Information, Variances and Covariances in Evolutionary Genetics

Warren J Ewens

Marseille, May 2009

The theme

Population genetics theory, with Fisher, Haldane and Wright, was mainly a theoretical exercise, aimed at validating and describing the Darwinian theory using the Mendelian hereditary mechanism. In this it was successful. But… there was not much data.

By contrast, we now have data, that is "information". In this talk I discuss the information concept in the context of standard "Genetics 101" material.

(Fisher) information

For a discrete random variable whose possible values have respective probabilities $p_1(\theta)$, $p_2(\theta)$, …, where θ is an (unknown) parameter, the Fisher information about θ provided by an observed value of the random variable is

$$E\ (\mathrm{d}/\mathrm{d}\theta \log p(\theta))^2$$

$$= \Sigma_j\, p_j(\theta)\ (\mathrm{d}/\mathrm{d}\theta \log p_j(\theta))^2$$

$$= \sum_j\ (\ \dot{p}_j(\theta))^2 / p_j(\theta)$$

The "dot" derivative is with respect to θ.

As an example, in the Poisson (θ) case,

$$d/d\theta \, (\log p_j(\theta) = d/d\theta \, \{-\theta + j \log \theta\} = -1 + j/\theta,$$

And so the information about θ provided by the observed value of a random variable having the Poisson distribution is

$$\sum_j [e^{-\theta} \theta^j/j!] \, [-1 + j/\theta]^2 = E(j - \theta)^2/\theta^2 = 1/\theta.$$

The information about θ provided by n iid observations is n/θ.

Simple haploid two-allele one-locus model

Standard deterministic evolutionary model (1920's):-

We have two alleles, A_1 and A_2, with fitnesses $(1+sa_1)/(1+sa)$ and $(1+sa_2)/(1+sa)$, where the parental generation frequencies are p_1 and p_2 and $a = p_1a_1+p_2a_2$. We can then compute the daughter generation frequencies p_1' and p_2' as

$$p_i' = p_i\,(1+sa_i)/(1+sa), \quad (i = 1,2).$$

The new approach: knowing a_1, a_2, p_1 and p_2 , we can compute s.

Note that

$$\Delta p_j = p_j' - p_j = s p_i (a_i - a)/(1+sa), \quad (j = 1,2).$$

Thus

$$\frac{(\Delta p_1)^2}{p_1} + \frac{(\Delta p_2)^2}{p_2} = \frac{s^2[p_1(a_1 - a)^2 + p_2(a_2 - a)^2]}{(1+sa)^2}$$

Also, σ^2 = variance in fitness

$$= \frac{s^2[p_1(a_1 - a)^2 + p_2(a_2 - a)^2]}{(1+sa)^2}$$

So " deterministic model information" = variance in fitness.

The variance in fitness can also be written as $\dfrac{p_1 p_2 s^2 (a_1 - a_2)^2}{(1+sa)^2}$

In the analogous continuous-time model, knowing p and s,

$$\dot{p} = sp(1 - p).$$

(dot derivative = rate of change with respect to time.)

The new perspective: knowing p and p', we can say that

$$s = \dot{p} / [p(1 - p)]$$

Roughly, $s\,\delta t = \delta p/[p(1\text{-}p)]$.

In a continuous-time evolutionary model, time and fitness differences are totally confounded. Thus making an inference about a time parameter is equivalent to making an inference about fitness differentials, and vice versa.

In the simplest continuous-time model,

$$\dot{p}_1 = sp_1(1 - p_1), \dot{p}_2 = -sp_2(1 - p_2).$$

A deterministic "Fisher-information-like" quantity derived from this is "Sum over alleles of the square of the rate of change of allele frequency divided by the current frequency", or

$$\sum_j (\dot{p}_j)^2 / p_j = s^2 p_1 p_2 .$$

The variance in fitness is

$$s^2 p_1 (1-p_1)^2 + s^2 p_2 (1-p_2)^2 = s^2 p_1 p_2 .$$

Thus, once again, "variance = information".

The stochastic (Wright-Fisher) case

Suppose that in a haploid population of N genes, there are i A_1 genes in some parental generation. Assume a Wright-Fisher model with the fitnesses above. The probability that these i A_1 genes give rise to j genes in the daughter generation is

$$p_{ij} = \binom{N}{j} \psi_i^j (1 - \psi_i)^{n-j}$$

where $\psi_i = (i/N)(1+sa_1)/(1+sa)$.

