A mathematical interlude: The infinitesimal model as the limit of Mendelian inheritance

Luminy, 18th June 2015

Additive traits in haploids (no mutation)

N_t = population size in generation t; M = number of (unlinked) loci.

- Trait value in individual j:

$$Z_j = \bar{z}_0 + \sum_{l=1}^{M} \frac{1}{\sqrt{M}} \eta_{jl},$$

where $\bar{z}_0$ = average value in ancestral population. Scaled allelic effects satisfy $|\eta_{jl}| \leq B$ (uniformly in j, l).

Additive traits in haploids (no mutation)

N_t = population size in generation t; M = number of (unlinked) loci.

- Trait value in individual j:

$$Z_j = \bar{z}_0 + \sum_{l=1}^{M} \frac{1}{\sqrt{M}} \eta_{jl},$$

where $\bar{z}_0$ = average value in ancestral population. Scaled allelic effects satisfy $|\eta_{jl}| \leq B$ (uniformly in j, l).

- Ancestral population. $\hat{\eta}_{jl}$ i.i.d (for different j). $\mathbb{E}[\hat{\eta}_l] = 0$ for all l.

Reproduction

$[1]$ and $[2]$ refer to the first and second parents of an individual.

- $\eta_{jl}[1]$ is the scaled allelic effect at locus l in the first parent of the jth individual. Similarly, $Z_j[1]$ will denote the trait value of the first parent of individual j.

- Write $X_{jl} = 1$ if the allelic type at locus l in the jth individual is inherited from the 'first parent' of that individual; otherwise it is zero. $\mathbb{P}[X_{jl} = 1] = 1/2 = \mathbb{P}[X_{jl} = 0]$.

$$Z_j = \bar{z}_0 + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \{X_{jl}\eta_{jl}[1] + (1 - X_{jl})\eta_{jl}[2]\}\,.$$

Conditioning

Let

- $\mathcal{P}^{(t)}$ denote the *pedigree* relationships between all individuals up to and including generation t;
- $Z^{(t)}$ denote the *traits* of all individuals in the pedigree up to and including the tth generation.

We would like to derive the distribution of trait values in generation t conditional on knowing $\mathcal{P}^{(t)}$ and $Z^{(t-1)}$.

Warmup: the ancestral population

For $\beta = (\beta_1, \ldots, \beta_{N_0}) \in \mathbb{R}^{N_0}$,

$$\begin{aligned}\sum_{j=1}^{N_0} \beta_j Z_j &= \bar{z}_0 \sum_{j=1}^{N_0} \beta_j + \frac{1}{\sqrt{M}} \sum_{j=1}^{N_0} \sum_{l=1}^{M} \beta_j \hat{\eta}_{jl} \\ &= \bar{z}_0 \sum_{j=1}^{N_0} \beta_j + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \left(\sum_{j=1}^{N_0} \beta_j \hat{\eta}_{jl} \right).\end{aligned}$$

CLT implies convergence of distribution function to that of Normal, mean $= \bar{z}_0 \sum_{j=1}^{N_0} \beta_j$, variance $= \hat{\Sigma} \sum_{j=1}^{N_0} \beta_j^2$. Error order $1/\sqrt{M}$.

($\frac{1}{M} \sum_{l=1}^{M} \mathrm{Var}(\hat{\eta}_l) \to \hat{\Sigma}$ as $M \to \infty$.)

Continuing the argument

Ancestral population:
$(Z_1, \ldots, Z_{N_0}) \to$ multivariate normal with mean vector $(\bar{z}_0, \ldots, \bar{z}_0)$ and covariance matrix $\hat{\Sigma}\,\mathrm{Id}$

What about later generations?

$$Z_j = \bar{z}_0 + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \{X_{jl}\eta_{jl}[1] + (1 - X_{jl})\eta_{jl}[2]\}\,.$$

Key: Need to be able to calculate the distribution of $\eta_{jl}[1]$ *conditional on* $Z^{(t-1)}$ (and show that it is almost unaffected by the conditioning).

