36

Comparisons of G and its Stability

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. — Tukey (1986)

Version 26 June 2014

As shown in the previous chapter, theory offers some hints that parts of $\mathbf{G}$ (its orientation) may be somewhat stable over modest amounts of time, but also suggests the $\mathbf{G}$ can dramatically change over short periods of time. Hence, the stability of $\mathbf{G}$ remains foremost an empirical question, and this chapter develops the tools to compare $\mathbf{G}$. We start by considering the expected change in $\mathbf{G}$ under drift, examining theoretical and (more importantly) empirical work on this subject. The next section forms the bulk of this chapter, and focuses on proposed methods to compare two (or more) $\mathbf{G}$ matrices. A number of approaches have been proposed, all with flaws, but some are still useful. We then consider the empirical evidence for stability of $\mathbf{G}$, which is a bit of a mixed bag. The basic conclusion is that the data do show that the orientation of $\mathbf{G}$ remains more stable than its eigenvalues, but this general observation could also be a result of weak power. We conclude with a brief discussion on methods for estimating the dimensionality of $\mathbf{G}$ and working with reduced-rank $\mathbf{G}$ matrices.

CHANGES IN G UNDER DRIFT

A few comments regarding the expected change in the additive genetic covariance matrix under drift are in order, as one issue when comparing $\mathbf{G}$ matrices from different populations is whether an observed difference is simply due to drift. Recall from Chapter 5 that the change in the genetic variance under a strictly additive model has very nice properties. First, additive variance is strictly decreasing over time, with $\sigma^2_A(t) = (1 - f_t)\sigma^2_A(0)$ where f_t is the inbreeding coefficient at time t. Second, if we consider the change in the mean under inbreeding (in a strictly additive model), there is no *net* change, but the distributions of means at time t has variance $2f_t\sigma^2_A(0)$. Finally, when nonadditive variance is present, the change in σ^2_A under inbreeding is *not* predictable simply given $\sigma^2_A(0)$ and f_t. Additive variance can actually *increase* for a time under inbreeding, either by epistatic variance being converted into additive variance and/or by the presence of dominance. Finally, when dominance is present, the mean can also show inbreeding depression, and hence itself show a net change under drift. How do these observations generalize when considering covariance matrices?

Under Additivity, G Shows an Expected Proportional Decrease

Under the strictly additive model (no dominance or epistasis), as we have already mentioned several times, the *expected* change in $\mathbf{G}$ is a proportional decrease,

$$E[\,\mathbf{G}(t)\,] = (1 - f_t)\mathbf{G}(0) = \left(1 - \frac{1}{2N_e}\right)^t \mathbf{G}(0)$$

as noted (in various forms) by Wright (1951), Lande (1979, 1980), Roff (2000), and López-Fanjul et al (2004). Indeed, Roff (2000) has even stressed that a proportional decrease in $\mathbf{G}$

is an indication that drift must be considered. In such a case, the matrices being compared would have the same orientation (same eigenvectors) and proportional eigenvalues (i.e., $\lambda_i(1) = c \cdot \lambda_i(2)$ where $\lambda_i(k)$ denotes the ith eigenvalue for the $\mathbf{G}$ matrix from population k). Likewise, if one has a set of replicates all originating from a common ancestral population and only drift has been operating, the distribution for the vector $\boldsymbol{\mu}(t)$ of means in generation i is multivariate normal with mean vector $\boldsymbol{\mu}(0)$ (the original mean, so no *net* change) and variance-covariance matrix $2f_t\mathbf{G}(0)$. Hence, $\mathbf{G}$ determines the distribution of means among replicate populations under drift, with the largest dispersal occurring along those directions of $\mathbf{G}$ showing the greatest variation — the leading eigenvectors of $\mathbf{G}$. Thus, under strict drift, populations tend to evolve along Schluter's (1996) genetic lines of least resistance (Chapter 30).

The Experimental Results of Phillips, Whitlock, and Fowler

Phillips, Whitlock, and Fowler (2001) tested the validity of the above theory using 52 lines of *Drosophila melanogaster* that had been inbred for one generation of brother-sister mating and then expanded to a large population size by two generations of random mating. They estimated $\mathbf{G}$ for all 52 lines using six wing traits (overall wing area and five different angles between wing landmarks). Genetic variances and covariances were estimated by parent-offspring regression, with 1945 families used to estimate the outbred $\mathbf{G}$ and a total of 4680 families over the 52 inbred lines (an average of 90 families per line). This experiment represents a truly heroic amount of work, and while it provides support for the above theory, it also offers some very strong caveats. Figures 29.1 and 29.2 illustrate the major findings.

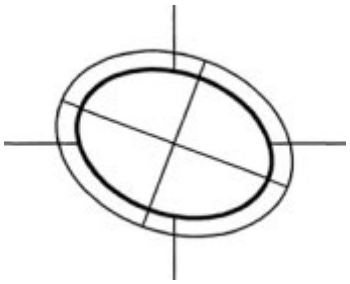


Figure 29.1. The 95% confidence ellipses (for Phillips et al.'s wing traits B versus F) using the outbred population estimate of $\mathbf{G}$ (the thin outer line) and the mean estimate of $\mathbf{G}$ over all 52 inbred lines (the thicker inner line). The axes of both ellipses are shown, and are essentially identical. However, the mean estimate of $\mathbf{G}$ over the inbred lines shows a proportional (0.7) reduction along both axes. Thus, the eigenvectors are identical, but the eigenvalues for the inbred lines are only 70% of those for the outbred lines. The other 14 two-trait combinations showed very similar results. After Phillips et al. (2001).

First, the *average* $\mathbf{G}$ matrix under inbreeding (obtained by using the average of the matrices taken over all 52 inbred lines) very nicely fits the theory. The orientation of the average of the inbred matrices conforms to the outbred control, and the eigenvalues were decreased by an amount corresponding to the $(1 - f)$ value for this mating system. Figure 29.1 shows this for two of the traits (different wing landmark angles), displaying the 95% confidence ellipses for the two covariance matrices. The outer ellipse (thin line) shows the ellipse associated with the outbred $\mathbf{G}$ while the darker inner ellipse corresponds to the *average* $\mathbf{G}$ matrix over the 52 inbred lines. Note that the orientation of both matrices is identical, and that the inner ellipse is proportionately reduced relative to that for outbreeding.

The reduction to about 0.7 is consistent with the amount of inbreeding and the expected population size. All 15 two-trait combinations showed very similar results. Further, the dispersion of means over all 52 lines also nicely follows the prediction of having covariance matrix consistent with $2f\mathbf{G}(0)$.

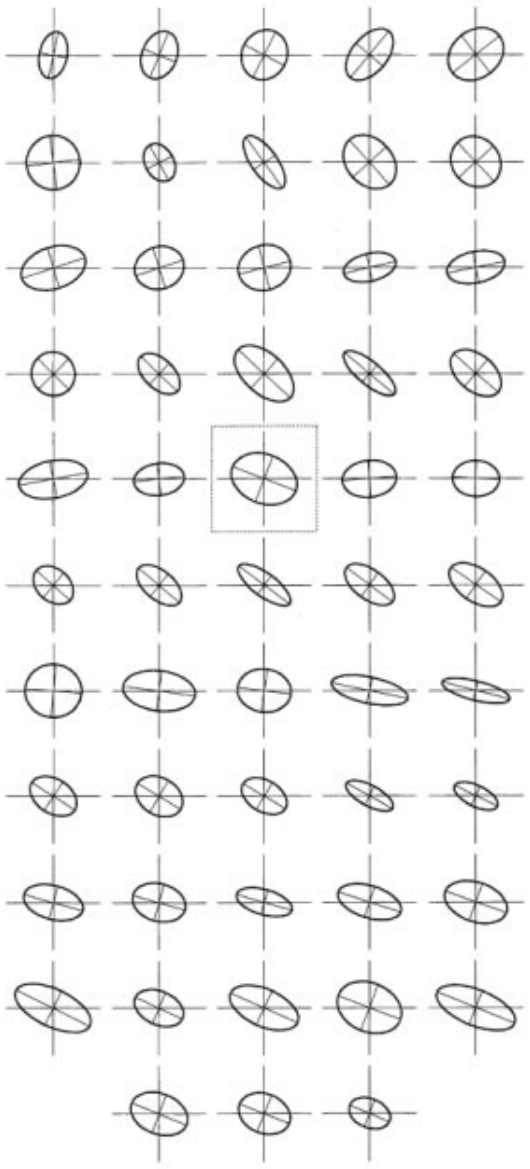


Figure 29.2. Individual realizations for **G** over the 52 inbred lines. The 95% confidence ellipse is shown for each of the estimated **G** matrices from the 52 inbred lines for wing traits B and F. The average of all matrices is given by the boxed ellipse in the middle of the figure. Note the extreme variation about the mean in both orientation and variation along axes. After Phillips et al. (2001).

However, when one considered the *realization* of **G** for a particular line, as opposed to

the *average value* of **G** over all lines, a different picture emerges. Figure 29.2 shows the 95% confidence ellipses for the same two traits compared in Figure 29.1 for the 52 estimated **G** matrices for the inbred lines. The result is quite considerable variation around this expectation, including cases where the genetic correlation changes signs.

A compact representation of this same information is given in Figure 29.3, which shows the angle and length of the principle axes for each of the 52 lines compared with that for the outbred **G** and mean **G** over all 52 lines. The tick on the horizontal axes corresponds to the variance in the outbred **G**, so that two of the lines actually showed an *increase* in the additive variance along the main axis relative to an outbred population. Note that some of the realizations have their major axis essentially orthogonal to the outbred major axis. This highlights the major point of the Phillips et al. paper, namely that "drift causes idiosyncratic changes in the variance-covariance structure of particular populations." Thus, even with near additivity (as the mean value nicely fits the additive theory), quite different **G** matrices can be generated by drift.

