The Evolution of the $G$-matrix:

Long-Term Characteristics of Multivariate Heritable Variation

Benjamin Logsdon

School of Molecular Biosciences, Washington State University, Pullman, WA 99164
TO THE UNIVERSITY HONORS COLLEGE:

As thesis advisor for Benjamin Logsdon,

I have read this paper and find it satisfactory.

Thesis Advisor

12 February 2006
Date
The research problem investigated in this paper concerns the nature of the evolution of heritable variation. Heritable variation, or additive genetic variation, is defined at the population level and is the component of observed phenotypic variation which is passed between generations. This paper deals specifically with the properties of the evolution of additive genetic variation for multiple traits. In the case of multiple traits not only does one have to worry about the variation for any particular trait and how that variation evolves, but also the covariance between two different traits. The hypothesis set forward in this paper postulates that the $G$-matrix stabilizes over evolutionary periods of time under specific (unknown) conditions. The $G$-matrix contains all the information about variation and covariation. The question of the evolution of the $G$-matrix is intriguing because an accurate model of its behavior over evolutionary periods of time would provide insights into the long term evolution of any population of organisms.

$G$-matrix stability is a huge question mark in the field of population genetics. If one could know the long-term evolution of the $G$-matrix, one could predict the evolutionary trajectory of a population across the space of all possible phenotypes. Unfortunately, this problem is very difficult and theoretical analyses are unable to predict $G$-matrix stability.

Computer simulations provide an attractive alternative for studying the structure of the $G$-matrix over evolutionary periods of time. There is precedent for these kinds of studies with various types of computer simulations having been developed to mimic biological processes in natural populations. The methods used in this paper were similar to methods used in previous papers but they were extended to any number of traits as
opposed to only one or two traits. The most significant problem was making sure the model was justifiable from a biological perspective. In the end it was easy to corroborate methods and results from this paper to others and feel confident in the results.

The most significant finding was an exponential distribution of average eigenvalues of the $G$-matrix for the neutral model (a population evolving without selection). This suggested that an exponential distribution of variation could be generated across any number of traits without selection in a finite population. This result does not answer the original hypothesis, but it provides a very significant insight into populations of organisms evolving without selection. Even though the traits were supposedly evolving independently of one another (no correlations in the selection or mutation functions) random correlations were still produced, and these random correlations generated the result of an exponential distribution of average eigenvalues. This research only skimming the top of the nature of the $G$-matrix, which is a very complex and deep problem, but the distribution of eigenvalues was a huge finding since it corroborates with experimental data. This corroboration suggests that perhaps the exponential distribution of variation is a robust phenomenon.

There are many potential further investigations for this research project. The most significant question is whether there is an easier way to generate the distribution of average eigenvalues for any number of traits without running the computer simulations to generate the data. Another interesting question is the effect of correlations in mutation and/or selection between traits on this distribution. If these questions could be answered a greater understanding of the long-term structure of the $G$-matrix could be elucidated.
List of Equations, Figures, and Tables

Equation (1): Genetic Structure of Individual 18
Equation (2): Definition of Breeding Value 18
Equation (3): Mutation Distribution 19
Equation (4): Phenotype Assignment Distribution 20
Equation (5): Fitness Function 20
Equation (6): G-matrix Definition 22

Figure 1: Distribution of Eigenvalues for Wing Shape in Fruit Flies (Mezey and Houle 1035) 25
Figure 2: Distribution of Average Eigenvalues from Computer Simulation (Population of 140) 25
Figure 3: Average Broken Stick Distribution 25
Figure A1: Distribution of Average Eigenvalues from Computer Simulation (Population of 342) 34
Figure A2: Distribution of Average Eigenvalues from Computer Simulation (Population of 683) 34
Figure A3: Distribution of Average Eigenvalues from Computer Simulation (Population of 1366) 35
Figure A4: Distribution of Average Eigenvalues from Computer Simulation (Population of 140) 35
Figure A5: Total Average Genetic Variation Regressions on Number of Traits 36
Table A1: Table of Slopes of Logarithmic Regressions 36
Figure A6: Histogram of 1000 Replicate Bootstrap Simulation 37
Figure A7: Histogram of 200 Replicate Bootstrap Simulation 37
Figure A8: Distribution of Predicted Eigenvalues for 1000 Replicate Bootstrap Simulation 38
Figure A9: Distribution of Predicted Eigenvalues for 200 Replicate Bootstrap Simulation

Table A2: Table of Comparison of the Results of the Simulations to Results of Jones et al.
Acknowledgements

Many thanks to Dr. Richard Gomulkiewicz for helping me with my first big research endeavor. When I started I knew basically nothing about quantitative genetics or G-matrices but now I know a lot more thanks to him. I would also like to thank Dr. Cortland Griswold for key programming tips and discussions. C++ Pointers, memory leaks, and Cholesky decompositions were not intuitive to say the least. Finally I would like to thank the College of Sciences and the University Honors College for the Summer Research Mini-Grant that allowed me to conduct the vast majority of this project.
Introduction

All populations of organisms exhibit variation in many of their complex quantitative traits. Some easily imaginable examples of these kinds of traits would be height, weight, or morphology of organisms in general. The innumerable examples of observed variation in quantitative traits causes frustration for many evolutionary theorists. Many models of evolution assume stabilizing selection, where the extreme values of any particular trait are selected against (so in the example of height, the shortest and tallest members of a population). The stabilizing selection theory shows that variation is destroyed every generation. Without a contrasting force that creates equally as much variation as selection destroys all variation would disappear. This is obviously not the case; there must be a contrasting force that creates variation at an equal rate that selection destroys it. There are quite a few theories about what causes the stability of heritable variation, but no definitive answers as of yet.