A toy example

Suppose η_l are i.i.d. with $\eta_l = \pm 1$ with equal probability, $\bar{z}_0 = 0$.

$$
\begin{aligned}
\mathbb{P}[\eta_1 = 1 | Z = k/\sqrt{M}] &= \frac{\mathbb{P}\left[\sum_{l=1}^{M} \eta_l = k \middle| \eta_1 = 1\right]}{\mathbb{P}\left[\sum_{l=1}^{M} \eta_l = k\right]} \mathbb{P}\left[\eta_1 = 1\right] \\
&= \frac{\mathbb{P}\left[\sum_{l=2}^{M} \eta_l = (k-1)\right]}{\mathbb{P}\left[\sum_{l=1}^{M} \eta_l = k\right]} \mathbb{P}\left[\eta_1 = 1\right] \\
&= \frac{\frac{1}{2^{M-1}}}{\frac{1}{2^{M}}} \frac{\binom{M-1}{(M+k-2)/2}}{\binom{M}{(M+k)/2}} \mathbb{P}\left[\eta_1 = 1\right] \\
&= \left(1 + \frac{k}{M}\right) \mathbb{P}\left[\eta_1 = 1\right].
\end{aligned}
$$

Toy example continued

If scaled allelic effects are i.i.d. Bernoulli,

$$\mathbb{P}\left[\eta_1 = 1 \middle| Z = \frac{k}{\sqrt{M}}\right] = \left(1 + \frac{k}{M}\right)\mathbb{P}\left[\eta_1 = 1\right].$$

For a 'typical' trait value, $k/M = \mathcal{O}(1/\sqrt{M})$.

For extreme values ($k = \pm M$), the trait gives complete information about the allelic effect at each locus.

For 'typical' k, the distribution of η_1 is almost unchanged because there are so many different configurations of allelic effects that correspond to the same trait value.

Strategy for a recursion

Want to show that at any generation, conditional on $\mathcal{P}^{(t)}$ and $Z^{(t-1)}$, as $M \to \infty$, $(Z_j)_{j=1,\ldots,N_t}$ converges to a multivariate normal random variable for which the variance-covariance matrix is conditionally independent of $Z^{(t-1)}$ given $\mathcal{P}^{(t)}$.

Key step: show that the variance-covariance matrix is independent of $Z^{(t-1)}$, for which we need conditional distribution of $(\eta_{jl}[1], \eta_{jl}[2])$.

Mathematical obstruction: CLT gives convergence of *distribution function* of each trait value, but for conditional probabilities would need convergence of the corresponding *density* functions. Not in general available.

Conditioning on approximate trait values

Write $Z^{(t)} \approx \underline{z}$ for $|Z_j^{(t)} - z_j| \leq \epsilon_M$, $\forall j$.

Conditional on $\mathcal{P}^{(t)}$ and $Z^{(t-1)} \approx \underline{z}$,

$$\left(Z_j - \frac{Z_j[1] + Z_j[2]}{2} \right)_{j=1,\ldots,N_t}$$

converges (in distribution) to mean zero multivariate normal with diagonal covariance matrix Σ_t.

$(\Sigma_t)_{jj}$ = *segregation variance* among offspring of the parents of individual j.

Generation one

Bayes' rule again

$$\mathbb{P}\left[\eta_{jl}[1] = x | \, \mathcal{P}^{(1)}, Z^{(0)} \approx \underline{z}\right] =$$

$$\mathbb{P}[\eta_{jl}[1] = x] \frac{\mathbb{P}[Z_j[1] - \frac{1}{\sqrt{M}}\eta_{jl}[1] \approx z_1 - \frac{x}{\sqrt{M}}]}{\mathbb{P}[Z_j[1] \approx z_1]},$$

We have

$$\left| \mathbb{P}\left[\frac{(Z_j[1] - \bar{z}_0)}{(\hat{\Sigma}^M)^{1/2}} \leq z \right] - \Phi(z) \right| \leq \frac{C}{\sqrt{M\hat{\Sigma}^M}} \left(1 + \frac{1}{\hat{\Sigma}^M}\right).$$

$$\mathbb{P}[Z_j[1] \approx z_1] = \left(\phi(y) + \frac{C}{2\epsilon_M\sqrt{M}}\left(1 + \frac{1}{\hat{\Sigma}^M}\right)\right)\frac{2\epsilon_M}{\sqrt{\hat{\Sigma}^M}},$$

for some y with

$$y \in \left(\frac{z_1 - \overline{z}_0}{\sqrt{\hat{\Sigma}^M}} - \frac{\epsilon_M}{\sqrt{\hat{\Sigma}^M}}, \frac{z_1 - \overline{z}_0}{\sqrt{\hat{\Sigma}^M}} + \frac{\epsilon_M}{\sqrt{\hat{\Sigma}^M}}\right).$$

$$\mathbb{P}\left[Z_j[1] - \frac{1}{\sqrt{M}}\eta_{jl}[1] \approx z_1 - \frac{1}{\sqrt{M}}x\right]$$
$$= \left(\phi(y^{(l)}) + \frac{C'}{2\epsilon_M\sqrt{M}}\left(1 + \frac{1}{\hat{\Sigma}^M}\right)\right)\frac{2\epsilon_M}{\sqrt{\hat{\Sigma}^M}},$$

for some $y^{(l)}$ with

$$y^{(l)} \in \left(\frac{z_1 - \overline{z}_0 - x/\sqrt{M}}{\sqrt{\hat{\Sigma}^M}} - \frac{\epsilon_M}{\sqrt{\hat{\Sigma}^M}}, \frac{z_1 - \overline{z}_0 - x/\sqrt{M}}{\sqrt{\hat{\Sigma}^M}} + \frac{\epsilon_M}{\sqrt{\hat{\Sigma}^M}}\right).$$