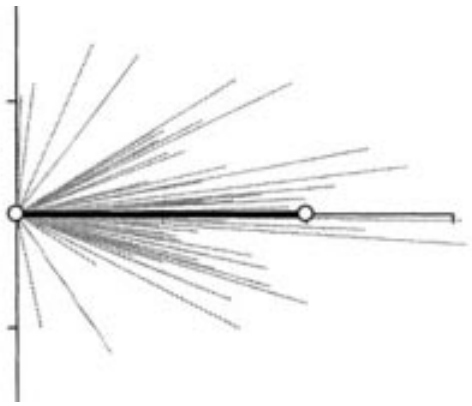


Figure 29.3. Differences in orientation and length of the major axis for each of the estimated 52 **G** matrices (for wing traits B and F). The horizontal axis sets the reference for the angle between the major axis of a replicate and that for the outbred control. The thick horizontal line ending in the open circle is the average over all 52 replicates and shows the same orientation as the control, but a smaller length (the tick on the horizontal axis corresponds to the length in the outbred control). Each thin line corresponds to the major axis for each of the 52 particular realizations, with the angle between each line and the horizontal axis representing the angle between its major axis and that for the outbred populations, while the length of the line indicates the eigenvalue for the major axis. Note that the orientations of the major axes for the individual **G** estimates for each of the 52 lines vary dramatically from the mean, as does the length of this axis. After Phillips et al. (2001).

Figure 29.4 shows how hidden pleiotropy (pleiotropic alleles present, but no *net* genetic correlation) can have a significant influence on changes in **G** under inbreeding. It shows the lengths and orientations of the major axis in each of the 52 estimates of **G** for two pairs of traits, both of which show no genetic correlation in the outbred population. One set of traits shows very little variation in orientation of the principal axis, and hence very little genetic correlations in the **G** estimated from the inbred lines. However, the second pair of traits, which also showed no genetic correlation in the outbred population, showed considerable variation in genetic correlations (both positive and negative) among the different realizations of **G** across the inbred lines, presumable reflecting a significant reservoir of pleiotropic alleles. Note that we would expect roughly equal numbers of positive and negative estimates of

genetic correlations if the *net* effect of all alleles is zero, as we would expect roughly equal numbers of samples with excess number of positive, or negative, alleles.

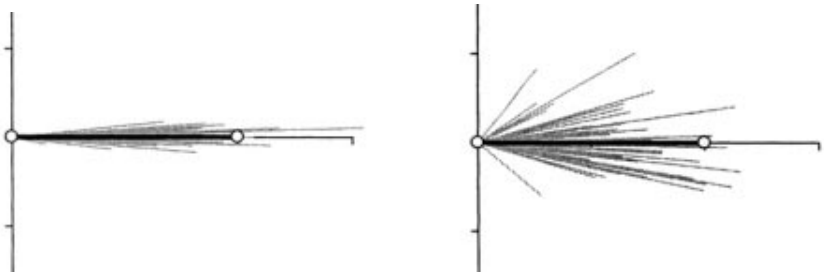


Figure 29.4. Indications of hidden pleiotropy among sets of wing traits in the Phillips et al. experiment. As in Figure 29.3, the length and orientation of the principal axis for the estimated **G** matrix from each of the 52 lines is compared with that for the outbred population. Both pairs of traits represented here showed zero genetic correlation in the outbred estimate of **G**. **Left:** For wing traits C and D there is very little variation on orientation among the estimated principal axes, although there is considerable length variation. Thus, the estimated **G** matrices all showed very little genetic correlations between these traits. **Right:** Wing traits E and F, on the other hand, show considerable variation in orientation, so that considerable genetic correlations between these traits are seen among the replicate population despite little genetic correlation observed in the outbred estimated.

Whitlock et al (2002) examined a subset of these lines after an additional 17 generations of random mating (20 generations past the inbreeding event), and found evidence that **G** has continued to change in some of these lines, although not in any way that returns it to its pre-bottleneck shape. They suggest that some of the continued change in the **G** matrices for the inbred lines was the result of decay of linkage disequilibrium generated during inbreeding. By contrast, **G** for the outbred population as also examined and showed no change.

The experimental results of Phillips et al. have very important implications for the role of drift in the dynamics of multivariate response. If indeed changes in **G** are simply proportional, then the effects of drift are simply to change the time scale of the dynamics, as the orientations (directions of genetic variation) are unchanged, and thus the trajectories would be unchanged, simply requiring longer time. Of course, as variation runs out due to drift, the population (in the absence of mutation) stops evolving. Against this simple model based on the *mean* change in **G** is the reality that any particular realization of **G** from sampling due to drift will likely show considerable deviation from $E(\mathbf{G})$. Indeed, when hidden pleiotropy is present, different realizations may have different signs for their genetic correlations even when the outbred population shows no correlation. Consider replicate populations first formed by drift that subsequently experience selection. The sampling of the founding population generates a variance in the means among replicates, starting them at different locations. It will also generate differences in their **G**, which can (in some cases) be rather dramatic. Thus, even replicate populations experiencing the same amount of selection can have rather different starting locations and subsequent evolutionary trajectories even in the simple case of a strictly additive model. As Phillips et al. correctly point out, one is "likely to underestimate the potential role of drift in explaining divergence among populations".

Changes in G When Non-additive Genetic Variance is Present

As discussed in Chapter 5, when non-additive variance is present, the additive genetic variance may actually initially *increase* under inbreeding. This can occur via dominance

when recessives are present, and it can also occur by the conversion of nonadditive into additive variance. Given the complications for individual realizations under even a simple additive model, the behavior of **G** in a single population under inbreeding when nonadditive variance is present is expected to be even more complex.

While there are no general results, an important start is provided by López-Fanjul et al. (2004), who used two-locus models (and ignored the effects of linkage) to considered the impact of inbreeding on genetic covariances when nonadditive variation is present in the ancestral base population. When only additive and additive-by-additive genetic variance is present in the base population, a simple relationship exists for the additive genetic covariance,

$$\sigma_t(A_1, A_2) = (1 - f_t)\sigma_0(A_1, A_2) + 4f_t(1 - f_t)\sigma_0(AA_1, AA_2) \tag{29.1a}$$

which relates the genetic covariance in generation t with the additive (A) and additive-by-additive (AA) genetic covariances in the base population (indicted by zero subscripts). Note that the later is simply the covariance between additive-by-additive values for two traits within an individual from the population. If both ancestral covariance components are positive, then the additive genetic covariance under inbreeding will exceed its value in the outbred population provided

$$\sigma_0(A_1, A_2) < 4(1 - f_t)\sigma_0(AA_1, AA_2) \tag{29.1b}$$

When Equation 29.1b holds, inbreeding initially increases the additive genetic covariance, with the maximum occurring at a level of inbreeding corresponding to

$$f_t = \frac{4\sigma_0(AA_1, AA_2) - \sigma_0(A_1, A_2)}{8\sigma_0(AA_1, AA_2)} \tag{29.1c}$$

Similarly, for a reserval of sign under inbreeding, the two genetic components must be of opposite sign with

$$\sigma_0(A_1, A_2) < 4f_t|\,\sigma_0(AA_1, AA_2)\,| \tag{29.1d}$$

when the additive genetic covariance is positive.

Equation 29.1a is the covariance analog for the conversion of additive-by-additive into additive variance under inbreeding (Chapter 5). Thus, the expected (additive) genetic covariance at generation t becomes

$$\mathbf{G}_t = (1 - f_t)\mathbf{G}_0 + 4f_t(1 - f_t)\mathbf{G}_{AA,0} \tag{29.2}$$

The genetic covariance matrix does not proportionately decrease with time, and individual elements may actually increase. López-Fanjul et al. also examined cases with dominance and dominance epistasis (i.e., dominance-by-dominance and additive-by-dominance terms), but no simple analytic expression emerges (although they present equations for simple numerical analysis). They make the important point that when dominance is present, there will often be inbreeding depression, so that even if genetic variances or covariances are increased by dominance or dominance-related epistasis, this is likely to be more than countered by strong inbreeding. They suggest that although occasional increases in additive variances and covariances may be observed upon inbreeding, this is unlikely to increase the rate of evolution.

The Eigenstructure of G Under Drift and Mutation

A common observation is that **G** shows a distribution of eigenvalues that is far from uniform (i.e., all having roughly equal values). Griswold et al. (2007) make the important finding that

such an eigenstructure can easily arise from drift even when the mutational matrix has a uniform distribution of eigenvalues (i.e., all eigenvalues of **M** are roughly equal). The authors used simulations and coalescent theory (Chapters 2, 5) to model the genealogical structure of alleles within a sample. They assumed a mutational matrix **M** with equal eigenvalues, and them allowed for drift and recombination when examining the estimated **G** matrix obtained using a sample of individuals from the population. While the *mean* genetic covariance matrix at equilibrium in a population of effective size N_e was $E[\mathbf{G}] = 2N_e\mathbf{M}$, which has the same eigenvalues as **M** (up to a constant), when the distribution of eigenvalues for any *particular* **G** was considered, it was highly nonuniform, showing close to exponential decay. The authors show that this skewing of the distribution of eigenvalue arises for genealogical reasons: drift imposes a dependence structure on the alleles in the sample due to shared common ancestry, and this in turn results in the distribution of the eigenvalues of **G** being highly nonuniform. While the nonuniform distribution arises in the estimated **G** matrix from a sample, the authors show that this effect still holds, albeit not as dramatically, in the population as well.