How much (Fisher) information about s do the observed values of i and j provide?

Using Fisher information, we get

$$\log p_{ij} = \text{const} + j \log(1+sa_1) + (N-j) \log(1+sa_2) - N \log(1+sa)$$

$$\mathrm{d} \log p_{ij} / \mathrm{d}s = \{j - \mathrm{E}(j)\} \{a_1/(1+sa_1) - a_2/(1+sa_2)\}$$

Therefore Fisher information $= \mathrm{E}[\, \mathrm{d} \log p_{ij} / \mathrm{d}s \,]^2$

$$= \frac{s^2 (a_1 - a_2)^2}{(1+sa_1)^2 (1+sa_2)^2} Var(j)$$

$$= \frac{i(N-i) s^2 (a_1 - a_2)^2}{N(1+sa_1)(1+sa_2)(1+sa)^2}$$

Note the close similarity between the "stochastic" model information and the deterministic model information (= the deterministic total variance in fitness).

Do these various results generalize to the *k* allele haploid model?

The case of *k* alleles

Consider k alleles $A_1, A_2, \ldots, A_k$ at the locus, with respective fitnesses $(1 + sa_i)/(1+sa)$, $i = 1, 2, \ldots, k$, where $\sum_j p_j a_i = a$. Then in the deterministic case,

$$\Delta p_j = p_j(1+sa_j)/(1+sa) - p_j$$

$$= p_j s (a_j - a)/(1+sa) .$$

Thus $$\sum_j (\Delta p_j)^2/p_j = s^2 \sum_j p_j (a_j - a)^2/(1+sa)^2$$

$$= \frac{s^2}{(1+sa)^2}[\{\sum_{j-1}^{k} p_j a_j^2\} - a^2].$$

This is also the total variance in fitness. So again, "information" = variance.

The stochastic case.

Suppose that the frequencies of the k alleles A_1, A_2, .., A_k in a population of N genes are $p_1, p_2, \ldots, p_k$, and that one generation later there are $n_1, n_2, .., n_k$ genes respectively of these types.

The probability of the values $n_1, n_2, \ldots., n_k$ is

$$\frac{N!}{n_1!\ldots n_k!}[\prod_j \{j(1+sa_j)^{n_j}\}]/(1+sa)^N.$$

Here, as before, $a = \sum p_j a_j$.

The log of this is const + $\sum n_j \log(1+sa_j) - N \log(1+sa)$.

Then the Fisher information about s is

$$\frac{N}{1+sa}\left[\sum_{j=1}^{k}\frac{p_j a_j^2}{1+sa_j}-\frac{a^2}{1+sa}\right]$$

This is "close to" to the deterministic $\sum_j (\Delta p_j)^2/p_j$,

And so also close to the total variance in fitness.

The diploid case (with k alleles)

Deterministic theory. The Fisher information

$$\Sigma_j \, (\dot{p}_j)^2 / p_j$$

(or something like it) arises in k-allele deterministic diploid evolutionary population genetics theory in at least two places – see later.

But first: average effects and the additive genetic variance

If P_{ij} is the ordered frequency of the genotype A_iA_j, and w_{ij} is the fitness of that genotype, and w is the mean fitness, we find the average effects of the alleles $A_1, A_2, \ldots, A_k$ by minimizing

$$\sum\sum P_{ij} (w_{ij} - w - \alpha_i - \alpha_j)^2$$

with respect to $\alpha_1, \alpha_2, \ldots, \alpha_k$. The additive genetic variance σ_A^2 is the sum of squares so removed. It is that component of the total variance in fitness of the various genotypes explained by the genes in those genotypes.

The minimizing values of α_1, α_2, …, α_k are the average effects of A_1, A_2, … A_k.

They can be thought of as the best we can do in assigning fitnesses to the various alleles (analogous to the fitnesses $1 + sa_1$, …, $1 + sa_k$ in the haploid case).

What information do we have about these average effects?

Fisher information result #1.

In discrete time, with fitnesses depending on the alleles at one locus and assuming random mating, and allowing an arbitrary number k of alleles $A_1, A_2, \ldots, A_k$ possible at the gene locus of interest, with respective frequencies $p_1, p_2, \ldots, p_k$, and with parental generation mean fitness of 1,

$$\sum_j (\Delta p_j)^2 / p_j = \tfrac{1}{2} \sigma_A^2.$$

Here σ_A^2 is the additive genetic variance in fitness and Δp_j is the change in the frequency of allele A_j between parent and daughter generation brought about by natural selection.

