The use of composite likelihood methods in population genetics

Paul Fearnhead

Department of Mathematics and Statistics,
Lancaster University

Outline

- Introduction to Population Genetics.
- Composite Likelihood methods for estimating recombination rates:
  - Two composite likelihood methods;
  - Overview of existing theory;
  - Open questions.

Introduction: Population Genetics

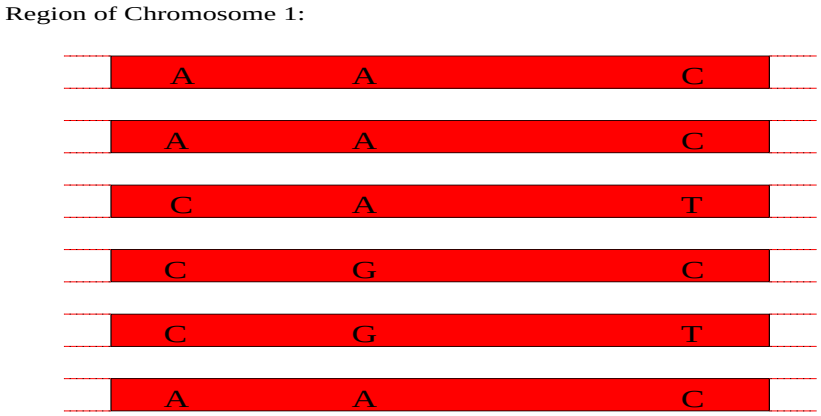
Population Genetics is the study/modelling of the genetic make-up of populations.

Data consists of some genetic information for each member of a randomly chosen sample from a population of interest.

Inference can be about both the Biological and Demographic forces that have affected the evolution of the population.

For concreteness, we will focus on the case of SNP haplotype data.

SNP (haplotype) data



SNP refers to ascertained positions of the genome where there is known to be polymorphism with the population. Haplotype refers to known the genetic type of each SNP on each of the two copies of a chromosome (for diploid organisms).

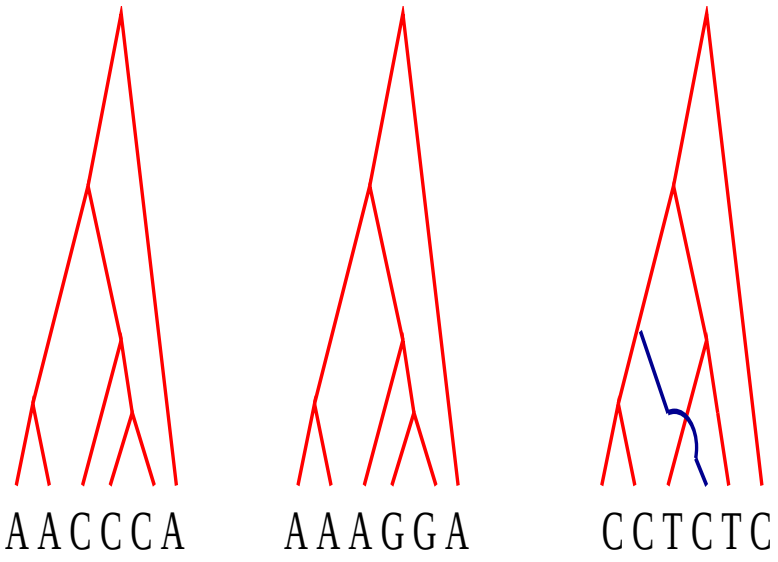
Coalescent Models

Statistical models for such data are often based on a stochastic process called the Coalescent.

The Coalescent is a probabilistic model for the genealogical relationship of a random sample of chromosomes from a population; and the dependence of the genealogies at different SNPS.

It is derived based upon models for the evolution of the population; and can account for biological factors (Mutation, Recombination, Selection), and demographic factors (Non-random mating, changes in population size).

Example: Genealogies



Recombination

The dependence between genealogies (and hence data) at different SNPs is governed by recombination: the greater the rate of recombination between two SNPs, the more independent the SNPs are.

Recombination is the process by which a parent's two copies of a chromosome is shuffled together when passed onto a child.

Recombination rates between SNPs are not proportional to their physical distance.

The Use of Composite Likelihood Methods

Likelihoods under coalescent models can only be calculated (computationally) for small data sets. By comparison, modern data sets can have data from 100s–1000s of individuals; each typed at 100,000s-1,000,000s of SNPs

This has lead to the use of Composite Likelihood Methods. Example applications include:

- Estimating Recombination Rates (Hudson 2001, Fearnhead and Donnelly 2002, Fearnhead 2003) These methods have produced the first fine-scale recombination map for humans (McVean et al 2004).
- Estimating Demographic Models (See Wiuf 2006 for theoretical results).
- Inferring genes under Selection (E.g. Nielsen et al. 2005).

Focus on Composite Likelihood for Estimating Recombination

We will consider two composite likelihood methods for estimating recombination rates:

(i) subdividing a chromosomal region into small sub-regions;

(ii) considering data from pairs of SNPs.