Griswold et al. showed that the nonuniformity decreases as the number of loci increases, but increases with tighter linkage among loci. The eigenvalue distribution in **G** is further skewed when **M** itself has a nonuniform distribution of eigenvalues. The more pleiotropic the mutational effects are (i.e., the more loci they influence), the stronger the departure from uniformity. They also examined the effect of the assumed mutations models on these results. Recall from Chapter 27 that the continuum-of-alleles (COA) model assigns the alleic effect of a new mutation as the sum of its current effect plus a random effect, while the house-of-cards (HOC) model creates a new effect that is independent of the current value. Both mutational models gave a highly nonuniform distribution of eigenvalues in **G**, but the COA model had a larger effect. This is not surprising, as the COA model contains some ancestry information (parent and mutational values are associated), while the HOC model has the value of its mutation independent of ancestry. The authors also found that the eigenvectors of **G** are also influenced by **M**. When **M** has a uniform distribution of eigenvalues, the angle of leading eigenvector of **G** is uniformly distributed (i.e., points in a random direction). Conversely, when **M** has an uneven distribution of eigenvalues, the leading eigenvector of **G** has a direction centered around the direction of the leading eigenvector of **M**.

Finally, the sample size used can influence the estimated dimension of **G**. Assuming a mutational matrix **M** of dimension 25, only 12 and 15 significant principle components for **G** were found when the sample size was 75 and 300, respectively. Thus, the dependence among some alleles imposed by the genealogy influences the observed dimensions of **G**. It is important to point out that Griswold et al. assumed no environmental variation, so that the estimate of **G** for a sample is exact, with no error in estimating **G** given the sample. One would imagine that the effect of sampling exaggerates most (if not all) of the above effects.

COMPARING COVARIANCE MATRICES: METHODOLOGY

There is a rich, often confusing, and occasionally contradictory, literature on methods for comparing **G** matrices. This reflects both a deep interest in such comparisons and the difficulty of the problem. Indeed, even a simple matrix is still a complex structure. If we rule out two matrices being identical, they may still share much in common. Much of literature reflects this tension of trying to obtain a simple metric that captures important biological information. We have taken a historical approach, starting with the initial methods and moving through more recent ones. A main take-home point is that while much progress has been made, comparison of **G** matrices still has significant unresolved problems. Short reviews of some of the methods are given by Roff (1997, 2000) and Steppan et al. (2002), although our following treatment is much more extensive.

General Issues of Inference on G Using a Population Sample

Most matrix-comparison procedures in the statistical literature were developed for **product-moment** estimates. The phenotypic covariance matrix is such an example, as its elements are (generally) computed directly from the data, e.g.

$$\widehat{P}_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ji} - \overline{x}_j)(x_{ki} - \overline{x}_k)$$

Conversely, the elements of **G** are typically obtained as **variance-component** estimates, meaning that we cannot observe the values of interest directly, but rather they must be inferred (for example, by ANOVA or REML). While we observe the phenotypes of sampled individuals directly, we cannot observed their vector of breeding values, but rather must estimate them (for example, by parent-offspring regression, or half- or full-sibs, see LW Chapter 17). Working with such an indirect estimator has two issues. First, the precision of product-moment estimates is much higher than variance-component estimators. Second, product-moment estimators (usually) ensure that the covariance matrix is semipositive-definite (contains no negative eigenvalues). Such is not the case when the elements of the covariance matrix are estimated through variance components procedures. Indeed, Hill and Thompson (1978) and Bhargave and Disch (1982) show that the probability of an estimated **G** matrix containing a negative eigenvalue can easily approach one, even with large samples. Further, the sampling properties of matrix comparison statistics using **G** have not been formally developed, as these depend on the details of the design used to estimate **G** in addition to any particular features about the choosen statistic.

The fact that **G** must be estimated using family information (or more generally, information from other sets of relatives) has two important implications for power and statistical inference. First, power is essentially set by the number of *families*, not the total number of individuals. One way to see this is to consider a situation where the mean offspring values of a sire are used to estimate his vector of breeding values, and the collection of such vectors over all sires is used to estimate **G**. Clearly, while adding more offspring increases the precision of the estimate of breeding value, the critical issue here is the number of sires, as these determine the number of independent values used to estimate **G**. The same general logic holds when other family designs are used.

Given that distribution theory for most of the proposed test statistics for matrices has either not been worked out, or only resolved under very strict assumptions, resampling methods are widely used. Hypothesis testing is typically performed using randomization methods to obtain p values, while bootstrap resampling is often used to obtain approximate standard errors and confidence intervals. The key with both these methods is the independent sampling unit, which is (usually) a family. For example, suppose n_1 and n_2 families are used to estimate the **G** matrices for two populations. A randomization test proceeds by choosing n_1 families at random (i.e., without regard to population) from the $n_1 + n_2$ collection of total families and assigns them to group 1, with the rest assigned to group 2. **G** is then estimated for both groups, and a test statistic generated under this null model of no association between **G** and group membership. Performing this (say) 5,000 to 10,000 times generates a distribution of the test statistic under the null hypothesis, and this is compared with the actual observed value. Specifically, suppose N randomized samples are generated, n of which have a test statistic as extreme (or more so) that observed in the original sample. The corresponding p value is given by

$$p = \frac{n+1}{N+1} \tag{29.3}$$

where, to be conservative, the extra one is added to the numerator and denominator to account for the original observation (Manly 1991). Note th4 nature of the null hypothesis in

this case: group membership has no impact of **G**, and hence the null is that **G** is the same for both groups.

Returning to the sampling unit, families are the independent data points (provided that no *related* families have been included). If one is using a nested full-sib/half-sib design (say with a common sire), then the level of sampling is at the sire level, i.e., the family unit is the collection of full and half-sibs from that sire (this assumes that dams are unrelated). If different families are related, then resampling cannot be used as the families are not indepedent. A further consideration are other potentially confounding effects in the experimental design. For example, due to the size of such experiment, it is fairly common to perform these in **blocks**, so that families for the two groups are raised in common batches, and hence common environmental effects. Since there can be changes between blocks (batches), resampling should occur within, and not between, blocks.

The other common resampling approach, the bootstrap, also has the family as the independent sampling unit when considering statistics on **G**. A bootstrap sample for a covariance matrix estimated from n_1 families is generating by sampling *with replacement* from the original families until a sample of size n_1 is generated. A number of such bootstrap samples would then be generated and the variance among these samples provides an approximation for the sampling variance, while the spread of these samples provides an estimate of an approximate confidence interval. For example, Spitze et al (1991) used a number of fairly nonstandard statistics (e.g., difference in leading eigenvalues, difference between determinants) to compare two *Daphnia* covariance matrices, using bootstrapping for hypothesis testing by constructing bootstrap confidence intervals. Jackknife approaches are also occasionally used (e.g., Brodie 1993, Holloway et al. 1993, Roff 2002), again the sampling unit (the items to be deleted one at a time) are family units. For a matrix of n families, n delete-one-family estimates are generated (a family is deleted, **G** estimated and the appropriate statistic computed) and the variance of these estimates about their mean can be used to estimate an approximate standard error. While both the bootstrap and jackknife approaches are potentially very powerful, they do not automatically work in all conditions, and thus some must be taken in their use (Miller 1974, Manly 1991). On the other hand, provided family units are independent (unrelated), p values obtained under randomization are usually fairly bulletproof, *provided* that care is taken to identity the appropriate unit for randomization.

Identity, Proportionality, Common Orientation, Common Scaling

What exactly are we comparing between two matrices? The simplest comparison, testing each element for identity between the two matrices (e.g., are the variance estimates for trait one significantly different between populations, what about the covariance estimates for traits 1 and 2, and so on through the rest of the elements) is fraught with problems for a number of reasons. An obvious one is multiple comparisons — when compairing two 10×10 matrices, there are $(10 \cdot 9)/2 = 45$ different tests, and thus we would expect at least two to be significant at the 5% level simply by chance. A more subtle issue is that since we are estimating **G** indirectly, the estimates for different elements of **G** are generally correlated and thus not independent. Third, performing all pairwise comparisons as separate tests seems to be wasting information. For example, for a 2 x 2 to be a proper covariance matrix we must have $G_{12}^2 \leq G_{11}G_{22}$. An element-by-element comparison does not use this information. Finally, what can we say if one (or more) of the pairs of elements are statistically different from each other? The matrices may still have very much in common. How do we quantity this?

The key is that a matrix is fundamentally a *geometric object*, and it is similarity of geometry that is of interest to us. Since the eigenstructure of a matrix defines its geometry, at the heart of matrix comparison procedures should be a comparison of some feature(s) of their eigenstructure (Figure 29.5). Although this concept was recognized even in some of the initial

papers on comparison of **G** (Lofsvold 1986, Cowley and Atchley 1992) it took time to be fully implemented. Broadly speaking, we can consider four different types of matrix similarity:

• **Identity**: The eigenvalues and their corresponding eigenvectors are the same for both matrices: $\lambda_i(1) = \lambda_i(2), \mathbf{e}_i(1) = \mathbf{e}_i(2)$ for $1 \leq i \leq n$.

• **Proportionality**: For some (positive) constant c, $\lambda_i(1) = c\lambda_i(2), \mathbf{e}_i(1) = \mathbf{e}_i(2)$ for $1 \leq 2 \leq n$. The eigenvalues all differ by a fixed constant (a proportional change), while the orientation (the eigenvectors) of the two matrices is unchanged. We have already seen that, *on average* drift (when only additive variance is present) is expected to produce proportional changes in **G**, albeit with a very high sampling variance. If nonadditive variance is present, drift is *not* guaranteed to produce proportional differences.

• **Common Orientation, Different Scaling**: The eigenvectors of both matrices are identical, but the eigenvalues may differ. Thus, the ordering of the eigenvectors (which is given by the ordering of their associated eigenvalues) may change over matrices, but the set of axes is the same, just scaled differently. Such matrices show the same independent combinations of traits (given by the eigenvectors that specify the axes), but the genetic variances (the eigenvalues) associated with these sets of traits differs.

