BAYESIAN INFERENCE IN GENETICS

Jaromir Sant, Paul Jenkins, Jere Koskela, Dario Spanò

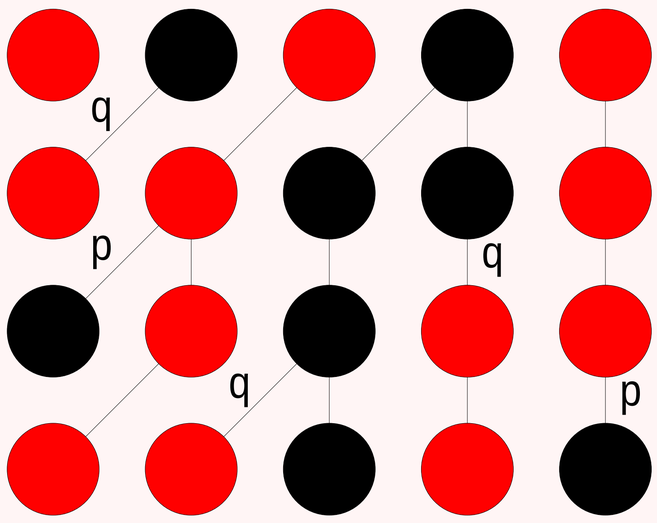
The Wright-Fisher Model

Consider a haploid population of fixed size N and consider only two alleles a and A. In the neutral Wright-Fisher model without mutation, parents are chosen uniformly at random and offspring inherit their type. To add mutation and selection:

Mutation - after choosing parent, flip a coin. If heads then the offspring mutates, otherwise it retains type of their parent

Selection - weight individuals by the relative fitness when choosing parents

Example of WF Model with Mutation and Selection



If we let Y_k^N be the number if individuals having allele A, and we assume that a and A have relative fitness $1 : 1 + \frac{s}{N}$, and mutation probabilities $\frac{1}{N}\ _{a\to A}$; $\frac{1}{N}\ _{A\to a}$, then

$$\mathrm{P}\left[Y_{k+1}^N = j \,|\, Y_k^N = i\right] = \binom{N}{j}\ _i^j (1 - \ _i)^{N-j}$$

with

$$_i := \frac{i(1+\frac{s}{N})(1-\frac{1}{N}\ _{A\to a}) + (N-i)\frac{1}{N}\ _{a\to A}}{i(1+\frac{s}{N}) + N - i}$$

The Wright-Fisher Diffusion

Rescaling the Wright-Fisher model leads to the Wright-Fisher diffusion:

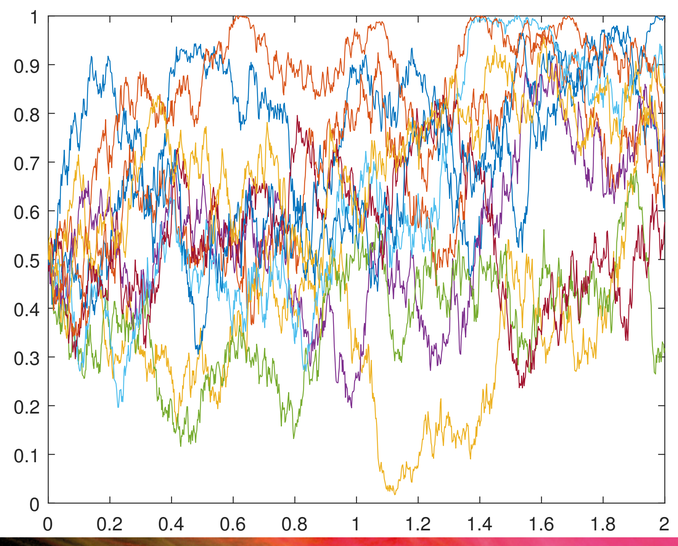
$$\frac{1}{N} Y_{\lfloor tN \rfloor}^N \to X_t$$

where the convergence is pathwise in $D_{[0;1]}([0;\infty))$, and X satisfies the SDE

$$dX_t = \frac{1}{2}\left(sX_t(1-X_t) - \ _{a\to A}X_t + \ _{A\to a}(1-X_t)\right)dt + \sqrt{X_t(1-X_t)}dW_t$$

$s \in \mathcal{S}$ is the selection parameter we wish to infer and $_{a\to A};\ _{A\to a} > 0$ are the corresponding mutation parameters which we assume to be known.