Taylor expansion of ϕ about y and using that

$$|y^{(l)} - y| \leq \frac{|x|}{\sqrt{M}} + \frac{\epsilon_M}{\sqrt{\hat{\Sigma}^M}}$$

$$\rightsquigarrow \text{errors order } \epsilon_M + \frac{1}{\epsilon_M \sqrt{M}}.$$

Choose $\epsilon_M = 1/M^{1/4}$

$$\frac{\mathbb{P}[Z_j[1] - \frac{1}{\sqrt{M}}\eta_{jl}[1] \approx z_1 - \frac{1}{\sqrt{M}}]}{\mathbb{P}[Z_j[1] \approx z_1]} = 1 + \frac{C}{M^{1/4}}\left(1 + \frac{1 + \Delta(|z_1 - \bar{z}_0|)}{\hat{\Sigma}^M}\right).$$

As with toy model, require that the trait we condition upon is not 'too extreme'.

Generation t

Aim: conditional on $\mathcal{P}^{(t)}$ and $Z^{(t-1)} \approx \underline{z}$,

$$\left(Z_j - \frac{Z_j[1] + Z_j[2]}{2} \right)_{j \in \{1,\ldots,N_t\}}$$

converges in distribution to mean zero, normally distributed random variable with diagonal variance-covariance matrix, Σ_t, which is *conditionally independent* of $Z^{(t-1)}$ given $\mathcal{P}^{(t)}$.

$$Z_j = \bar{z}_0 + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \{X_{jl}\eta_{jl}[1] + (1 - X_{jl})\eta_{jl}[2]\}$$

$$= \frac{Z_j[1] + Z_j[2]}{2} + R_j.$$

The limiting variance

$$R_j = \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \left(X_{jl} - \frac{1}{2} \right) \eta_{jl}[1] + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \left((1 - X_{jl}) - \frac{1}{2} \right) \eta_{jl}[2].$$

$\mathbb{E}[R_j | \mathcal{P}^{(t)}, Z^{(t-1)}] = 0$. Limiting variance:

$$\lim_{M \to \infty} \frac{1}{M} \sum_{l=1}^{M} \left(\frac{\mathbb{E}[\hat{\eta}_l^2]}{2} - \frac{\mathbb{E}[\eta_{jl}[1]\eta_{jl}[2] | \mathcal{P}^{(t)}]}{2} \right).$$

This is just $\hat{\Sigma}(1 - F_j)/2$ where F_j is the probability of identity of the two parents of the jth individual, i.e. we recover the segregation variance among offspring of the parents of the jth individual.

What have we proved?

Convergence of distribution of traits conditional on $\mathcal{P}^{(t)}$ and $Z^{(t-1)}$ to multivariate normal whose variance-covariance matrix is conditionally independent of $Z^{(t-1)}$ given $P^{(t)}$, with error order $1/M^{1/4}$ provided that

- Traits in pedigree not too extreme: i.e. $|Z_j - \frac{Z_j[1]+Z_j[2]}{2}|$ not too big;
- Probability of identity not too close to one.

So under these conditions, infinitesimal model valid for $\mathcal{O}(M^{1/4})$ generations.

What have we proved?

Convergence of distribution of traits conditional on $\mathcal{P}^{(t)}$ and $Z^{(t-1)}$ to multivariate normal whose variance-covariance matrix is conditionally independent of $Z^{(t-1)}$ given $P^{(t)}$, with error order $1/M^{1/4}$ provided that

- Traits in pedigree not too extreme: i.e. $|Z_j - \frac{Z_j[1]+Z_j[2]}{2}|$ not too big;
- Probability of identity not too close to one.

So under these conditions, infinitesimal model valid for $\mathcal{O}(M^{1/4})$ generations.

BUT $M^{1/4}$ is not very big.

Toy example tells us that sometimes can do better ($M^{1/2}$).