• **Common Scaling, Different Orientation**: The two matrices have the same set of eigenvalues $\lambda_i(1) = \lambda_i(2)$ for $1 \leq i \leq n$, but not the same set of eigenvectors (their orientations differ). Both matrices have the same distribution of variances over their independent trait combinations, but those trait combinations (eigenvectors) vary over matrices.

Of course, the most realistic situation is where matrices share *some* common eigenstructure (for example the first couple of eigenvalues and eigenvectors), but not all.

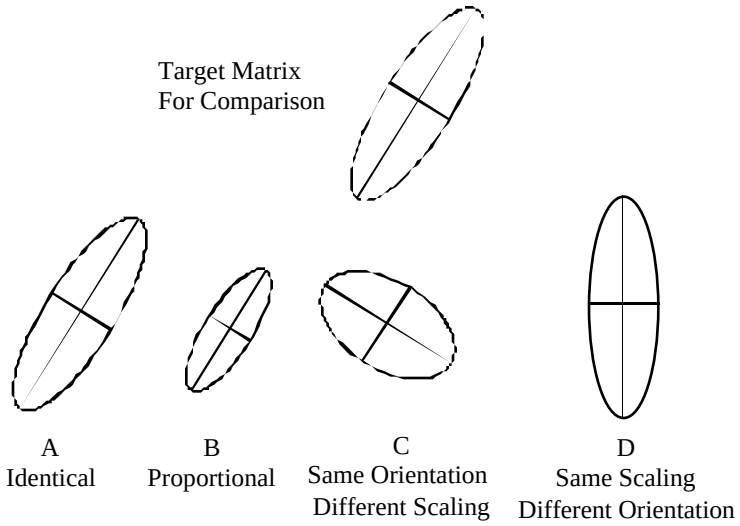


Figure 31.11. Comparison of matrices is best done through comparison of their *geometry* (i.e., their eigenstructure). Matrix A and the target matrix are identical, having the exact same eigenstructure. Matrix B is proportional to the target, having the same eigenvectors (orientation), and the same set of eigenvalues, just scaled by a constant. Matrix C has the same eigenvectors (compare the principal axes), but different eigenvalues. This can result in a dramatically different appearance between the two matrices. Matrix D has the same eigenvalues (the lengths of the principal axes are the same for both matrices), but the orientation of those axes differs. More generally, for two n-dimensional matrices, some subset of their geometry may be in common, for example, they may share the first k principal components.

Element-by-element Tests

The first attempts to compare **G** between populations were by Arnold (1981) and Atchley et al. (1981). Arnold examined the estimated **G** matrices for a series of chemoreceptive traits between two populations of garter snakes (*Thamnophis elegans*), noting that heritabilities were not significantly different across populations and the covariances appeared to be similar (although no formal tests were applied). While Arnold's comparison was within a species, Atchley et a. (1981) compared **G** matrices for common skull features across species (rat and mouse). Again, no formal statistical tests were applied. These papers set the tone for more formal comparisons, which were initially based on element-by-element comparisons. Any number of tests can be used to test significance between a pair of elements. For example, Carr and Fenster (1994) performed pair-wise comparison of individual heritiabilities and genetic correlations using t-tests based on large-sample approximations for standard errors (i.e., assumptions of normality). Brodie (1993) used t-tests with a (delete-one-family) jackknife approach to generate their approximate standard errors. He found no heterogeneity of estimates between populations for a suite of antipredator traits in the garter snake *Thamnophis ordinoides*. Shaw and Billington (1991) used REML to compare individual variance terms for growth and reproduction traits in two populations of the grass *Holcus lanatus*, finding no significant differences.

One obvious issue with element-by-element tests is multiple comparisons. Using the standard Bonferonii correction (setting the critical value for each comparison as α/n in order to achieve an α-level p value over n comparisons) is both too stringent and likely not appropriate given that tests are likely to be correlated. While sequential Bonferonii approaches such as the methods of Holm (Holm 1979), Simes-Hochberg (Simes 1986, Hochberg 1988), or Hommel (Hommel 1989) can be used to (somewhat) alleviate the stringency, the problem of lack of independence still remains. A more robust approach would be to use the binomial test to see if the number of significant tests exceeds that expected by chance.

Example 29.1. Paulsen (1996) compared elements between two estimated **G** matrices for wing pattern traits in two sibling species (*Precis coenia* and *P. evarete*) of buckeye butterflies, using bootstrap resampling (with the family being the unit of sampling) to generate confidence intervals. She found 9 of 45 comparisons of heritability to be significantly different at the 5% level. Is this number significantly greater than expected from binomial sampling? Here $n = 45$, and $p = 0.05$, for an expected number of 2 significant tests. The probability of seeing 9 (or more) significant tests (under the null of no difference) is given by the binomial with success probability $p = 0.05$ and sample size $n = 45$,

$$\sum_{i=9}^{45} \Pr(k \text{ significant tests}) = \sum_{k=9}^{45} \frac{45!}{k!(45-k)!}\, 0.05^k\, 0.95^{45-k} = 0.0003$$

Thus, among the heritabilities, there clearly are an excess of significant differences. There were 990 pairwise comparison involving correlations, of which 41 where significant at the 5% level. The expected number under the binomial is 49.5. Thus, there is no support for an excess of significant tests involving correlations.

When considering all of the elements together, 50 of 1035 tests were significant, while under binomial sampling by chance we expect 52 to be declared significant. Thus, if the focus is entirely on the heritabilities, we can declare a significant difference, while if the focus is on the entire matrix, we cannot do so. However, it is worth mentioning that heritabilities are estimated with greater precision that correlations, and hence with greater power. Thus, the lack of an excess of significant results among the correlation comparisons may simply be a reflection of lack of power, not lack of biological differences.

As mentioned, the results of these test can be difficult to interpret. First, low power coupled with the multiple-comparisons issue make a non-significant result (the null here is that the matrices are identical) not very impressive. Conversely, a significant difference in one (or more) elements is difficult to interpret, as we must translate this into changes in the eigenstructure to fully understand its biological implications.

Example 29.2 An interesting variant of element-by-element tests was used by Roff and Mousseau (1999) who examined femur and ovipositor length in two species of crickets. They had estimates of $\mathbf{G}$ from eight populations of *Allonemobius socius* and from one population of *A. fasciatus*. Two different element-by-element comparisons were used to test for element heterogeneity. The first was simply to compute a variance over all nine populations for a specific estimate. Suppose x_i denotes the estimated genetic covariance between the two traits in population i, then

$$Var(x) = \frac{1}{8}\sum_{i=1}^{9}(x_i - \overline{x}\,)^2$$

The observed value of this statistic for all three elements in $\mathbf{G}$ was compared to the distributions of values generated by randomizing families over the nine populations. Significant heterogeneity was found for all three estimates (the two variances and the covariance). Heterogeneity could arise because of variation in the estimates within *socius* and/or variation in estimates between species. For a test of species differences, the authors constructed a simple t test where the standard deviation was estimated from the empirical values for the eight *socius* samples. Again, let x denote the genetic parameter of interest, with the estimated value for the different species (and population for the *socius*) samples indicated by subscripts

$$t = \frac{x_{fas} - \overline{x}_{soc}}{SD_{soc}\sqrt{9/8}}, \quad \text{where} \quad \overline{x}_{soc} = \frac{1}{8}\sum_{i=1}^{8} x_{soc,i}, \quad SD^2_{soc} = \frac{1}{7}\sum_{i=1}^{n}(x_{soc,i} - \overline{x}_{soc})^2$$

Again, significant between-population differences were found for all three components of $\mathbf{G}$.

Roff's Jackknife MANOVA Approach

A final variant of element-by-element tests was suggested by Roff (2002) and Bégin et al. (2004), who placed this problem in MANOVA (**multivariate analysis of variance**) framework. Recall that MANOVA is simply the multivariate extension of ANOVA to test whether a vector of means changes over treatments. Normally, one has (say) $n_1, \cdots n_k$ vectors each of length m for treatments one to k, and (under appropriate normality assumptions) a number of tests for equality of the mean vectors over the treatments have been proposed (Tabachnick and Fidell 2006). Clearly we can consider all $n(n+1)/2$ elements in a covariance matrix to be the vector of means to be contrasted across treatments, but how do we generate the vectors of observations for each treatment in order to use standard MANOVA? Roff offered the nifty solution of using jackknife pseudovalues to generate the required vectors.

Recall that a pseudovalue ϕ from a jackknife resampling scheme is a measure of the variation about the true value and is obtained as follows. Suppose there are n_1 families used to estimate $\mathbf{G}_1$. Focus on a particular element of $\mathbf{G}_1$, say the covariance between traits two and three, and denote the estimate for this value using all of the families as $\widehat{\theta}$. Likewise,

denote the estimate of this value when family i has been removed as $\widehat{\theta}_{-i}$. The jackknife pseudovalue ϕ_i associated with family i for this estimate is given by

$$\phi_i = n\widehat{\theta} - (n-1)\widehat{\theta}_{-i}$$

as obtained by Tukey (1958), also see Miller (1974) and Manly (1991). For the n_1 vectors associated with treatment one, Roff proposed using the n_1 vectors of pseudovalues for estimated genetic covariance elements (one vector for each family). The n_2 vectors for treatment 2 follow similarly. Note that this method easily allows comparison of k treatments. Further, it is easy to implement using standard statistical software, and that the multiple corrections problem is accounted for within the MANOVA single test statistic. Standard MANOVA test statistics (e.g., Wilk's λ, Lawley-Hotelling and Pilai-Bartlett traces) rely heavily of multivariate normality assumptions (Tabachnick and Fidell 2006). While small-scale simulations appear to show that the jackknife returns the correct standard errors for heritabilities (Simons and Roff 1994) and genetic correlations (Roff and Preziosi 1994), the issue of normality was not addressed, and indeed application of this method often generates a non-normal distribution of pseudovalues (Bégin and Roff 2003). If this is a concern, again a resampling method can be used to generate a p value. As before, families are randomly assigned to groups, and then entire procedure of generating jackknife pseudovalues for each of these new family assignments is done, and an approximate MANOVA summary statistic computed. This is redone several thousand times to generate a distribution under the null of matrix equality across treatments. While computationally intense, this is straightforward.