Note the changes from the (haploid) total variance in fitness to the (diploid) half the additive component of this variance.

We can therefore think of (1/2) σ_A^2 as the "information" about the average effects of the *k* alleles available from allelic frequencies in successive generations.

(This involves a leap of faith – not proven yet.)

This is useful since we can estimate σ_A^2 from data.

Equivalently …..

The additive genetic variance and parent/offspring correlations between relatives and covariances in fitness.

In very simple models, with *k* alleles at the locus of interest,

correlation (parent/offspring) = $\frac{1}{2}\,\sigma_A^2 / \sigma^2$

or equivalently

covariance (parent/offspring) = $\frac{1}{2}\,\sigma_A^2$.

So in these simple models we can think of the P/O covariance as providing information about the average effects.

Does equating P/O covariance with "information" make sense?

Yes - A covariance between (say) height and weight is a measure of how much information about weight there is in a height measurement. In this case the P/O covariance is a measure of how much information about average effects is obtained by comparing parental and daughter allelic frequencies.

Special case: at a stable (internal) equilibrium, allelic frequencies do not change, and σ_A^2 , the P/O covariance and all average effects are zero. Thus the (zero) P/O covariance gives the information that the average effects are all zero.

Fisher information result #2.

We have just seen that the natural selection allelic frequency changes (Δp_1, Δp_2, …., Δp_k) satisfy the equation

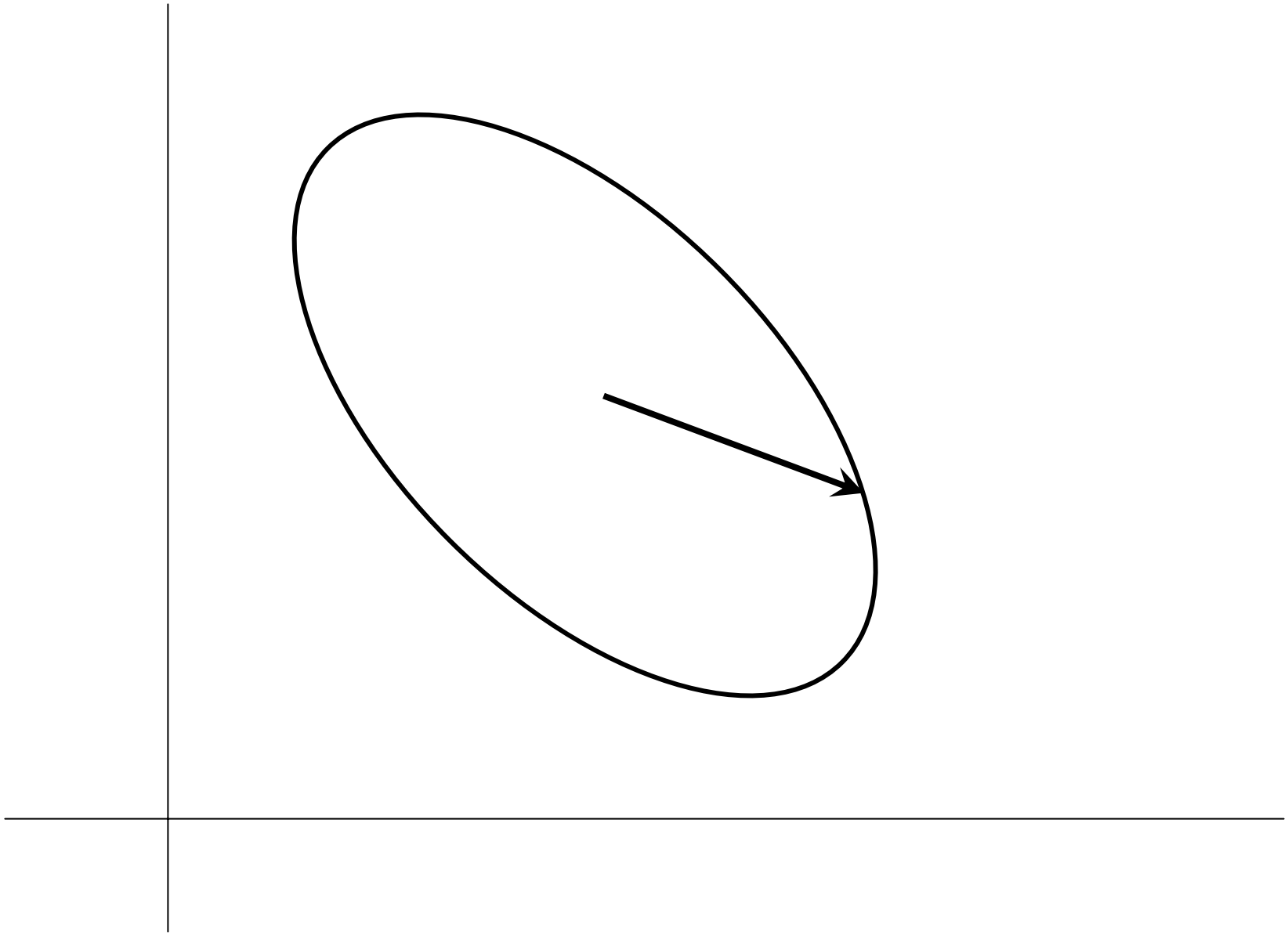
$$\sum_j (\Delta p_j)^2 / p_j = (½)\, \sigma_A^2$$

