance matrix between these observed minus expected counts. In other words,

these are partial covariances.

The matrix of covariances between counts of allele 1 at these unlinked loci is calculated as matrix $A = \{a_{ij}\}$ where

$$a_{ij} = \sum_k [X_{ik} - \mathrm{E}(X_{ik} \mid \theta_k)][X_{jk} - \mathrm{E}(X_{jk} \mid \theta_k)]$$

At each realization of the sampled alleles and allele frequencies, we compute the first eigenvalue of this matrix, divided by the trace (sum of eigenvalues). If this proportion is higher than expected by chance, it indicates that there is an underlying factor giving rise to allelic associations between unlinked loci. This ratio T is computed as T_{obs} for the observed alleles and as T_{rep} for a replicate dataset with alleles sampled conditional on the ancestry-specific frequencies and gamete admixture proportions. The posterior predictive check probability is the posterior frequency with which $T_{rep} > T_{obs}$.

15.4. Posterior predictive check test for dispersion between prior and observed allele frequencies

n_{jk} - number of individuals with ancestry k at marker j

r_{jk} - vector of counts of observed alleles at marker j with ancestry k

ϕ_{jk} - allele frequency at marker j in population k

Generate

$$r'_{jk} \sim \mathrm{Mu}(\phi_{jk}, n_{jk})$$

Compare T_{obs} (likelihood of r_{jk}) and T_{rep} (likelihood of r'_{jk}).

This test statistic is computed for each locus in each subpopulation, and as a summary test over all loci in each subpopulation.