Roff's T test

Willis et al. (1991) suggested that one test for a measure of association between two matrices is the sum of the absolute differences of their elements, while Spitze et al. (1991) considered the sum of squared differences and used bootstrapping for confidence intervals and hypothesis testing. Roff et al. (1999) and Bégin and Roff (2001) implemented this class of comparisons using randomization tests. Roff's T statistic is given by

$$T = \sum_{i \le j} E_{ij}, \quad \text{where} \quad E_{ij} = |\, G_{ij}(1) - G_{ij}(2)\,| \tag{29.4a}$$

Hence T is simply the sum of the absolute differences over all the unique elements of **G**. A related metric was also considered by Steppan (1997b), and Cheverud and Marroig (2007) note that T can easily be extended to comparisons of three (or more) populations by computing all pairwise combinations. The observed value of T is then contrasted with the observed distribution of T values under the null hypothesis (both matrices are identical) generated by randomization. If a significant T value is found, the impact of individual differences in elements can be ascertained by comparing the observed E_{ij} values with their distribution in the randomized sample. As a standardization for comparing results from different experiments, Bégin and Roff (2001) proposed a scaled version of T that yields the absolute difference between elements as a percentage of the average size of the matrix elements,

$$T\% = \frac{T/m}{[\,\overline{G}_{ij}(1) + \overline{G}_{ij}(2)\,]/2}, \quad \text{where} \quad \overline{G}_{ij}(k) = \frac{2}{n(n+1)} \sum_{i \le j} G_{ij}(k), \tag{29.4b}$$

Here $m = n(n+1)/2$ is the total number of unique comparisons. Note that $\overline{G}_{ij}(k)$ is just is the average size of an element from matrix k. Bégin and Roff (2001) did not specify if this is the average of all unique elements or the average of the absolute values of these elements. The later seems the more appropriate metric.

Mantel's Test and Other Matrix Correlation Approaches

The early literature of comparisons of **G** contains numerous references to **matrix correlation** tests, which examine the association between the corresponding elements of two matrices (Lofsvold 1986, Kohn and Atchley 1988, Fong 1989, Shaw 1992, Cowley and Atchley 1992, Steppan 1997a). The basic idea of this class of tests is straightforward (reviewed Hubert 1983 and Dietz 1983): corresponding elements in the two matrices are treated as paired observations, some measure of association is generated, and some procedure is used to obtain measures of association under the null. Standard correlation coefficients such as Pearson's product-moment and Spearman's rank have been used as the measure of association, as have other similar measures (such as Dietz's 1983 K_c statistic). The significance of the test statistic is assessed by randomly permuting the rows and then rearranging the elements within a row to recovery symmetry. As a toy example, consider

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 1 \end{pmatrix}$$

The three rows are picked in random order (say 2,3,1) and then the columns adjusted to recover a symmetric matrix,

$$\mathbf{A}_r^* = \begin{pmatrix} 2 & 1 & 4 \\ 3 & 4 & 1 \\ 1 & 2 & 3 \end{pmatrix}, \quad \text{recovering symmetry gives} \quad \mathbf{A}_r = \begin{pmatrix} 1 & 4 & 2 \\ 4 & 1 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

Note that there are only $n!$ possible permutations (and hence only six for our toy example). Such a permutation procedure generates a null where the matrix elements are independent of position, and hence the two matrices being compared are *independent of each other*. This general approach often goes under the name of the **Mantel test** (Mantel 1967), although originally this test was restricted to the case where the Pearson correlation was used for association.

While the motivation for using Mantel-type tests was admirable (trying to generate a single test statistic based on comparing the entire matrix as a single object), they have a number of fatal flaws when applied to estimated covariance matrices. First, there is the issue of whether the appropriate null is that two matrices are dissimilar (versus the null that they are identical), which has been spiritedly debated (Cowley and Atchley 1992, Shaw 1992, Marroig and Cheverud 2001). To us, it seems that the most natural null for populations that have shared a common ancestor is that the matrices are similar. This issue aside, the fatal problem is that the various elements along a row are *not* exchangeable, as each column corresponds to a trait that may be measured on a very different scale from the others (Turelli 1988, Cheverud 1996, Cheverud and Marroig 2007). Use of correlations or other scalings (e.g., Goodnight and Schwartz 1997) does not necessarily mitigate this issue, as the different elements may be estimated with very different levels of precision. A further issue is that the size of the matrix sets the total number of possible permutation (at $n!$), and thus the number of traits must be reasonably large ($n \geq 7$) to generate a reasonable empirical distribution under the null. Finally, the measure of association used can infleunce the outcome (Cheverud et al. 1989, Cowley and Atchley 1992). For example, Pearson's correlation is much more senstive to skewness than Spearman's or K_c. Lofsvold (1986) found no evidence for association between two species of *Peromyscus* mice using Pearson, while a reanalysis of the same data by Kohn and Atchley (1988) using Dietz's K_c statistic did.

Given that these issues largely concern the permutation procedure, what if we access significance using an appropriate randomization approach? Here, *families* are randomly

assigned to the two groups being compared, an association statistic computed, and this procedure repeated to generate a distribution for the test statistic under the null hypothesis. In this case, since the group membership is randomized, the null hypothesis is that the matrices for both groups are *identical* (Goodnight and Schwartz 1997). Note that is completely opposite of the null under element (as opposed to family) randomization.

Regression Methods: Tests of Proportionality

Closely related to correlation-based tests are those based on regressing the values of the elements from the two matrices on each other (Carr and Fenster 1994, Roff et al. 1999, Bégan and Roff 2001). As with correlation analysis, each corresponding element in the two matrices being compared is treated as a paired data point, e.g., $[G_{ij}(1), G_{ij}(2)]$, and a regression line is fit through the data. Matrix similarity is indicated by a slope of one, while good fit to a slope different from one is consistent with the two matrices being proportional. Since both elements are measured with error, standard regression is not appropriate, rather major axes or similar methods of regression are used.

Part of the motivation for this approach is a test for changes due to drift. While we have seen that under strict drift, the *expected* **G** is proportionately reduced, the variation about this expected value is enormous, casting some uncertainly on its usefulness. Further, when nonadditive variance is present, proportional changes are *not* expected. However, with these caveats in mind, Roff and colleagues (building on his T statistic, Equation 29.4) have proposed a nested series of tests to look for signatures of drift (Roff et al. 1999, Roff 2000, Bégan and Roff 2001). Consider three sums of squares. The first is just a T-like statistic with square differences replacing absolute differences.

$$\mathrm{SS}_1 = \sum_{i \le j} \big(G_{ij}(1) - G_{ij}(2)\big)^2 \tag{29.5a}$$

The second is based on the notion that under drift, elements should be proportionally equal, i.e. $G_{ij}(1) = bG_{ij}(2)$, leading the sum of squares

$$\mathrm{SS}_2 = \frac{1}{2}\sum_{i \le j} \big(G_{ij}(1) - bG_{ij}(2)\big)^2 + \frac{1}{2}\sum_{i \le j} \big(G_{ij}(2) - (1/b)G_{ij}(1)\big)^2 \tag{29.5b}$$

The two terms represent the reduced major axis regression (both forced through the origin) of elements of one on two and elements of two on one (as these regressions can differ). If the two matrixes are sufficiently similar, the reduction in the sums of squares going from model one to model two will be nonsignificant (again, we can test this using randomization of the families to generate null distribution). However, if the reduction in the sums of squares is significant when going to model two, there is support for proportionality. If the elements are truly proportional, then the best-fitting line should have no intercept. Motivated by this, Roff's final proposed sum of square is

$$\mathrm{SS}_3 = \frac{1}{2}\sum_{i \le j} \big(G_{ij}(1) - [a + bG_{ij}(2)]\big)^2 + \frac{1}{2}\sum_{i \le j} \big(G_{ij}(2) - [A + BG_{ij}(1)]\big)^2 \tag{29.5c}$$

where a and b are the intercept and slope of the regression of matrix one on matrix two, and A and B the corresponding values for the complementary regression. No assumption is made about how these regression parameters are related to each other (unlike for model two where $b = 1/B$). Roff suggests that if model two shows an improved fit, but no significant additional improvement is offered by model three, this can be taken as evidence for proportionately. However, if the fit is further improved by model three, the changes between the two **G**

matrices are not due to drift alone. This later statement appears to be not well supported, given the behavior for samples of $\mathbf{G}$ under drift. Despite this concern, this approach may prove useful for certain questions.

Likelihood-based Tests Assuming Multivariate Normality: Variance Components

Several tests for matrix equality avail themselves when we can assume that the distribution of breeding values is multivariate normal, allowing us to use the machinery of maximum likelihood estimation and hypothesis testing (LW Appendix 3).