When considering what basic biological mechanisms could create variation in a trait or suite of traits mutation comes to mind as the most obvious. Mutation can create new alleles at genetic loci. A greater set of alleles with different potential effects on traits increases the variation of the phenotypes in the population. The equilibrium between variation destroyed by selection and variation created by mutation is known as mutation-selection balance (Lande, The maintenance of 222). There are some confounding factors in this model, one of which is linkage disequilibrium which concerns the effect of linkage between loci on variation. Linked loci are genes that during recombination are not freely recombined and have some probability of staying on the same chromosome after
transmission to eggs or sperm. This can create a similar effect on variation as in the
infinitesimal model (discussed below) (Lande, The maintenance of 224).

Up until Russell Lande proposed an actual theoretical model of mutation-selection
balance, most theorists assumed mutation could not produce enough new variation to
counteract the loss in variation. One of the earlier and less intuitive ways of explaining
the lack of loss in variation is to assume each trait is affected by an effectively infinite
number of loci. Under this assumption selection does not actually destroy variation each
generation; it merely rearranges the variation into correlations between pairs of loci. A
similar event occurs with a finite number of loci, but only if they are linked. Even though
the observed variation is decreased, the total potential equilibrium variation is the same,
and if selection is removed the population will return to that equilibrium variation after a
certain number of generations. This is also known as the infinitesimal model (Bulmer
201-203).

There are unfortunately many complicating factors which make theoretical
analysis of the problem of heritable variation incredibly difficult. One of these factors is
random genetic drift. Random genetic drift occurs in finite populations. Basically, in a
finite population, a portion of genetic information is lost every generation because every
gene in a parent generation can not be transferred and expressed the same way in an
offspring generation. Offspring generations will have a random sampling of the alleles
(different forms of a gene) in the parent generation based on the size of the population.
For larger population sizes the representative frequencies of alleles in the offspring and
parent generation will become more and more similar. On the other hand, the smaller the
population is, the greater the loss of information every generation and the more variable
the combinations of frequencies of alleles will be. This loss of information decreases total variation in a similar way to stabilizing selection (Lande, *Natural selection and* 320-323). Therefore the equilibrium genetic variation for a particular trait may be more accurately described as a mutation-selection-drift equilibrium.

Unfortunately, organisms do not consist of a finite number of independently evolving traits where each trait has genes which only affect that trait. Because of the vastly complex relationships between genes and phenotype, most genes affect a broad suite of traits. This phenomenon where alleles at a specific locus affect multiple traits is known as pleiotropy (Lande, *The genetic covariance* 204). Pleiotropy also means a mutation at a single locus can cause effects on many traits simultaneously. It can also confound analysis of single traits because when many genes affect multiple traits the genetic variation for one trait may not be independent of the genetic variation of another trait. Therefore if selection affects one trait it will indirectly affect the other trait merely because of the connection between the two from pleiotropy (Lande, *The genetic covariance* 204-205).

Other complicating factors include dominance and epistasis. Dominance and epistasis can be characterized as interactions within a locus or between loci respectively. Dominance can be summed up as the tendency for one allele to overpower the expression of another allele and epistasis is the extension of this non-linearity to interactions between loci (Rice, *A general population* 15518). Mendel’s peas are a classic example, where the wrinkled peas are the recessive phenotype and the smooth peas are the dominant phenotype. Each phenotype is created by a specific genotype, which in this case consists of two alleles, or versions, of the pea shape gene. This applies to diploid
organisms, or organisms with two complementary sets of genes. Hence, there are four possible combinations of alleles. With $W =$ smooth and $w =$ wrinkled, the combinations are $WW$, $wW$, $Ww$, and $ww$. $WW$, $wW$, and $Ww$ will all have the same smooth phenotype or observed characteristic of producing smooth peas. On the other hand $ww$ will produce wrinkled peas. Therefore the $W$ allele is dominant because it overpowers the $w$ allele. In this example there is complete dominance because $W$ always completely overpowers $w$ (Hartl and Jones 36). Yet there can be varying degrees of dominance, especially in more complicated traits (such as height, weight, wing shape, etc...). All this means is that one allele always (of which there can be many possible, not just two) has a greater effect than other alleles when the effects of the alleles are combined to produce a genotype (Falconer 119).

For many quantitative genetic traits the assumption is made that the effects of the alleles at every locus affecting the trait can be added together to produce the total genetic contribution to the phenotype (Lande, The maintenance of 223). This assumption ignores any dominance or epistasis which can be good or bad, depending on the type of genetic system one chooses to model. For many general theoretical models this is not necessarily a bad assumption since many genetic systems can be approximated by this assumption especially if there are many genes of small effect (Lande, The maintenance of 221). Aside from the assumption of additivity, other assumptions sometimes are made. One such assumption is the independence of genetic variation and environmental variation. Under this assumption the total variation in a trait is merely the sum of the genetic and environmental components of variation. This occurs because the genotype does not always completely determine the phenotype; variations in phenotype are also influenced
by the environment. The independence of environment and genotype basically means that the environment does not have an effect on genotype or vice versa (Lande, *Quantitative genetic analysis* 405).