The Kimura optimization principle, for the case of random mating, with fitnesses depending on one locus only, states that:-

Let ($d\,p_1$, $d\,p_2$, …., $d\,p_k$) be an *arbitrary* vector of changes of allelic frequencies, subject (of course) to the requirement that $\sum_j d\,p_j = 0$. Then this principle states that, subject to the constraint

$$\Sigma_j (\,d\,p_j\,)^2/p_j = (½)\,\sigma_A^2,$$

inspired by the equation satisfied by the natural selection changes, the allelic frequency changes which maximize the between-generation increase in mean fitness are the natural selection values ($\Delta\,p_1$, $\Delta\,p_2$, …., $\Delta\,p_k$).



That is, thinking of (½) σ_A^2 as information, then for a given amount of information about the average effects of the alleles that drive allelic frequency changes, the natural selection changes maximize the increase of mean population fitness.

It is convenient to move to matrix and vector notation to carry these ideas to the whole genome. So we define M as a diagonal matrix whose typical element is p_i, $\boldsymbol{\Delta}$ as the vector $(\Delta p_1, \ \Delta p_2, \ldots, \Delta p_k)'$ and $\boldsymbol{d}$ as the arbitrary vector $(d p_1, \ d p_2, \ldots, dp_k)'$.

Then the above results can be written as

1. $$\boldsymbol{\Delta}' \mathrm{M}^{-1} \boldsymbol{\Delta} = \tfrac{1}{2}\sigma_A^2,$$

2. The Kimura principle is: subject to the constraint

$$\boldsymbol{d}' \mathrm{M}^{-1} \boldsymbol{d} = (1/2)\, \sigma_A^2,$$

for an arbitrary vector of allelic frequency changes $\boldsymbol{d}$, the frequency changes that maximize the increase in mean fitness is the natural selection vector of changes $\boldsymbol{\Delta}$.

(Remember the quadratic forms $\boldsymbol{\Delta}' \mathrm{M}^{-1} \boldsymbol{\Delta}$ and $\boldsymbol{d}' \mathrm{M}^{-1} \boldsymbol{d}$.)

An unanswered question. Why impose the actual constraint that Kimura imposed? It has been widely criticized in the literature as being "ad hoc", and with no extrinsic justification.

To answer this, and to consider the whole "information" question much more generally, we (i) remove the assumption of random mating, and (ii) consider the entire genome, not just one gene locus.

Frequency changes written in Greek (either $\boldsymbol{\delta}$ or $\boldsymbol{\Delta}$) in what follows are assumed to be those brought about by natural selection. $\boldsymbol{\delta}$ is a "within-generation" change, $\boldsymbol{\Delta}$ is a between-generation change. Frequency changes written in Roman (i.e. $\boldsymbol{d}$) are arbitrary changes.

We list all the (approximately) $(4^{5,000})^{30,000}$ whole-genome genotypes as genotypes G_1, G_2, ……,

with parental generation population frequencies g_1, g_2, …, and fitnesses w_1, w_2, …

These are (again) normalized so that the parental generation mean population fitness $\boldsymbol{w}$ ($= \Sigma_s \; g_s \, w_s$) is 1.