Shaw (1991) proposed a likelihood-based comparison test based on estimated variance components. For example, with (say) four traits, there are four estimated variances and six estimated covariances required to fill out $\mathbf{G}$. The idea is to use REML methodology to estimate the individual (co)variance components, using all of the information about relatedness among individuals in the sample. Since this approach is based on the full animal model (Chapter 16, LW Chapter 27), it allows general pedigrees to be used (as opposed to specific designs such as half or full sibs). As such, it fully accounts for all (known) relationships among different families. Further, it easily extends to comparing more than two populations (e.g., Service 2000). It should be stressed that if a REML procedure is used that estimates each of these components separately, there is no guarantee they will be consistent (i.e., yield a non-negative definite, and hence a proper, covariance matrix). If the resulting $\mathbf{G}$ matrix is not a proper covariance matrix, it is often adjusted to be so, for example by rounding negative variance estimates to zero and also by "bending" the estimated matrix (Chapter 33) to ensure no negative eigenvalues. This is accomplished (Equation 33.33) by adding a bending factor γ which is increased until the eigenvalues are all non-negative (or strictly positive in some cases), where the bent $\mathbf{G}$ is given by $\mathbf{G}_b = (1-\gamma)\mathbf{G} + \gamma\overline{\lambda}\mathbf{I}$, where $\overline{\lambda}$ is the mean of the eigenvalues.

Shaw's method, which is essentially an element-by-element approach, proceeds as follows. First, a model is fitted assuming the same variance components for both populations, and then a subsequent model is fitted where the components of $\mathbf{G}$ are allowed to be population-specific. The ratio of these two likelihoods follows (under some significant assumptions) a chi-square distribution whose degrees of freedom is the number of extra parameters between the models, providing the basis for the test. Because of the use of REML estimates (and hence the full relationship matrix $\mathbf{A}$ among all individuals in the sample), this procedure can be computationally quite demanding, and indeed convergence may not occur for certain components if the sample size is sufficiently small (Shaw and Billington 1991). This method allows for specific tests such as proportionality, as one could assign all variance components in the second population a value of c times their values in the first population, and perform a LR test for c significantly different from one.

A major caveat to keep in mind when applying this (or any other matrix test) is *power*, which is in essence a function of the number of families. The number of families in each group should be in the hundreds to have any significant power (Shaw 1991). Even in such cases, the likelihood surface for variance components can be rather flat and thus require very large differences for significance. If sample size is sufficiently small, the lack of significant curvature of the likelihood surface may cause lack of convergence of the algorithm. Since the null hypothesis here is matrix identity, this test defaults to the matrices being equal when the sample sizes are small. When families are related, the sample size for power lies between the number of families and the number of individuals, in part depending on the strength of relatedness between different families. A final issue is that the likelihood ratio test is a *large-sample* approximation, and how large is large is very unclear. There are also issues of the validity of this approximation when estimates lie on (or are forced to) the boundaries of the parameter space (i.e. a zero or negative variance estimator).

Despite these concerns, the ability to frame matrix comparisons in a rigorous hypothesis-

testing framework made this method fairly popular (e.g., Billington et al. 1998, Shaw and Billington 1991, Platenkamp and Shaw 1992, Podolsky et al. 1997, Service 2000). In theory, one could avoid many of the assumptions by assigning significance values through randomization (assigning pedigrees at random to the two groups, then computing the test statistic under this null), but the computational demands for even a single run to estimate values often make this impractical.

Likelihood-based Tests Assuming Multivariate Normality: Bartlett's Modified Test

While Shaw's approach is quite elegant, its high computational costs are only worth being spend when a number of the families in the sample are fairly closely related. In this case, its full use of the relationship matrix **A** makes it a powerful method, provided we are willing to live with the assumptions, and have very large samples to ensure power. With smaller samples, this method often fails to converge. Further, given the high computational costs for even a single estimation run, resampling methods cannot easily be applied.

Paulsen (1996) faced these issues when comparing wing pattern traits in buckeye butterflies (Example 29.1). Her solution was to use another likelihood-based method that assumed multivariate normality, but which had much lower computational costs so that resampling approaches could be used. She chose **Bartlett's modified likelihood ratio test**, an extension of a univariate homogeneity of variance test to homogeneity of covariance matrices. One advantage of this method is that it, like Shaw's method, also allows multiple matrices to be compared. The structure of the test is as follows: Suppose $\mathbf{G}_1, \cdots, \mathbf{G}_m$ are $p \times p$ sample covariance matrices, where n_i is the sample size (number of families) for sample matrix i. Compute the weighted average matrix over all samples by

$$\overline{\mathbf{G}} = \sum_{i=1}^{m} \frac{n_i - 1}{N - m} \mathbf{G}_i, \quad \text{where} \quad N = \sum_{i=1}^{m} n_i \tag{29.6a}$$

Bartlett's modified likelihood ratio test statistic Λ is based on comparing the determinants $|\,\mathbf{G}_i\,|$ of each sampled matrix to the determinate $|\,\overline{\mathbf{G}}\,|$ of a weighted-average matrix (based on sample size),

$$\Lambda = 2\left(\frac{N-m}{2}\ln(|\overline{\mathbf{G}}|) - \sum_{i=1}^{m}\frac{n_i - 1}{2}\ln(|\mathbf{G}_i|)\right) \tag{29.6b}$$

Recalling that the determinant of a matrix is the product of it eigenvalues, we see three immediate features. First, when zero eigenvalues are present, the method cannot be used. Second, the determinate does provide a metric of the marix geometry. Third, different covariance matrices (i.e., containing very different eigenstructures) can still have the same determinant. If the estimated breeding values in the sample follow a multivariate normal distribution, then for large samples Λ follows a χ^2 with $p(p+1)(m-1)/2$ degrees of freedom. The first application of this test to the comparison of genetic covariance matrices appears to be Holloway et al. (1993), who mistakenly referred to it as a Mahalanobis D^2 test (which is a test for equality of multivariate means).

Compairing this test statistic with a critical χ^2 value is, however, not recommended. For starters, the test quite sensitive to departures from normality (Zhang and Boos 1992, 1993). Further, Equation 29.6b assumes a product-moment estimate, while breeding values are not observed directly, but rather estimated to construct **G**. This adds an additional level of error (confidence limits for variances estimated from ANOVA are much larger than those when variances can be estimated directly) and likely generates further depature from normality. Finally, sample covariance matrices are often adjusted to be positive-definite and this introduces additional depatures from the assumptions.

Paulsen's solution was to estimate the null distribution (equal matrices) of the test statistic through bootstrapping the two covariance matrices (this approach was independently suggested by Goodnight and Schwartz 1997). While the bootstrap is a powerful approach, it is not distribution-free (Manly 1991), and thus requires some justification for use. Zhang and Boos (1992, 1993) showed that the large-sample distribution of Λ under the null hypothesis can be generated by using a bootstrap with randomization: All families are placed into a single pool, and n_1 families are drawn (*with replacement*) for construction of $\mathbf{G}_1$, and this procedure continued to create a bootstrap sample for all m sample covariance matrices. A Λ value is then generated, and this procedure performed several thousand items to generate an appropriate null distribution. An even simpler (and statistically more robust) procedure is to just randomize without bootstrapping — simply assign families at random (without replacement) over the m matrices. If sample covariance matrices with negative eigenvalues are adjusted for the original sample, then the same procedures should also be used adjust such matrices in the randomization/bootstrap sample.

Random Skewers: Probing the Geometry of G With Responses to Selection Response

Which matrix comparison method one chooses is greatly influenced by the question(s) being asked. From an evolutionary standpoint, one of the most interesting questiosn is whether two different covariance matrices have similar *evolutionary potential*. That is, if one picks a random direction for selection, do both populations (given their $\mathbf{G}$ matrices) respond in similar ways? **Random skewers** offer one method to probe this question. Here, 500 to 1000 (or more) random $\boldsymbol{\beta}$ vectors (of unit length) are generated, and these are projected onto both $\mathbf{G}$ matrices to generate expected vectors of response ($\mathbf{R}_1 = \mathbf{G}_1\boldsymbol{\beta}$, $\mathbf{R}_2 = \mathbf{G}_2\boldsymbol{\beta}$). Comparisons are them made among this paired set of response vectors. This approach was first introduced (into community ecology) by Pielou (1984), and informally discussed by Willis et al. (1991) as a way of comparing genotypic and phenotypic covariance matrices. Its formal use to compare genetic covariance matrices starts with Cheverud (1996), Marroig and Cheverud (2001), and Cheverud and Marroig (2007), who considered the distribution of angular differences among the paired sets of response vectors. Note that a random $\boldsymbol{\beta}$ is easily generated by generating a uniform (0,1) random variable for each element, randomly (50/50) assigning it a sign, and then normalizing the final vector of elements to one.

Recalling Equation A4.4, the angle θ between the pair of response vectors for a particular $\boldsymbol{\beta}$ is given by

$$\cos(\theta) = \frac{\mathbf{R}_1^T\mathbf{R}_2}{||\mathbf{R}_1||\ ||\mathbf{R}_2||} = \frac{\boldsymbol{\beta}^T\mathbf{G}_1\mathbf{G}_2\boldsymbol{\beta}}{||\mathbf{G}_1\boldsymbol{\beta}||\ ||\mathbf{G}_2\boldsymbol{\beta}||} \tag{29.7}$$

One then compares some appropriate test statistic for the observed θ and compares this to a critical value under either the null of no association (random θ) or the null of complete association ($\theta = 0$). Note that the correlation coefficient between these two vectors is exactly given by Equation 29.7, as $r = \cos(\theta)$, and so one can also frame random skewers in terms of the average correlation among response vectors.

Considering the distribution of angular difference among random skewers is a powerful approach for examining the similarity of potential evolutionary trajectories. Note that either identical or proportional matrices should return a distribution of θ centered around zero (subject to sampling error in estimating $\mathbf{G}$). While this approach compares differences in the trajectories, it does not account for differences in the *lengths* of those trajectories. For example, $\mathbf{G}_1$ may give a random vector that exactly aligns with the corresponding projection through $\mathbf{G}_2$, but is only $1/4$ the length.