One more assumption is known as the continuum of alleles assumption which plays a large role in the justification of the model used in this paper. Basically, with the continuum of alleles one assumes that there are an infinite number of possible alleles at every locus. The distribution of alleles at any loci across a population is an approximately normal distribution (Crow and Kimura 495).

For multivariate analysis of traits one studies the G-matrix. The G-matrix is a matrix of the additive variances in genotype across a population for specific traits (along the diagonal) and additive covariances in genotype across a population between specific traits (the respective off-diagonals). Since the covariance between trait one and trait two will be the same as the covariance between trait two and trait one, the matrix is symmetric. This matrix carries all the information about genetic variation and covariation in a population for an entire suite of characters and can be used to predict the response in mean phenotype to selection (Lande, *Quantitative genetic analysis* 406). The structure of the G-matrix is very important when considering the evolution of a population of organisms. The basic fuel of evolution is variation. This makes intuitive sense because the greater the variation, the faster a population will approach an optimum phenotype (Steppan et al. 321). The G-matrix defines the structure of possible evolutionary trajectories across the space of all possible phenotypes. Certain combinations of traits may have lots of variation, yet at the same time certain combinations of traits may have little or no variation. What this effectively means is
there may be some directions in the space of all possible phenotypes in which a selection function might try to send a population that are prohibited by the $G$-matrix. This means that the population cannot evolve in that direction and this limits the possible phenotypes allowed for any future population (Steppen et al. 321).

To predict the response of the mean phenotype to selection, one constructs a gradient vector $\beta$ which specifies the direction in phenotype space which selection would like to send the phenotype (or the direction towards the optimum phenotype). Then multiply the gradient vector $\beta$ by $G$ to acquire a vector which contains the change in mean phenotype for each trait. This is also known as the breeder’s equation (Lande, Quantitative genetic analysis 406). In terms of prediction, this is a very powerful and applicable tool to predict how a population will evolve over short periods of time. Hence, if one could predict the long term stability of the $G$-matrix, one could theoretically know the set of all possible evolutionary trajectories based upon whatever arbitrary selection function utilized. But as described earlier, the evolution of heritable variation, or the $G$-matrix, is a very complex problem, and whether or not long term $G$-matrix stability is a robust phenomenon is not known (Jones, et al. 1757-1758).

The $G$-matrix is influenced by a number of factors, including multivariate fitness functions (Arnold, et al. 17), pleiotropy, and linkage disequilibrium (Lande, The genetic covariance 204). A multivariate fitness function includes modified fitness for combinations of traits that is different than the fitness of the two traits taken independently (Arnold, et al. 17). Russell Lande developed a model to try to explain the evolution of the $G$-matrix that included these effects in his 1980 paper “The Genetic Covariance between Characters Maintained by Pleiotropic Mutations,” but many people
later criticized this model. One of the major criticisms is of the continuum of alleles assumption, where the distribution of allelic effects at all loci is approximately Gaussian. This may not be a good approximation for mutation rates lower than $10^{-4}$ according to Michael Turelli. Therefore other models of the distribution of allelic effects at loci were created that were more complicated but worked with lower mutation rates (Turelli 166-167). Another problem with Lande’s model is that it does not take into account random genetic drift. Lande does analyze multivariate drift in an earlier 1979 paper “Quantitative Genetic Analysis of Multivariate Evolution Applied to Brain: Body Size Allometry,” but it is not incorporated explicitly into his later paper.

If a stable equilibrium solution to all contributing factors of $G$-matrix structure could be deduced then extrapolations of the evolution of populations could be made. The more stable the $G$-matrix is over a period of time, the longer the extrapolations based upon that $G$-matrix can be. This is important when one tries to build a bridge between microevolution and macroevolution. Predicting short term responses to selection exists in the domain of microevolution, whereas predicting the long term evolution of a population, including speciation, is the domain of macroevolution. Being able to connect the two with the $G$-matrix is a powerful, though currently incomplete approach (Jones, et al. 1747).

There also exists another important theory of evolution, known as the neutral theory of molecular evolution, which can establish a baseline to predict whether observed variation is the product of just mutation and drift (Lynch 915). In essence, the neutral theory assumes no selection, just a population diffusing across a phenotypic landscape with no directing force. One of the most interesting aspects of this type of variation is
that within a population a stochastic equilibrium is reached for the observed variation (Lynch 928). Variation does not grow without bound from mutation because random genetic drift eventually removes variation at the same rate which mutation produces variation. Hence, in a two dimensional phenotypic space, one could imagine an ellipse which describes the variation for both traits and their respective correlations fluctuating in size and shape, but staying approximately the same shape over time. Yet, these ellipses would randomly walk over the landscape like particles diffusing in a liquid. Therefore the phenotypic values of two populations starting at the same point would eventually travel away each other, since there would be no force of selection to push towards an optimum phenotype. But at the same time their G-matrices would be proportional to the original G-matrix when the two populations diverged (Arnold et al. 14). The neutral theory will be important later on, since most of the results from the various experiments were conducted without the force of selection.