By definition of the fitness w_s of the (whole genome) genotype # s, the within-generation change of the frequency of this genotype, that is the change in frequency between the time of conception and the age of reproduction, is

$$\delta(g_s) = g_s \{w_s - 1\}.$$

By simple summation, the within-generation change in the frequency of p_{ku} of the allele A_{ku}, allele k at gene locus u, is

$$\delta(p_{ku}) = (1/2) \Sigma_s c_{kus} g_s \{w_s - 1\},$$

Where c_{kus} is the number of times (0, 1 or 2) that this allele occurs in whole genome genotype # s.

For allelic frequencies, within-generation changes are identical to between-generation changes. Thus

$$2\,\Delta p_{ku} = \Sigma_s\, c_{kus}\; g_s\; \{w_s - 1\},$$

where Δp_{ku} is to the "between-generation" change in the frequency of allele # *k* at gene locus # *u* brought about by natural selection.

We now have to define the average effects of all the $(4^{5,000})^{30,000}$ alleles in the entire genome. This is done by a weighted least squares procedure, generalizing that for the "two alleles at one single locus" case. Specifically,…..

If α_{ku} is the average effect of allele *u* at locus *k,* then the various average effect values α_{ku} (u = 1,2,.., k = 1,2,..), are found by minimizing

$$\sum^{s} g_s \left(w_s - w - \sum^{(s)} c_{kus} \alpha_{ku} \right)^2$$

subject to the constraint $\sum_u p_{ku} \alpha_{ku} = 0$ for all *k*.

(The inner sum contains α_{ku} once, twice or not at all, depending on how many times the allele A_{ku} arises in genotype *s*.)

Why do we impose this constraint? It leads to uniqueness: if we did not do it, we could add any constant *c* to the average effects at one locus, and subtract this same constant *c* from the average effects at some other locus, we do not change the sum of squares removed by the regression. The constraint puts all loci "on an equal footing".

.

This leads to a set of non-singular equations

$$M\,\boldsymbol{\alpha} = \boldsymbol{\Delta}$$

Where $\boldsymbol{\alpha}$ is a (huge) vector of average effects of all alleles at all gene loci, $\boldsymbol{\Delta}$ is a (huge) vector of the between-generation natural selection changes in the frequencies of these alleles, M is a huge matrix, whose form is known and can be written down. It is the direct generalization of the one-locus matrix M which had the allelic frequencies displayed along its main diagonal.

From the equation $\mathbf{M}\boldsymbol{\alpha} = \boldsymbol{\Delta}$, the average effects are given by

$$\boldsymbol{\alpha} = \mathbf{M}^{-1}\boldsymbol{\Delta}$$

So that $$\boldsymbol{\alpha}' = \boldsymbol{\Delta}' \mathbf{M}^{-1}.$$

The sum of squares removed by fitting the average effects α_{ku} is the whole genome additive genetic variance σ_A^2. It is found from standard least squares theory that

$$\boldsymbol{\alpha}'\boldsymbol{\Delta} = \tfrac{1}{2}\sigma_A^2$$

Then we get, eventually, $\boldsymbol{\Delta}' \mathbf{M}^{-1}\boldsymbol{\Delta} = \tfrac{1}{2}\sigma_A^2$.

We continue to think of the left-hand side in the equation

$$\boldsymbol{\Delta}' \, \mathrm{M}^{-1} \boldsymbol{\Delta} = \tfrac{1}{2}\sigma_A^{\,2}$$

as "information". (Is it Fisher information? This is not yet shown.) Before seeing what this "information" is telling us, we have to take up another theme.

Theme 3. Fisher's "Fundamental Theorem of Natural Selection"

For convenience we continue to fix the parental generation mean fitness mean fitness at the value 1. However, we do not fix the daughter generation mean fitness at this or any other value. The theorem says:

No matter what form of mating exists (random or otherwise), the whole genome PARTIAL increase in mean fitness (i.e. the increase due to "genes within genotypes", defined later) is exactly the (whole genome) additive genetic variance σ_A^2.