An alternative to comparing by angular difference is to instead consider the *distance* between vectors, which takes into account both direction and length (Hansen and Houle

2008). From Equation A4.1b, the distance between the two responses is

$$d = ||\mathbf{G}_1\boldsymbol{\beta} - \mathbf{G}_2\boldsymbol{\beta}|| = \sqrt{(\mathbf{G}_1\boldsymbol{\beta} - \mathbf{G}_2\boldsymbol{\beta})^T (\mathbf{G}_1\boldsymbol{\beta} - \mathbf{G}_2\boldsymbol{\beta})}$$
$$= \sqrt{\boldsymbol{\beta}^T (\mathbf{G}_1 - \mathbf{G}_2)^T (\mathbf{G}_1 - \mathbf{G}_2)\boldsymbol{\beta}} \quad (29.8)$$

Hansen and Houle (2008) refer to the average of Equation 29.8 over a random set of $\boldsymbol{\beta}$ as the **response difference** between two $\mathbf{G}$ matrices, and show that it is (approximately) a function of the squared eigenvalues of the matrix $\mathbf{G}_1 - \mathbf{G}_2$. The mean of the d values forms one test statistic, and this can be robustly tested by a randomization. As before, families are randomly assigned to the two groups, and the above procedure run. Although a new random set of $\boldsymbol{\beta}$ could be generated for each run, a slightly more precise p value is obtained by first randomly generating a set of $\boldsymbol{\beta}$, and then applying the same set in the original and all randomized samples.

Random skewers attempt to quantity *random* selection. However, selection is far from random, and it is really the response in the direction(s) of actual selection that are of concern. For those admittedly rare cases where one has estimates of both $\mathbf{G}$ and $\boldsymbol{\beta}$ for two (or more) populations of interest, a more appropriate measure might be built around the estimated $\boldsymbol{\beta}$. If there are two (or more) estimates of $\boldsymbol{\beta}$ (for example, one from each population), then the average angle or distance based on this set of $\boldsymbol{\beta}$ can be used. Again, randomization tests (randomly assigning families to create new $\mathbf{G}$ matrices) can again be used to set appropriate p values. Likewise, since $\boldsymbol{\beta}$ itself is estimated with error, the "random" set of $\boldsymbol{\beta}$ chosen for the skewers might be the $\boldsymbol{\beta}$ arising from a bootstrap sample (Chapters 28, 29).

Comparison of Shared Geometry: The Flury Hierarchy

As mentioned on several occasions, our real interest in matrix similarity is the similarity of their geometry. For example, is variation oriented along the same direction (similar eigenvectors)? Is the variation similar along these axes (similar eigenvalues)? While some of the early comparisons were certainly aware of these issues (e.g., Lofsvold 1986 examined the angles between eigenvectors but offered no formal tests), Phillips and Arnold (1999) and Arnold and Phillips (1999), following a suggestion by Cowley and Atchley (1992), were the first to do so in a systematic fashion for comparisons of $\mathbf{G}$.

The basis for their approach is the **Flury hierarchy** (Figure 29.6), a nested set of hypotheses about matrix relationships proposed by Flury (1987, 1988). This approach is also referred to in the literature as **common principal components** or **CPC**. When comparing two n dimensional covariances matrices, Flury noted that we can order their similarity from unrelated to complete similarity in a hierarchy based on shared orientations, i.e., their eigenvectors/principal components (PCs). At the bottom of the hierarchy are matrices that share no eigenvectors and hence are regarded as being completely unrelated. Next, the two matrices could share one eigenvector, and Flury denoted this by **CPC(1)**, showing that they share one common principal component. One can proceed up the hierarchy to sharing two CPC(2) or more PCs. Elements in this set are often called **partial common principal components** or **PCPC**. **Full CPC** means that the two matrices share all the same eigenvectors. Note for an n dimensional matrix that the step is from CPC(n-2) to Full CPC, as if they share $(n-1)$ PCs, because of orthogonality they also share the last PC. Full CPC is not the top of the hierarchy, as the matrices could be proportional (also share all eigenvalues, up to a multipicative constant) and finally at the top of the hierarchy is equality where all eigenvalues and eigenvectors are shared.

Example 29.3. Arnold and Phillips (1999) applied the Flury hierarchy to compare $\mathbf{G}$ matri-

ces for six morphological traits between a coastal and inland population of the garter snake *Thamnophis elegans*. These two Californian populations are separated by nearly 300km and occupy very different habitats. Preliminary mitochondrial DNA analysis placed the divergence time as roughly 2 million years. Further, both populations showed considerable mean divergence in these traits. Separate covariance matrices were estimated for both sexes, and two different methods used for estimation — parent (female)-offspring regression and within-family variation. The Flury results (using the jump-up method with critical values determined by randomization) for comparing all combinations (males vs. female, coastal vs. inland) of the four groups were follows:

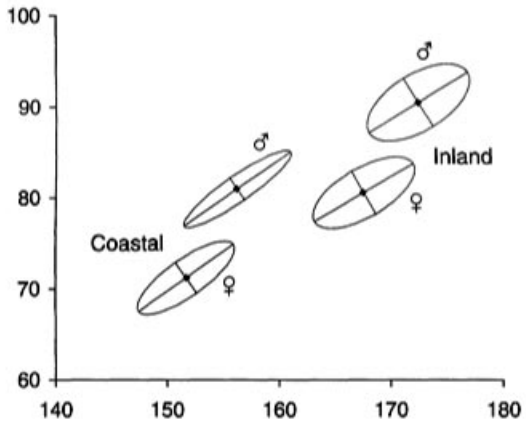
Method	Coastal M vs. F	Inland vs. Coastal M	Cos M vs. Inl F	All others
Regression	CPC(2)	CPC(1)	Full CPC	Full CPC
Family-mean	Full CPC	CPC(3)	CPC(4)	Full CPC

For three of the tested combinations, the pair of $\mathbf{G}$ matrices showed common principal components (Full CPC), so that the orientation was the same, but the eigenvalues differed (and not in a proportional fashion). Recall that CPC(4) is one step below Full CPC. For the other three combinations, at least one PC was shared between matrices, with the number of shared PCs depending on the method used to estimate $\mathbf{G}$ (but not in any consistent fashion).

Population	λ_1	λ_2	% first 2 PCs	trace($\mathbf{G}$)
Coastal males	7.61 (61%)	3.42 (28%)	88.69	12.43
Inland males	10.16 (53%)	6.51 (34%)	88.06	18.93
Coastal females	5.73 (45%)	5.24 (41%)	86.04	12.75
Inland females	5.85 (46%)	5.27 (41%)	87.28	12.74

As shown above, the leading eigenvalues differed between the four groups, but the first two eigenvalues in all cases account for almost 90% of the total variance in $\mathbf{G}$. Recall that the trace of a matrix (the sum of its diagonal elements) equals the sum of its eigenvalues, and hence the total variance of $\mathbf{G}$. Even if only the first two PCs are in common, this still accounts for the vast majority of the genetic variation.

The following figure plots the 95% confidence ellipsoids for the four populations using these first two PCs. Note that all these $\mathbf{G}$ matrices display a common orientation, but differ in the amounts of variation along the common axes. Thus, despite large differences in the means, the orientation of the major variation in $\mathbf{G}$ remains very similar. Further note that the divergence in the populations (given by the four centroids for the ellipses) also generally falls along the orientation of the first PC, in account to Schulter's notion of response along lines of lease resistance.



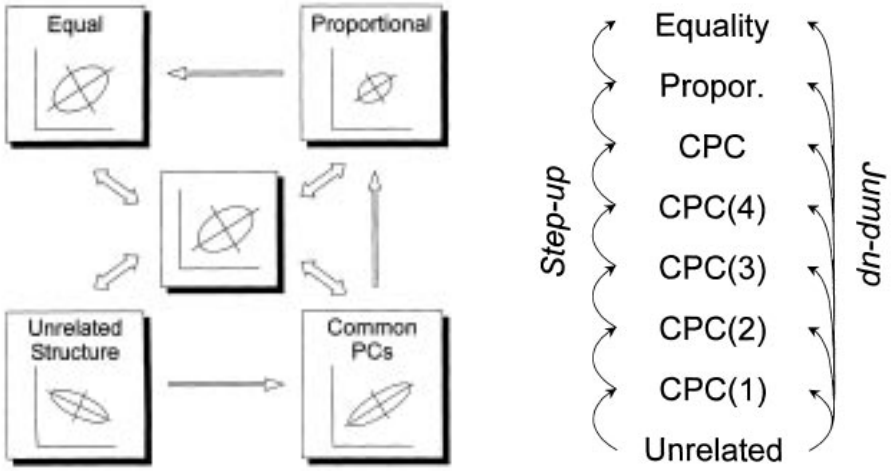


Figure 29.6 The Flury hierarchy of possible relationships between two matrices, building around shared eigenvectors (common orientation). **Left:** The similarity between two matrices can be classified into levels of increasing geometric similarity. Comparing one matrix with another (middle) can, at one extreme, result in no shared eigenvectors, with the two having unrelated structure. Next, one or more eigenvectors can be shared (common principal components, or CPC), up to complete sharing of all eigenvectors (Full CPC). Next up the hierarchy is that the eigenvalues of the two matrices are proportional, and at the top of the hierarchy is that both matrices are equal (same eigenvectors and eigenvalues). **Right:** Testing where a matrix comparison falls in this hierarchy can take several approaches. One is the so-called step-up approach where we move up the hierarchy one test at a time, stopping when the difference is significance. An alternative approach is the jump-up, where a test for each level is compared against the hypothesis of unrelated (see Example 29.4). After Phillips and Arnold (1999).