Because of the difficulty of predicting the long-term stability of the G-matrix with either analytical or numerical techniques, it is often useful to employ computational modeling. Individual based models or Monte Carlo simulations can be used to model the dynamics of artificial populations of organisms and study phenomenon such as the G-matrix as a function of time. Other phenomenon can be studied as well, such as specific aspects of the G-matrix, like its eigenvalues, eigenvectors and their interrelationships (Jones, et al. 1748-1750). Plus, one can run many replicate runs of an evolving population using the same parameters and then construct distributions for different results. Since the evolution of the G-matrix in finite populations is an inherently stochastic process, what with mutation, recombination, and random mating, it is useful to
do replicate runs and study the distribution of results produced to search for any general overarching phenomenon. Adam Jones, et al., in their 2003 paper, “Stability of the G-Matrix in a Population Experiencing Pleiotropic Mutation, Stabilizing Selection, and Genetic Drift,” developed a framework for studying the stability of the G-matrix using Monte Carlo simulations, but only for a bivariate case (two traits). They looked at the effects of varying population sizes, varying strengths of selection, varying correlations in selection and varying correlations in mutation (1748-1750). The results of their paper were very interesting, and support some of the results from the experiments produced in this paper.

Many of the methods and parameters used in this paper are derived from the methods and parameters in their paper so the results from both models can be compared to check for validity of this model. For their model they assumed additivity of mutation: whenever a mutation event happened, the effect of the mutation was sampled from a normal distribution with a mean of zero and a variance of 0.05, and added to the value of the allele at the locus where the mutation event occurred. Their mutation rate was 0.0002, which they justified from an earlier study. Their selection function was a Gaussian style function, like the selection function described in the methods. They ran almost all of their experiments for 12,000 generations. The first 10,000 generation allowed the population to reach mutation-selection-drift equilibrium, then for the last 2,000 generations statistics about the population were collected. Their statistics illustrated whether or not the G-matrix fluctuated wildly or mildly over time (Jones, et al. 1748-1750).
Two other significant computational simulations are illustrated in (Buerger, et al. 1753-1754) and (Buerger and Lande 903-904). Jones’s methods are similar to these earlier models, though Buerger and Lande use a different mutation scheme, which is also utilized in this paper. The different mutation scheme is as opposed to adding a random variable sampled from a distribution, just sampling a variable from a distribution for the mutated value of an allele at a loci. In their paper they use multiple different kinds of distributions, including normal distributions and gamma distributions (Buerger and Lande 905).

**Hypothesis**

The additive genetic variance-covariance G-matrix stabilizes over evolutionary periods of time under specific conditions.

**Methods**

This individual based model is a Monte Carlo simulation created to observe mutation-drift-selection balance for a population of individuals with any number of polygenic traits and with any number of pleiotropic loci. It is an iterative model, characterized by the following sequence of events: assignment of phenotype for each individual, assignment of relative fitness for each individual, mutation, recombination, and random mating. This process is repeated until the population reaches mutation-drift balance or mutation-drift-selection balance based upon the type of experiment being performed.
Since this model is diploid, the genetic structure of each individual is stored as an (m)x(n)x(2) array, where m is the number of traits and n is the number of loci as shown in (1). K is the genetic contribution from the maternal set of alleles and K' is the genetic contribution from the paternal set of alleles.

\[
K = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1n} \\
  x_{21} & x_{22} & \cdots & x_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{m1} & x_{m2} & \cdots & x_{mn}
\end{bmatrix}
\]

\[
K' = \begin{bmatrix}
  x'_{11} & x'_{12} & \cdots & x'_{1n} \\
  x'_{21} & x'_{22} & \cdots & x'_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  x'_{m1} & x'_{m2} & \cdots & x'_{mn}
\end{bmatrix}
\]

All of the loci are pleiotropic, which means that any locus (column in K) can potentially affect any trait (row in K). Only additive effects are considered in this model and the element x_{ij} is the effect of an allele at i^{th} locus on the j^{th} trait. The breeding value of the individual for each trait is defined as the sum of effects at each locus for a specified trait (2). It is assumed that both “male” and “female” alleles are summed to acquire the breeding value for any particular trait. The vector \(x_i\) represents all the effects at one locus on all traits, or the i^{th} column of K.

\[
x = \sum_{i=1}^{n} x_i + x'_i
\]

The additive effect at each locus can potentially take on any value since two mutation continuous mutation schemes are utilized (as described below). These values are sampled from a distribution (3) (below) with a specific mutational covariance matrix.
\( V_m \). It is assumed that \( V_m \) is the same for each locus. All simulations described in this paper assume that \( V_m \) is diagonal with the same value (variance) along the diagonal.

\[
\xi(x_i) = \frac{1}{\sqrt{(2\pi)^m |V_m^{-1}|}} \exp\left\{ -\frac{(x_i)^T V_m^{-1} (x_i)}{2} \right\}
\]

(3)

The phenotype for each individual is sampled from a distribution which is dependent on breeding value and environmental variance. Using these definitions the genetic structure of the entire population is defined by an \((m)\times(k)\times(2)\times(N)\) array, where \( N \) is the number of individuals. The original population can be generated with any distribution of effects across all loci, but since the optimum for the population (in simulations with selection) is set to zero, most simulations are run with an initial population where the effects at all loci are arbitrarily set to zero. The original distribution of effects should not matter when considering the long term equilibrium of the additive genetic variance. It will not matter because any original distribution of effects will be destroyed by mutation, random genetic drift, and selection.