Local Limit Theorem (David McDonald 1979)

$\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$ be independent, integer-valued.

$$p_n(x) = \mathbb{P}[\xi_{n1} + \cdots + \xi_{nn} = x],$$

$$\mu_{nm} = \mathbb{E}[\xi_{nm}], \quad B_n^2 = \sum_{m=1}^{n} \mathtt{Var}(\xi_{nm}), \quad A_n = \sum_{m=1}^{n} \mu_{nm}.$$

(I) $\limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}[\exp(a\xi_{nm})] < \infty$ for some $a > 0$;

(II) $\liminf_{n\to\infty} B_n^2/n \geq c > 0$;

(III) $\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \left(\sum_{k=-\infty}^{\infty} \min\{\mathbb{P}[\xi_{nm} = k], \mathbb{P}[\xi_{nm} = k+1]\} \right) > 0.$

Then $p_n(x) = \frac{1}{\sqrt{2\pi} B_n} \exp\left(-\frac{(x - A_n)^2}{2B_n^2} \right) \left(1 + \frac{C}{\sqrt{n}} \right)$

uniformly in $|x - A_n| \leq \sqrt{n}$.

Why does this work?

Convergence was fast for our toy model because when the trait value is a sum of independent Bernoulli random variables, many different combinations of allelic effects lead to the same trait value.

Condition (III) guarantees that we can 'find' $\mathcal{O}(n)$ independent Bernoulli random variables lurking inside the sequence $\xi_{n1}, \ldots, \xi_{nn}$.

The Bernoulli part of a random variable

X random variable with mass function f.

$\alpha(k) := f(k) \wedge f(k+1)$, $q = \sum_{k=-\infty}^{\infty} \alpha(k)$.

Define T, U, ϵ, L by

$$\begin{aligned} f_T(k) &= \frac{\alpha(k)}{q}, \quad f_U(k) = \frac{1}{1-q}\left(f(k) - \frac{\alpha(k-1)+\alpha(k)}{2}\right), \\ f_\epsilon(0) &= 1-q, \quad f_\epsilon(1) = q, \quad f_L(0) = \frac{1}{2}, \quad f_L(1) = \frac{1}{2}. \end{aligned}$$

$$X \sim (1-\epsilon)U + \epsilon T + \epsilon L.$$

Applying this to our sum of independent variables

For variables $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$,

$$S_n := \Sigma_{m=1}^{n} \xi_m = \sum_{m=1}^{n} \left[V_m + \epsilon_m L_m \right].$$

Set $M_n = \sum_{m=1}^{n} \epsilon_m$,

$$S_n \stackrel{d}{=} Z_n + \sum_{m=1}^{M_n} L_m^*,$$

where L_m^* i.i.d. Bernoulli. $\quad M_n = \mathcal{O}(n).$

$$\text{(III)} \quad \liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \left(\underbrace{\sum_{k=-\infty}^{\infty} \min\{\mathbb{P}[\xi_{nm} = k], \mathbb{P}[\xi_{nm} = k+1]\}}_{q_m} \right) > 0.$$

Beyond the additive

House of cards mutation

Mutation probability per locus per generation μ. Scaled allelic effect mutant at locus l, $\tilde{\eta}_l$.

Environmental noise

Trait value of offspring

$$Z_j = \bar{z}_0 + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \{X_{jl}\eta_{jl}[1] + (1 - X_{jl})\eta_{jl}[2]\} + E_j,$$

E_j Gaussian 'environmental noise'.

Beyond the additive

House of cards mutation

Mutation probability per locus per generation μ. Scaled allelic effect mutant at locus l, $\tilde{\eta}_l$.

Environmental noise

Trait value of offspring

$$Z_j = \bar{z}_0 + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \{X_{jl}\eta_{jl}[1] + (1 - X_{jl})\eta_{jl}[2]\} + E_j,$$

E_j Gaussian 'environmental noise'.

. . . and then life much easier.

Epistasis

For a set $U \subseteq \{1, 2, \ldots, M\}$ of loci, write χ_U for the allelic *states* and $f_U(\chi_U)$ for the corresponding scaled epistatic effects.

$$Z = \bar{z}_0 + \sum_U a_U f_U(\chi_U).$$

Expected offspring trait no longer simply the mean of the parental values.

Epistasis

For a set $U \subseteq \{1, 2, \ldots, M\}$ of loci, write χ_U for the allelic *states* and $f_U(\chi_U)$ for the corresponding scaled epistatic effects.

$$Z = \bar{z}_0 + \sum_U a_U f_U(\chi_U).$$

Expected offspring trait no longer simply the mean of the parental values.

However, provided that $f_U = 0$ for $|U| > D$ and

$$\sum_{U \cap U' \neq \emptyset} a_U a_{U'} < \infty,$$

offspring traits still follow a normal distribution with variance conditionally independent of $Z^{(t-1)}$ given $\mathcal{P}^{(t)}$.