In each case we will consider inference for a scalar recombination rate, ρ. (E.g. under the assumption of a constant recombination rate across the region of interest.)

Asymptotic Regime and Dependence

We consider the asymptotic regime where the size of the chromosomal region we have data from goes to infinity.

[In most cases there is a finite amount of information about parameters in the limit as the number of individuals tends to infinity.]

Main Result. If we have data from two sub-regions, $\mathcal{D}_1$ and $\mathcal{D}_2$; these sub-regions are disjoint, and separated by a recombination rate ρ_b. Then

$$|\pi(\mathcal{D}_1, \mathcal{D}_2) - \pi(\mathcal{D}_1)\pi(\mathcal{D}_2)| = O(1/\rho_b).$$

Method of Fearnhead and Donnelly (2002)

This method splits the data into identically distributed sub-regions. Let $\mathcal{D}_i$ be the data for the ith sub-region; and $l_i(\rho) = \log \pi(\mathcal{D}_i|\rho)$ the corresponding log-likelihood function for the (per kb) recombination rate ρ..

Fearnhead and Donnelly (2002) suggest basing inference on

$$\ell_{\mathbf{CL}}(\rho) = \sum_{i=1}^{R} l_i(\rho).$$

In the limit as $R \to \infty$ (and under mild regularity conditions); $\hat{\rho} = \arg\max \ell_{\mathbf{CL}}(\rho)$ is a consistent estimator.

Outline of Proof

The proof of consistency is based on three results:

(i) $\mathbb{E}(\log \pi(\mathcal{D}_i|\rho)$, viewed as a function of ρ has a maximum at the true parameter value.

(ii) For any ρ,

$$\frac{1}{R}\ell_{\mathbf{CL}}(\rho) = \frac{1}{R}\sum_{i=1}^{R} l_i(\rho) \to \mathbb{E}(\log \pi(\mathcal{D}_i|\rho)$$

as $R \to \infty$.

(iii) Some technical results that deal with the state-space being continuous.

Comments

- The asymptotic regime is unrealistic (as recombination rates are not constant across a chromosome).
- Interest lies in extending the result to get asymptotic normality of the estimator; and hence a justifiable method for obtaining confidence intervals (which works well for realistic sample sizes).
- In practice, estimating the variance of the estimator is problematic – assumptions about the dependence of the score function for different sub-regions is required.

Pairwise Method of Hudson (2001)

Now let $\ell_{i,j}(\rho)$ be the log-likelihood function based on data solely on data from ith and jth SNPs.

A pairwise likelihood is

$$\ell_{\mathbf{PL}}(\rho) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} w_{i,j} \ell_{i,j}(\rho),$$

for some fixed weights $w_{i,j}$. Hudson (2001) and McVean et al. (2002) propose using a pairwise likelihood with $w_{i,j} = 1$.

[This is the more popular method of the two: due to its computational properties.]

Results

Define $\tilde{\rho} = \arg\max \ell_{\mathbf{PL}}(\rho)$

Consider the limit as the chromosomal region increases (and $N \to \infty$):

(i) If $w_{i,j} = 1$ then it has not been proven that the $\tilde{\rho}$ is consistent. (Theory suggests it is not, and this is supported by empirical evidence; see Smith and Fearnhead (2005).)

(ii) If $w_{i,j}$ decay sufficiently quickly with the distance between the ith and jth SNPs, then $\tilde{\rho}$ is consistent.

Note (ii) has been used the implementation of pairwise methods (McVean et al. 2004 and Li et al. 2006).

Comments

- Again, the asymptotic regime is unrealistic (as recombination rates are not constant across a chromosome).
- Interest lies in how best to choose $w_{i,j}$.
- Also, how to estimate the variance of $\tilde{\rho}$ (or calculate confidence intervals).

Discussion

- The only existing theoretical results relate to consistency of composite likelihood estimators.
- Even for these, the asymptotic regime is in some cases no realistic (though it is for estimating demographic parameters).
- Interest in theoretical results which:
  - guide the implementation of composite likelihood methods;
  - help with the production of confidence interval or evaluating uncertainty in the estimator.

References

Theory:

Fearnhead (2003) Consistency of estimators of the population-scaled recombination rate. *Theoretical Population Biology* 64, 67–79.
Wiuf (2006) Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology*, 53, 821–841.

Applications:

Fearnhead and Donnelly (2002) Approximate likelihood methods for estimating local recombination rates. *JRSS B*, 64, 657–680.
Hudson (2001) Two-Locus Sampling Distributions and Their Application. *Genetics, 159, 1805–1817.*
Li et al. (2006) A New Method for Detecting Human Recombination Hotspots and Its Applications to the HapMap ENCODE Data. *American Journal of Human Genetics*, 79 628 – 639.
McVean et al. (2002) A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. *Genetics* 160, 1231–1241.
McVean et al. (2004) The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science*, 304, 581 – 584.
Nielsen et al. (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, 15, 1566-1575.
Smith and Fearnhead (2005) A Comparison of Three Estimators of the Population-Scaled Recombination Rate: Accuracy and Robustness. *Genetics*, 171, 2051–2062.