After the initial genesis of the population, phenotypes of each individual are assigned by sampling from a random multivariate normal distribution with a mean equal to the breeding value and covariance matrix equal to the environmental covariance matrix (4). \( V_e \) is the environmental covariance matrix, which is assumed to be unity along the diagonal.
It is assumed there are no genotype environment interactions. The phenotype for each individual is characterized by a vector of length \( m \). To assign a relative fitness to each individual, a multivariate Gaussian-like function with a mean of zero and a specified variance \( (V_s) \) is applied with no correlations between fitness of different traits \( (5) \).

\[
W(z) = \exp{\left[ -\frac{(z)^T V^{-1}_s (z)}{2} \right]}
\]

\( V_s \) of this distribution is directly related to the relative strength of selection. The higher \( V_s \) is, the weaker selection is. Most of the simulations discussed in this paper were done with a pure drift model, hence selection was not considered so the relative fitness of all individuals was equal to one.

Once relative fitness is assigned then the overall mating function is applied. To try to minimize computation time, only individuals that are selected to mate are recombined and mutated, as opposed to recombining and mutating the entire population. The recombination frequency is one half for all loci, and two different mutation functions are simulated. The first mutation scheme is mutation scheme A, where a mutation event occurs with a specified frequency, and the magnitude of the change in effect is sampled from \( \xi(x_i) \) and then added to the original effect at that locus. The other mutation scheme is mutation scheme B. In mutation scheme B, the change in effect replaces the original
effect at a locus instead of being added to it. Since this is a pleiotropic model, when a mutation event occurs at a locus it changes the effects of that locus on all the traits, so the entire vector $x_i$ is changed. Each element of $x_i$ changes independently of the other elements of $x_i$ because of the assumption of no correlations in mutation between traits (e.g. $V_m$ is diagonal with mutational variance along the diagonal).

The mating scheme is a roulette wheel style system. An individual is selected at random from the population. Then a random number between zero and one is generated. If the random number is greater than the relative fitness of that individual then that individual is returned to the population. On the other hand, if the random number is less than the relative fitness of that individual, the individual is kept, and another individual is sampled randomly from the population and selected in the same way to mate with the first individual. This model does allow for selfing. Because this is a diploid model independent segregation occurs and the four different combinations of male and female sets of alleles from each parent occur with equal frequency. Mating events continue to occur until the population size of the new population is equal to that of the old population. Since individuals are sampled randomly, they can be sampled more than one time. When the new population replaces the old population completely the process is started all over again with the assignment of phenotype.

For all the simulations discussed in this paper the populations were modeled for 12,000 generations. The $G$-matrix was defined as the covariance matrix of additive genetic effects across the population (6) (below). Each vector $x_{i\text{th}}$ was the breeding value $x$ for the $i^{th}$ individual as defined in (2).
The $G$-matrix was evaluated and stored for each generation. The properties of the $G$-matrix were studied for the last 2,000 generations since it was assumed after 10,000 generations the population reached mutation-drift balance. This was done originally to compare results of this model with a similar two trait model developed by Jones, et al. In fact, equations like (1) through (6) are used in Jones’s paper as well, except they only consider two traits, and the equations in this paper are generalized to any number of traits. Of course these equations are derived from earlier papers such as Russell Lande’s 1975 and 1980 papers. Other parameters used were very similar to parameters in their model. The mutation frequency was set at 0.0002, the mutational variance was set at 0.05, and the effective population sizes studied were the same as in their paper for population sizes of 342, 683, and 1366. The number of loci was set at 50 just as in their simulation and Buerger and Lande’s simulation.

The same analysis functions for the stability of the $G$-matrix were applied as in Jones’s paper. For all of these analysis functions the $G$-matrices for each generation from twenty separate simulations of populations with the exact same parameters were collected and then averaged so as to remove temporal auto-correlations. The analysis functions included the average $G$-matrix, the average eigenvalues, the average total amount of variation, and the average ratio of large to small eigenvalues. They also included the average difference between two successive generations for the $G$ matrix, the eigenvalues, and the ratio of large to small eigenvalues. Finally the average change in direction of the principal component between generations was also measured. The
functions that were of the greatest interest were the average eigenvalues and the average total amount of variation (sum of the eigenvalues). The other analysis functions were not of as much concern for most of the experiments. Hence the results for all of the analysis functions from a similar experiment to one that Jones, et al., did are compared in Table A2, but are not emphasized in the results section, aside from the average eigenvalues and average total variance.

As will be presented later, the average eigenvalues for many different parameter sets appeared to fit an exponential distribution. In an attempt to explain this distribution using the idea of a broken stick heuristic (Figures 3, A8, and A9), two new simulations were developed. In these simulations a large number replicate populations were run with a single trait, no selection, 6000 and 12000 generations with 1000 and 200 replicates respectively, and all other parameters the same as before. Within each simulation for every replicate the genetic variation at the last point in its evolution was sampled and a frequency distribution was constructed from all the replicates. From this frequency distribution a bootstrapping technique was used, where \( n \) values were sampled from the distribution and summed together, where \( n \) was the number of traits in question. This sum was then broken at \( n-1 \) random locations. The pieces were then ordered by size and stored in an array. This process was repeated between 200 and a 1000 times and the average taken for all the pieces. The idea was that this would produce a similar distribution as the eigenvalue distribution. With the justification that all the traits should be evolving independently, and the most random way to distribute variation is to repeatedly sample from the distribution of variation, sum the samples, then break the sum up randomly, assuming that variation explained by the distribution of eigenvalues
distributed as randomly as possible. The results of this experiment will be discussed in the results section.
Results and Discussion

Figure 1- Log$_{10}$(eigenvalues) from G-matrix for twenty quantitative traits from fruit flies. (slope of -0.133 ± 0.005) (Mezey and Houle 1995)

Figure 2- Normalized average eigenvalue distributions from Monte Carlo simulations of a population of 140 with 50 additive pleiotropic loci utilizing mutation scheme A.

