A Review and Comparison of Estimators of the Population-scaled Recombination Rate

Paul Fearnhead

Department of Mathematics and Statistics,
Lancaster University

(Joint work with Nick Smith.)

Outline

Look at three estimators of the recombination rate: Composite Likelihood, Pairwise Likelihood, and PACL:

(i) Theoretical properties;

(ii) Comparison of performance;

(iii) Robustness properties (to deviations from the Null model).

Concentrate on estimating a constant recombination rate (per kb) across a region of length, L.

Composite Likelihood (CL)

Idea:

(i) Split a long-region into sub-regions;

(ii) Calculate $l_i(\rho)$ an (accurate) approximate log-likelihood for sub-region i;

(iii) Set $\mathrm{Cl}(\rho) = \Sigma_i \, l_i(\rho)$.

Choose sub-regions with approx. 8 segregating sites (excluding singletons).

This is the slowest and least flexible of the three methods.

Composite Likelihood (theory)

CL gives a consistent estimator (as the length, L, of data increases).

Information increases like $L/\log L$.

Unless only small numbers of sub-regions are used, the related Likelihood Ratio statistic gives anti-conservative confidence intervals.

Pairwise Likelihood (Pwl)

Focus on Hudson (2001)'s approach.

Idea:

(i) For segregating sites i and j, define $l_{ij}(\rho)$ to be the log probability of the data at sites i and j conditional on segregation;

(ii) Set Pwl$(\rho) = \Sigma_{i,j}\, l_{ij}(\rho)$.

Assumes no repeat mutations.

This is the quickest and most flexible of the three methods.

Pairwise Likelihood: Theory

Pwl is a special case of a class of approximate log-likelihoods of the form

$$l(\rho) = \sum_{i,j} w_{ij} l_{ij}(\rho).$$

If weigths w_{ij} decay sufficiently quickly as the distance between sites i and j increase: consistent estimator as L increases.

Simulation suggests that a scaled Likelihood Ratio

$$2(\text{Pwl}(\hat{\rho}) - \text{Pwl}(\rho))/S,$$

where S is the number of segregating sites, is approximately χ^2_1.

PACL

Approach of Li and Stephens (2003) based on an approximate (and tractable) approximation of

$$\mathsf{P}(h_i|h_1,\ldots,h_{i-1};\rho).$$

For an ordered data-set $h_1,\ldots,h_n$, the PACL is obtained by the Product of these Approximate Conditional Likelihoods.

The PACL curve depends on the order of the sample: Average over a sample of orderings. (Asymptotic properties as sample size increases are incorrect for a specific ordering, but correct if averaged over all orderings.)

Corresponding Likelihood Ratio statistic is approximately χ^2_1 (though discrepancy in the tail).

PACL: Theory

Consider data at two sites a distance $d$kb apart, with alleles a/A and b/B respectively.

The Approximate Probability is

$$\mathsf{P}(h_i = ab | h_1, \ldots, h_{i-1}; \rho) \approx f_a f_b + p_d(\rho)(f_{ab} - f_a f_b),$$

where:

$p_d(\rho) = \exp(-d\rho)$ is the probability of no recombination between the two sites;

and f_{ab} is the frequency of haplotype ab in the sample $h_1, \ldots, h_{i-1}$ (etc.)

Comparison (1)

For small data-sets CL appears best (and PACL appears to substantially under-estimate ρ).

	ρ per kb	CL		PwL		PACL		Average		CL+PwL	
		RMSE	g	RMSE	g	RMSE	g	RMSE	g	RMSE	g
10kb	1/4	1.04	0.43	0.64	0.63	0.68	0.56	0.64	0.73	-	-
10kb	1	0.57	0.80	0.58	0.86	0.47	0.90	0.50	0.89	0.51	0.86
10kb	4	0.32	0.95	0.37	0.93	0.23	0.98	0.25	1.00	-	-
25kb	1	0.33	0.71	0.36	0.73	0.31	0.85	0.28	0.82	0.29	0.82

$\theta = 1$ per kb. Estimates of ρ/ρ_0. RMSE = Root Mean Square Error (scaled by true ρ); g is the proportion of estimates within a factor of 2 (3/2 for 25kb).

Comparison: PWL - inclusion of pairs

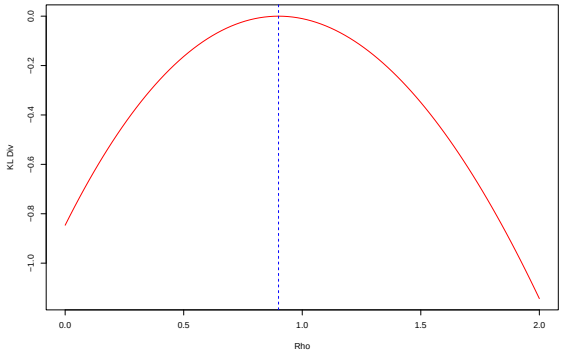
	1kb	5kb	10kb	50kb	All
MAE	0.21	0.17	0.16	0.18	0.18

Significant reduction of Mean Absolute Error (MAE) for only using pairs of sites within 10kb.

$\rho = \theta = 1$ per kb for 100kb. Estimates of ρ per kb.

Robustness: Method

Asymptotic properties of the mle is given by the expected log-likelihood curve.



This curve can be estimated by calculating the average of the log-likelihood curve over a large number of data sets. We report the value of ρ or ρ/θ which maximises this curve.

(θ estimated from segregating sites.)

Robustness: 10kb with $\rho/\theta = 1$

Scenario	CL	PwL	PACL
Growth	1.10	1.16	1.30
Bottleneck	0.65	0.60	0.56
4-Island (even)	0.80	0.84	0.88
4-Island (uneven)	0.65	0.62	0.62
Rate Variation	0.60	0.66	0.68

Results are the value of ρ (or ρ/θ) which maximises average log-likelihood.

Truth is 1 in each case.

Robustness: Scaling

	CL		PwL		PACL	
Scenario	0.25 to 1	1 to 4	0.25 to 1	1 to 4	0.25 to 1	1 to 4
Growth	0.79	1.09	0.82	1.36	0.91	1.18
Bottleneck	0.85	0.84	1.25	1.27	0.97	1.32
2-Island (even)	1.00	1.06	1.02	0.81	0.81	0.95

Columns are for increasing rate of ρ relative to θ. 10kb data.

Figures show relative increase/decrease in estimate of ρ compared to a four-fold increase in ρ. (1 is ideal.)

Robustness: Island Model (Pwl)

ρ per kb	10kb	50kb	100kb
1/4	0.75	0.75	0.75
1	0.80	0.70	0.65
4	0.70	0.50	0.45

Values show estimate relative to true ρ value.

Results for a 2-Island Model; even sampling.

Robustness: Mutation Rate (Pwl)

θ	Constant	Variable
0.001	1.00	1.00
0.01	1.05	1.05
0.1	3.45	6.60

Finite sites mutation model; Variable assumes a Gamma(1/2) distribution of mutation rate at each site.

Truth is 1.

Robustness: Sweep

ρ per site	CL	Pwl	PACL
0.25	1.00	2.80	0.42
1.00	0.65	0.94	0.36
4.00	0.60	0.75	0.62

Estimates relative to true ρ; 10kb data.

Absolute estimate of ρ when there is no recombination is larger for Pwl than for $\rho = 1$ case.

Appears to be an artefact of conditioning on segregation.

Summary

PACL appears to be the most accurate method for larger data.

For most deviations to the underlying model, the effect is similar on all three methods.

All three methods appear robust at estimating relative recombination rates.

Pwl can be improved by downweighting pairs of distant sites.