Sample paths from a Wright-Fisher Diffusion with $s = 1,\ _{a\to A} = 0.5;\ _{A\to a} = 0.8$



The Inferential Setting

We observe one whole sample path $(X_t)_{t\in[0;T]}$ continuously through time, and consider the properties of the Bayesian estimator $\tilde{s}_T$ for s in the asymptotic limit $T \to \infty$. We define

$$\tilde{s}_T = \arg\min_{\bar{s}_T} \int_{\mathcal{S}} \mathrm{E}_{\mathrm{P}_s^{(T)}}\left[`\left(\sqrt{T}\,(\bar{s}_T - s)\right)\right] p(s)ds$$

where p and l are a suitably chosen prior and loss function respectively. The idea is to obtain bounds on what we can learn from the data in the absence of observational error, as done in [1] & [2]. The object of interest is the **likelihood ratio function**

$$Z_{T;s}(u) := \frac{d\mathrm{P}_{s+\frac{u}{\sqrt{T}}}^{(T)}}{d\mathrm{P}_s^{(T)}}(X^T)$$

where we look at an order $\frac{1}{\sqrt{T}}$ perturbation around a fixed s.

Properties of the Bayesian Estimator

The Ibragimov-Has'minskii Conditions & Theorem

C1 : $\forall K \subset \Theta$ compact, $\exists\ a; B \in \mathrm{R}$ s.t. $\forall R > 0$, $\forall u_1; u_2$ s.t. $|u_1| < R$, $|u_2| < R$, and $q > 0$

$$\sup_{s\in K} \mathrm{E}_{\mathrm{P}_s^{(T)}}\left[\left|Z_{T;s}(u_2)^{\frac{1}{2}} - Z_{T;s}(u_1)^{\frac{1}{2}}\right|^2\right] \leq B(1+R^a)|u_2-u_1|^q$$

C2 : $\forall K \subset \Theta$ compact, $\exists g_T(\cdot)$ a suitable monotonically increasing continuous function s.t. $\forall u \in \mathrm{U}_{T;s} := \{u : s + \frac{u}{\sqrt{T}} \in \Theta\}$

$$\sup_{s\in K} \mathrm{E}_{\mathrm{P}_s^{(T)}}\left[Z_{T;s}(u)^{\frac{1}{2}}\right] \leq e^{-g_T(|u|)}$$

C3 : The random functions $Z_{T;s}(u)$ have marginal distributions which converge uniformly in $s \in K$ as $T \to \infty$ to those of the random function $Z_s(u)$

C4 : The random function

$$(v) = \int_{\mathrm{R}} `(v-u)\frac{Z_s(u)}{\int_{\mathrm{R}} Z_s(y)dy}du$$

attains its minimum value at the unique point $\tilde{u}(s) = \tilde{u}$ with probability 1

Theorem : If $\tilde{s}_T$ is the Bayesian estimator and C1-C4 hold, then we have that $\tilde{s}_T$:

is uniformly consistent in $s \in K$
is uniformly asymptotically normal
displays moment convergence for any $p > 0$ uniformly on compacts $K \subset \Theta$
is asymptotically efficient for a suitable choice of loss function

References

[1] Y. A. Kutoyants, *Statistical inference for ergodic diffusion processes.* Springer Series in Statistics. London, 2004

[2] G. A. Watterson. Estimating and testing selection: the two-alleles, genetic selection diffusion model. *Adv. in Appl. Probab.*, 11(1):14-30, 1979.

MASDOC Mathematics and Statistics Centre for Doctoral Training

WARWICK THE UNIVERSITY OF WARWICK

EPSRC Pioneering research and skills