Figure 3 – Average broken stick distribution for stick of length one with 19 break points. These lengths are the result of breaking a “stick” of fixed length 19 times, ordering the lengths by size, and repeating the process over and over again to attain an average value for each ordered piece.
The distribution of average eigenvalues approaches an exponential distribution from the forces which create mutation-drift balance (Figures 2, A1, A2, A3, and A4). The source of the non-uniformity in eigenvalues (or non-equal partitioning of total genetic variation between combinations of traits) is from the temporal random correlations created between traits because of a finite population size from mutation and drift. The data supports this conclusion, because as the population size increases, the slope of the logarithmic regression for eigenvalues decreases. For an infinite population variation should be evenly distributed among each trait, assuming no correlations in mutation or selection. Also, the slope of the logarithmic regression of average eigenvalues decreases as the number of traits increases, though whether the slope of the logarithmic regression for either very large populations or populations with many traits reaches a finite asymptotic value or zero is an open question. Also, adding more traits increases the total average genetic variation in a linear fashion, such that the total average genetic variation for any number of traits is a linear function of the number of traits. This makes sense intuitively, because each trait is evolving independently of all other traits and if each trait has an associated amount of genetic variation associated with it then the total genetic variation should be the sum of the individual trait’s genetic variation (Figure A5).

There also appears to be deviations away from a purely exponential distribution as the number of traits increases in size (Figures 2, A1, A2, A3, and A4). The deviation is subtle, with the second largest eigenvalue systematically below the logarithmic regression best fit line, and the second smallest eigenvalue systematically above the logarithmic regression best fit line. The deviations are qualitatively similar in structure to
qualitative deviations from an exponential distribution in the average broken stick
distribution (Figure 3). This suggests that the distribution of variation across
combinations of traits is as random as possible. Yet the slopes for the logarithmic
regressions are much too low in comparison to a logarithmic regression of the average
broken stick distribution. For example, the data from the experiments for five traits gives
a regression with a slope of -0.480 to -0.205 for population sizes from 140 to 1366
respectively (Table A1), but the logarithmic regression of the average broken stick
distribution of five pieces gives a slope of -0.5918. This suggests the lack total
independence in breaking up the total variation which a broken stick distribution would
support.

The results of the bootstrapping experiment (which attempted to use the broken
stick distribution to explain the distribution of average eigenvalues) failed. The
distribution of average values it produced was merely the average broken stick
distribution scaled by a constant factor (Figures A8 and A9). The distribution of
variation which the bootstrapping experiment was sampled from is depicted in Figures
A6 and A7. It failed for both the 200 replicate experiment and 1000 replicate
experiment. This indicates the relationship between average eigenvalues for many traits
cannot be generated using the distribution of genetic variation of a single trait along with
a broken stick heuristic. As discussed earlier, there are qualitative features of the
distribution of average eigenvalues that are similar to the average broken stick
distribution. There still may be some significant relationship between the broken stick
distribution or a modification of the broken stick distribution and the distribution of
average eigenvalues.
The question of under what exact conditions the G-matrix is stable has not yet been answered, and for the neutral model the G-matrix is very unstable, with the average eigenvalues following the exponential distribution, but the eigenvectors swinging wildly about. But there are still properties about the G-matrix which are stable even with random fluctuations in direction (or apparent instability). One of the most exciting results of these experiments is that the exponential distribution of eigenvalues has been observed in nature. One significant example is from a paper on the dimensionality of genetic variation for wing shape in fruit flies (Mezey and Houle 1035). Mezey and Houle found an exponential relationship for a 20 eigenvalue system which described the variation in wing shape for 24 metric traits (Figure 1). The range of their calculations for the slope of the logarithmic regression (base 10) were between -0.130 and -0.183. In natural logarithmic base, this range corresponds to slopes between -0.299 and -0.422. Interestingly enough, the slopes produced in the experiments in this paper by drift and mutation have a range of -0.190 to -0.684 for different population sizes and two through six traits (Table A1). It seems surprisingly coincidental that the range of slopes in their paper falls within the range of slopes produced by these computational experiments. This supports the conclusion that genetic variation is distributed exponentially and the exponential nature of this distribution is a stable phenomenon. In a different paper an exponential distribution of average eigenvalues was also observed for morphological traits, though the slopes were much steeper, with the leading eigenvalue accounting for most of the variation (Kirkpatrick and Lofsvold 960). Finally, there is a theoretical basis for the distribution of eigenvalues being approximately exponential derived from random correlation matrices. Wagner argues that deviations from the expected distributions of
eigenvalue for a random correlation matrix can occur when functionally different characters are measured. For example, if thorax shape and eye color in insects were measured. He also argues that within a morphologically similar set of characters, the distribution of eigenvalues will not deviate from the distribution of eigenvalues for a random correlation matrix. In this case different characters from a mammalian skeletal system would be a good example (Wagner 92-93). It is hard to say whether or not the data from these experiments supports Wagner’s conclusions. Obviously random genetic drift is a significant force in tuning the slope of the logarithmic regression of the average eigenvalues and Wagner admits his models of random correlation matrices can not account for either selection or random genetic drift.