The "partial increase" in mean fitness $\Delta_P(w)$

In the one-locus case, we replace the standard equation

$$w = \sum_i \sum_j P_{ij} w_{ij}$$

by the equally correct

$$w = \sum_i \sum_j P_{ij} (w + \alpha_i + \alpha_j)$$

Then
$$\begin{aligned}
\Delta_P(w) &= \sum_i \sum_j \Delta P_{ij} (w + \alpha_i + \alpha_j) \\
&= \sum_i \sum_j \Delta P_{ij} (\alpha_i + \alpha_j) \\
&= 2 \sum_i \alpha_i \sum_j \Delta P_{ij} \\
&= 2 \sum_i \alpha_i \Delta p_i \\
&= \sigma_A^2 / w.
\end{aligned}$$

(Why do this? Fisher (1930) thought that a concept of the fitness of the genotype A_iA_j that is more useful than the "standard" w_{ij} is $w + \alpha_i + \alpha_j$.)

(But Fisher (1941) seemed to describe the Fundamental Theorem of natural Selection in a different way. His 1941 interpretation is captured in a very elegant way by Lessard. In Lessard's interpretation, the partial change involves an "allele-based" change in genotype frequencies.

$$\Delta_P(w) = \sum_i \sum_j (\Delta P_{ij})(w_{ij})_\alpha = \sigma_A^2/w,$$
$$\text{with } (w_{ij})_\alpha = w + \alpha_i + \alpha_j.$$

$$\Delta_P(w) = \sum_i \sum_j (\Delta P_{ij})_\alpha \, w_{ij} = \sigma_A^2/w,$$
$$\text{with } (\Delta P_{ij})_\alpha = P_{ij}(\alpha_i + \alpha_j)/w.$$

The Fundamental Theorem of Natural Selection can be immediately generalized to the whole genome level. So we now return to a consideration of the theory at that level.

Recall that, at the entire genome level,

$$\boldsymbol{\Delta}' \, \mathrm{M}^{-1} \, \boldsymbol{\Delta} = \tfrac{1}{2}\sigma_A^2.$$

The "entire genome, no assumption about the mating scheme" generalization of the Kimura principle is that, of all arbitrary allelic frequency changes $\boldsymbol{d}$ such that

$$\boldsymbol{d}' \, \mathrm{M}^{-1} \, \boldsymbol{d} = \tfrac{1}{2}\sigma_A^2,$$

the changes maximizing the PARTIAL increase in mean fitness are the natural selection changes.

Interpretation in terms of information…….

Of all possible sets of allelic frequency changes that have the same information content about the average effects of all alleles at all loci in the genome as is provided by the natural selection changes, the natural selection changes maximize the partial increase in mean population fitness.

There is a continuous time parallel result.

What about the constraint $\boldsymbol{d}' \mathrm{M}^{-1} \boldsymbol{d} = \frac{1}{2}\sigma_A^2$?

It is found that maximizing the partial increase in mean fitness subject to this constraint is equivalent to the least squares definition of the average effects, subject to the constraint

$$\sum_u x_{ku} \alpha_{ku} = 0$$

For all loci *k*. This latter constraint is "natural". So the constraint is "natural", not arbitrary.

In the continuous-time case,

$$(1/2)\sigma_A^2 = \dot{\Delta} M^{-1} \dot{\Delta}$$

so that the amount of information about whole genome genotype the average effects of all alleles at all loci in the genome provided by the rates of change of allelic frequencies is half the additive genetic variance.

The continuous-time result parallel to the one just given for discrete time is:- of all arbitrary allelic frequency changes for which

$$(1/2)\sigma_A^2 = \dot{d}' M^{-1} \dot{d}$$

that is, which give the same information about whole genome average effects as do the changes brought about by natural selection, the natural selection changes maximize the partial rate of increase of mean fitness.

More about the P/O covariance.

In more realistic cases, for example with epistatic (interactive) effects between genes at different loci, the parent-offspring (P/O) covariance in fitness is not given by the simple formula $(1/2)\sigma_A^2$.

However, the formula $(1/2)\sigma_A^2$ DOES apply to the P/O correlation when we replace the actual fitness of each genotype, (as did Fisher) by

$$w + \sum_s c_{kus}\, \alpha_{ku},$$

where the sum is taken over all whole-genome genotypes (*s*), and $c_{kus} = 0$, 1 or 2, depending on how many times the allele A_{ku} arises in genotype *s* and the α_{ku} values are assumed to be constants (over time).

A generalization of this applies when we consider any character, e.g. height, not just fitness: evolution gives us the information, as quantified by half the covariance between this character and fitness, about the average effects of this character.