momi: a new method for computing the multipopulation sample frequency spectrum

John Kamm\(^1\), Jonathan Terhorst\(^1\) and Yun S. Song\(^{1,2,3}\)

Departments of \(^1\)Statistics, \(^2\)EECS, and \(^3\)Integrative Biology, University of California, Berkeley

Sample Frequency Spectrum (SFS)

- Distribution of counts of mutant alleles observed at a site
- Used to summarize genetic data and infer biological parameters
- momi (MO\(_{\text{Ran}}\) Models for Inference) is a program to compute the SFS for a neutral site under:
  - population size changes (including exponential growth)
  - population splits and mergers
  - pulse migration and admixture events

Our approach

- View demography as graphical model, apply variable elimination
  - aka “tree-peeling” when the demography is a tree
- Represent allele frequencies with Moran model
  - Equivalent to using the coalescent
- Polanski-Kimmel equations
  - Quickly and stably compute mutations arising in each subpopulation
- Automatic differentiation to compute gradient and Hessian

Demographic history as graphical model

(a) A demographic history with 18 parameters, very loosely based on human history. All parameters are in coalescent-scaled units.

(b) The same history, represented as a graphical model. The SFS is then computed via variable elimination. Each vertex represents the allele frequency of a particular subpopulation at a particular point in time.

Moran model

Moran model is a finite population model where lineages copy alleles onto each other at some rate \(\lambda\).

We model alleles within each vertex \(v\) by a Moran model with \(n_v\) lineages.

- \(n_v\) = number of samples with some ancestry in \(v\)
- Copying rate \(\lambda = \frac{1}{N(t)}\) inverse population size

Kingman’s coalescent embedded within Moran model via sample genealogy.

- \(\Rightarrow\) Moran is equivalent to using coalescent

Comparison with other population genetic models

- Moran model: \(O(n_v)\) states per vertex
  - # derived alleles at time \(t\)
- Coalescent: \(O(n_v^2)\) states per vertex
  - # ancestors and # derived alleles at time \(t\)
- Diffusion: \(O(D)\) states per vertex
  - Continuous state space: fraction of population with derived allele
  - “Discretize” into \(D\) states
  - Typically \(D \gg n_v\)

Inference

Use automatic differentiation to compute gradient and Hessian information

- Infer parameters via hill-climbing algorithm

Example demography with 18 parameters:

- Simulated 10 datasets with ms
  - \(n = 10\) samples per deme in Africa, East Asia, Melanesia, Europe.
  - \(n = 2\) samples per deme in Neanderthal, Denisova.
- For each dataset:
  - Choose random initial parameters (shown in blue)
  - Find local optimum (shown in red) with single run of a conjugate gradient method
- On average, each dataset had 185815.1 SNPs and 1436.6 observed SFS entries.
- Average running time of parameter search on a single dataset (start to finish) was 23.1 hours.

This research is supported by the NIH, the Packard Fellowship for Science and Engineering, a Miller Research Professorship, and a Citadel Fellowship.

\{jkamm,terhorst,yss\}@stat.berkeley.edu