Many interesting questions remain open. One of the most significant questions is whether or not correlations in selection and/or mutation influence the exponential distribution of eigenvalues. It seems logical that selection and mutation would conform the G-matrix to the representative structures of V_s and V_m or some function of the two. This was one of the results of Jones, et al. They showed that the stronger the correlations, the greater the stability of the G-matrix, especially with mutational correlations (Jones, et al. 1755). Obviously for the neutral case the eigenvectors are fluctuating wildly in direction, but the average relationship between lengths of eigenvectors follows an approximately exponential distribution.

Many other parameters in this experimental setup could be varied to study equilibrium G-matrix dynamics. The strengths of mutation and/or selection are parameters which could be varied. Another interesting question would be to see the effect of the number of loci on these results. Fifty loci is a fairly large number of loci, so
it would be interesting to see the dynamics of the eigenvalues for fewer loci. Finally, discrete allelic models, such as biallelic models or triallelic models could also be investigated, where instead of an infinite number of possible alleles at every locus; only two or three discrete alleles would be allowed. This type of model could be reconciled more easily with broad theoretical results since more analysis have been done on discrete allelic genetic systems.

Since both these simulations and Jones’s simulations gives similar results for the same parameters (Table A2), the methodology utilized in this paper appears to be robust. Hence, under specifically different but functionally similar setups similar results can be expected. For example, Jones’s mating scheme and this mating scheme are different, yet random mating still occurs in general. This project could have been approached from either a more theoretical approach or a numerical approach. Unfortunately both of those approaches are much more difficult than running Monte Carlo simulations. A theoretical approach would most likely have involved attempting to derive partial differential equations for the diffusion of alleles throughout the population. Yet that would be a very non-trivial endeavor. A numerical approach would probably have focused on developing a Markov Chain approach. This would involve setting up appropriate transition kernels, assuming the ergodic theorem, and studying multiple chains after thousands of generations. This would also be a non-trivial endeavor.

The many assumptions made in this paper do add certain biases to the results. One of the major assumptions is the continuum of alleles assumption, where an infinite number of possible alleles can exist at any locus. This may or may not be a good assumption based upon what system one is trying to model. Another assumption is the
inherent notion of additivity of the effects of alleles at all loci. This assumption makes modeling much easier, but is probably not very realistic. Dominance is a significant effect that includes within locus interactions and the number of possible epistatic interactions or between loci interactions of alleles could be very vast. Hence these assumptions weaken the broad generality of the observed results. But in another sense, this model is more general than Jones’s model because any number of traits can be modeled, allowing for dynamics beyond the range of a bivariate system.

Conclusion

This project illustrates the power of Monte Carlo simulations when considering very difficult, highly complex models of systems where there are many possible deterministic and random interactions. Individual based models or Monte Carlo simulations give a more intuitive feel to the dynamics of such complicated models, and salient as well as novel features of the model can be discovered. In this case, the exponential distribution of average eigenvalues for the G-matrix was merely a phenomenon that arose in the results of the data that was being collected. Obviously Monte Carlo simulations have their weaknesses, inasmuch as they can be computationally expensive, as well as they are difficult to program. This model was eventually programmed in C++ to reduce running time. They also do not give general results and there are often many assumptions invoked when creating them.

The larger ramifications of the stability of the exponential distribution of average eigenvalues of the G-matrix could be profound. This would be true especially if this distribution’s dynamics in the case where correlated mutation and selection functions
were applied was also exponential, or deviated from exponential in a predictable fashion. In which case, a general theory of distribution of variation across any number of traits could be formulated and tested on real populations. If this distribution could be predicted from a few measurable parameters, then it could be possible to predict the general structure of a $G$-matrix for a population of organisms. If this was the case one could work towards the ultimate goal of predicting all possible evolutionary trajectories through the space of all possible phenotypes, and connect microevolution to macroevolution.
Appendix

Definitions

- **Mutation Scheme A**: Mutation by sampling a value from a multivariate normal distribution then adding it to the value of the current allele at a particular locus.

- **Mutation Scheme B**: Mutation by sampling a value from a multivariate normal distribution then replacing the value of the current allele at a particular locus.

- **Normalized Eigenvalues**: Each eigenvalue is divided by the sum of the eigenvalues such that their new sum adds up to one.

- \( \lambda_i \): the mean \( i^{th} \) eigenvalue of the \( G \)-matrix

- \( \Sigma \): the mean sum of the diagonals of the \( G \)-matrix (total variation)

- \( \varepsilon \): the mean smaller eigenvalue divided by the mean larger eigenvalue (for two traits only)

- \( G_{ij} \): the \( i,j^{th} \) value of the \( G \)-matrix

- \( \Delta G_{ii} \): the average change between two successive generations for the \( i,i^{th} \) value of the \( G \)-matrix, divided by the mean value

- \( \Delta G_{ij} \): the average change between two successive generations for the \( i,j^{th} \) value of the \( G \)-matrix

- \( \Delta \lambda_i \): the average change between two successive generations for the \( i^{th} \) eigenvalue divided by the mean \( i^{th} \) eigenvalue

- \( \Delta \Sigma, \Delta \varepsilon \): the average change between two successive generations for \( \Sigma \) and \( \varepsilon \) divided by their respective means.

- \( \Delta \varphi \): the average change between two successive generations between the principal components of the \( G \)-matrix.
Figure A1-Normalized distribution of average eigenvalues utilizing mutation scheme A at mutation-drift equilibrium for a population of 342

Figure A2-Normalized distribution of average eigenvalues utilizing mutation scheme A at mutation-drift equilibrium for a population of 683
Logarithm of Average Eigenvalues for Population Size of 1366

![Graph showing logarithm of average eigenvalues for population size of 1366.

Figure A3-Normalized distribution of average eigenvalues utilizing mutation scheme A at mutation-drift equilibrium for a population of 1366.

Logarithm of Average Eigenvalues for Population Size of 140

![Graph showing logarithm of average eigenvalues for population size of 140.

Figure A4-Normalized distribution of average eigenvalues utilizing mutation scheme B at mutation-drift equilibrium for a population of 140.
Figure A5- Regressions of total average genetic variation versus number of traits for different population sizes

<table>
<thead>
<tr>
<th>Mut. A</th>
<th># of Traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop. Size</td>
<td>2</td>
</tr>
<tr>
<td>140</td>
<td>-0.684</td>
</tr>
<tr>
<td>342</td>
<td>-0.516</td>
</tr>
<tr>
<td>683</td>
<td>-0.381</td>
</tr>
<tr>
<td>1366</td>
<td>-0.314</td>
</tr>
<tr>
<td>Mut. B</td>
<td>-0.698</td>
</tr>
</tbody>
</table>

Table A1- Table of slopes of normalized logarithmic regressions of distribution of average eigenvalues for mutation schemes A & B
Figure A6-Histogram of genetic variation for a single trait population of 140 with mutation scheme B with 1000 replicates

Figure A7-Histogram of genetic variation for a single trait population of 140 with mutation scheme B with 200 replicates
Figure A8-Normalized distributions of predicted average eigenvalues for a population of 140 with mutation scheme B from the bootstrapping experiment with 1000 replicates.

Figure A9-Normalized distributions of predicted average eigenvalues for a population of 140 with mutation scheme B from bootstrapping experiment with 200 replicates.
Table A2: Comparison of Jones et al. data and data from the multivariate extension in this paper. In this case, the population size is 342, the strength of selection is 9 (strong), and the number of traits is two. The analysis functions are described earlier in the appendix in the definitions section.

<table>
<thead>
<tr>
<th></th>
<th>$V_g$</th>
<th>$G_{11}$</th>
<th>$G_{22}$</th>
<th>$G_{12}$</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\Sigma$</th>
<th>$\varepsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones et al.</td>
<td>9</td>
<td>0.201</td>
<td>0.192</td>
<td>0.008</td>
<td>0.24</td>
<td>0.15</td>
<td>0.39</td>
<td>0.64</td>
</tr>
<tr>
<td>Multivariate Extension</td>
<td>9</td>
<td>0.203</td>
<td>0.183</td>
<td>0.004</td>
<td>0.24</td>
<td>0.15</td>
<td>0.39</td>
<td>0.63</td>
</tr>
<tr>
<td>$\Delta G_{11}$</td>
<td></td>
<td>$\Delta G_{12}$</td>
<td>$\Delta \lambda_1$</td>
<td>$\Delta \lambda_2$</td>
<td>$\Delta \Sigma$</td>
<td>$\Delta \varepsilon$</td>
<td>$\Delta \varphi$</td>
<td></td>
</tr>
<tr>
<td>Jones et al.</td>
<td>0.074</td>
<td>0.075</td>
<td>0.010</td>
<td>0.062</td>
<td>0.091</td>
<td>0.055</td>
<td>0.10</td>
<td>9.51</td>
</tr>
<tr>
<td>Multivariate Extension</td>
<td>0.082</td>
<td>0.082</td>
<td>0.011</td>
<td>0.082</td>
<td>0.079</td>
<td>0.061</td>
<td>0.11</td>
<td>18.63</td>
</tr>
</tbody>
</table>
Bibliography


REQUEST FOR PERMISSION TO INCLUDE YOUR HONORS STUDENT'S THESIS IN THE WSU RESEARCH EXCHANGE

Washington State University Libraries

Permission is requested for a non-exclusive license to post the Honors thesis described below in digital form in the Honors College community within the WSU Research Exchange <https://research.wsulibs.wsu.edu:8443/dspace/community>. Posting in the Research Exchange will make the material publicly available as part of the Washington State University Research Exchange digital repository of research-related documents. Additional information about Research Exchange can be viewed at <http://research.wsulibs.wsu.edu>.

Thesis Author Logsdon, Benjamin

Thesis Title The Evolution of the G-Matrix: Long-Term Characteristics of Multivariate Heritable Variation

Date of Thesis Spring 2006

I grant a non-exclusive right to include this item in the Research Exchange. All other rights under copyright law are retained.

Permission granted by:

Richard Gomulkiewicz
Name (please print)

Signature (not computer generated)

Telephone / Email

Date of Signature

Please return the form with your written signature via fax, email, or postal service to:

Kay Vyhnaneck, Scholarly Communication Librarian
120F Terrell Library / PO Box 645610
Washington State University
Pullman, WA 99164-5610 USA
Fax: +1-509-335-672