The Flury hierarchy is a vast improvement over tests only returning binary answers (similar/not similar, proportionate/not proportionate) and is very appealing because it attempts to capture more closely our notion of matrix similarity, namely shared geometry. However, it is not without problems. The first issue concerns determining which level of the hierarchy two matrices share, which involves a comparison of sets of hypotheses. Flury was originally designed for product-moment covariance matrices. In this case, under normality assumption, likelihood-ratio tests between hypotheses follow (for large samples) χ^2 distributions whose degrees of freedom are set by the difference in number of fitted parameters. Conversely, **G** is a variance-component covariance matrix (the elements are estimated indirectly) and frequently contains negative eigenvalues (which can be adjusted by bending). The appropriate degrees of freedom for Flury when using a variance component **G** estimate is often unclear (Phillips and Arnold 1999). For example, Blows and Higgie (2003) conservatively used the number of sires in their nest half-sib/full sib design as a lower limit on the number of degrees of freedom, as it was unclear how Flury takes into account number of observations per sire.

This important issue aside, Phillips and Arnold (1999) suggest three different comparison procedures. The first is the **step-up** approach. Here, adjacent steps in the hierarchy are tested, starting at unrelated, and when a significant p value is reached, one stops and accepts the previous level as the appropriate hierarchical position for the comparison. Alternatively, the **jump-up** procedure involves comparisons of each level with the unrelated level, where we accept the highest non-significant level. Figure 29.6 and Example 29.4 illustrates these. Alternately, the approach recommend by Flury was to use a model-comparison statistic, such

as **Akaike's information criteria** (Akaike 1973), which penalizes model fit by the number of fitted parameters. The model with the smallest AIC score is taken as the best fit. The problem is that this is not a formal statistical procedure. As the following example shows, Flury can return different answers for the same dataset, depending on which comparison procedure is used.

Example 29.4. Phillips and Arnold (1999) applied the step-up, jump-up, and AIC approaches to comparison of the $\mathbf{G}$ matrices for inland versus coastal females in garter snakes (Example 29.3). Parameteric likelihood-ratio tests were use, which (under appropriate assumptions) are χ^2 distributed for large sample sizes. In the step-up comparison approach, one moves up the hierarchy by testing each successive level against the pervious lower one, stopping once a significant value is reached. For these data, the authors observed:

Comparison	df	χ^2	p
CPC(1) vs. Unrelated	5	5.10	0.4035
CPC(2) vs. CPC(1)	4	2.00	0.7357
CPC(3) vs. CPC(2)	3	7.02	0.0714
CPC(4) vs. CPC(3)	2	0.54	0.7648
Full CPC vs. CPC(4)	1	6.30	0.0121*

Hence, the step-up criteria gives the level in the hierarchy as CPC(4). A randomization test (randomizing families over the two groups) also returns CPC(4) using step-up. Using the jump-up approach (each likelihood comparison is with respect to the unrelated model) gives

Hierarchy	df	χ^2	p
Proportionality	20	73.77	<0.0001
Full CPC	15	20.95	0.1384
CPC(4)	14	14.65	0.4020
CPC(3)	12	14.12	0.2931
CPC(2)	9	7.10	0.6264
CPC(1)	5	5.10	0.4035

Jump-up thus returns a slightly different answer, Full CPC. Randomization tests also return Full CPC using jump-up. Finally, comparing model fit using Akaike's information criteria,

Model:	CPC(1)	CPC(2)	CPC(3)	CPC(4)	FullCPC	Proportionality
AIC value	37.1	31.1	32.1	28.7	33.0	75.8

AIC chooses the CPC(4) model as the best fit. Different answer result from the three different criteria. The step-up and AIC approaches suggest CPC(4), while jump-up suggests Full CPC. An extreme example of this was offered by Ackermann and Cheverud (2000), who found the Flury step-up returns unrelated, AIC returns CPC(3) for their (phenotypic) data.

The most critical unresolved issue with Flury is power. It does seem that power is higher at the top of the hierarchy (rejection of equality or proportionality) than it is at lower levels such as partial CPCs, which appear to be more readily accepted (Phillips and Arnold 1999, Houle et al. 2002, Blows and Higgie 2003). Thus, especially for $\mathbf{G}$ matrices, Flury tends to reject equality and proportionality but accepts some level of CPC. Reducing the number of families and/or increasing the number of traits tends to decrease power, and this lack of power can result in the wrong model being favored. Conversely, when power is high, Flury almost seems too discriminating, rejecting common structure even when strong similarities exist among covariance structures (Steppan 1997a, Marroig and Cheverud 2001, Houle et

al 2002, Cheverud and Marroid 2007). Hence, there is a bias towards common principal components in small samples and a bias away from then in large samples.

Differences in the performance of the various model-selection approaches are likely also due to power issues. Houle et al. (2002) found that (for smaller samples) jump-up and AIC have different performance, for example jump-up (in their simulations) rejected the hypothesis of equality more frequently than expected. As sample size increases, these differences go away. A subtle issue related to power is that often Flury returns rather different results using $\mathbf{P}$ instead of $\mathbf{G}$ on the same dataset. $\mathbf{P}$ is estimated with more precision than $\mathbf{G}$ and it is often seen that common structure is rejected for $\mathbf{P}$, but accepted for $\mathbf{G}$ (e.g., Bégin and Roff 2001, 2003). While this could reflect important biological differences, it could also be simply a lack of power when using $\mathbf{G}$ and hence a default towards the null, which (under randomization) is matrix equality. A common observation when using Flury on $\mathbf{G}$ is full (or nearly so) CPC (Table 29.1). Whether this is reflecting deep issues in biology or trivial issues related to power is unclear and further work need to be done in quantifying the implications of low to modest power.

A related issue is that programs for computing Flury often default to testing the leading eigenvectors of the two matrices first. If these are not equal, transition up the hierarchy stops. This can significantly bias results (Houle et al. 2002, Blows and Higgie 2003), and a manual reordering of which vectors are to be tested is recommended. For example, Blows and Higgie (2003) found that PC1 for their data was generally not equal across matrices, but that PC2 was. A formal approach for comparing all subsets of vectors has not yet been proposed, but if a defined procedure is used, a randomization test should returned an appropriate p value, provided the same procedure is used on all randomized samples.

Scale issues can also be important, for example for morphological traits PC1 usually represents a general size measure, and so comparison of the leading eigenvectors amounts to a comparison of general size and scaling relationships with size. Blows and Higgie (2003) also caution that large differences in scale between traits can influence the structure of $\mathbf{G}$, and for this reason caution the use of CPC on life-history traits which are often measured over very different scales.

A final concern does not really apply when our interest is simply a comparison of $\mathbf{G}$, but it is still biologically important. CPC has been use to attempt to make inferences about the nature of factors underlying the observed pattern of variation (e.g., Klingenberg et al. 1996). Example 31.4 makes the key point made by Houle et al. (2002). Recall in this example that the genotypic values were given as $\mathbf{g} = \mathbf{F}\mathbf{f} + \mathbf{e}$, the effects of a vector of underlying factors $\mathbf{f}$ plus additional effects $\mathbf{e}$, with resulting covariance matrix $\mathbf{G} = \mathbf{F}\boldsymbol{\Sigma}_f\mathbf{F}^T + \boldsymbol{\Sigma}_e$. The orthogonal axes associated with $\mathbf{G}$ usually do not correspond to contributions from the true underlying true orthogonal factors $\mathbf{f}$. The effect of an underlying factor might be spread out over several orthogonal axes in $\mathbf{G}$ (the same, of course, holds for $\mathbf{P}$). Since CPC focuses on orthogonal comparisons, it tends to underestimate the degree of structure, and that two matrices with many casual factors in common may not have common CPC.

While Flury is a quantum improvement over other tests, it is not without interpretation issues. In particular, how similar is similar? For example, Caruso et al. (2005) examined seven physiological traits between two species of perennial wildflowers, *Lobelia siphilitica* and *L. cardinalis*. Flury finds they share two common PCs. At first, this does not sound all that impressive, but these two PCs account for 87% of the total genetic variation. When Flury results are presented, it would be helpful to mention what fraction of total variation the common (if any) PCs explain. A related issue is that the power for comparing eigenvectors is thought to decrease with the size of their eigenvalues. Thus, Flury places a bias on orientations with large amounts of variation. On one hand, as we have seen that these are genetic lines of least resistance, and populations often appear to diverge along directions close to these lines. On the other hand, we have also seen examples (Chapter 30) where natural selection is acting

in directions that are nearly orthogonal to the major axes of variation, perhaps because the past history of selection has eroded usable variation in these directions. If this later view is correct, then comparisons of these orientations with very small eigenvalues may be *the* critical comparisons among covariance matrices. Finally, there is a irony with Flury in that it is underpowered for $\mathbf{G}$ (which likely results in it returning more CPC than are actually present), but it may be overpowered for $\mathbf{P}$ in that subtle differences in the eigenstructure, which may be biological irrelevant, can still be scored as differences. When using phenotypic covariance matrices, a number of studies have obtained results where matrices were highly similar by matrix correlation tests but unrelated under Flury (Steppan 1997a, Ackermann and Cheverud 2000, Cheverud and Marroig 2007). As a result, one should use other tests (such as random skewers) in addition to Flury to obtain a better understanding of the nature of matrix similarity (e.g., Bégin and Roff 2003).

It is important to stress that the above issues simply point to areas needing improvement when applying the general *concept* of the Flury hierarchy. Viewing similar matrices as a series of shared geometric components, and trying to estimate these, is clearly the way forward for comparison of $\mathbf{G}$ matrices. Flury is an important, but initial, step along this path, and much traveling lays ahead.

Comparison of Shared Geometry: Krzanowski Subspace Compairson

A question of interest when comparing matrices is whether the subspace containing the majority of their variation is in common. For example, most of the genetic variation might reside on the first few PCs, which may be very different between matrices, and thus Flury may return unrelated. However, if such a subset of the PCs for both matrices essentially spans the same space, then both matrices have comparable available variation. Krzanowski (1979) suggests the following approach for examining this issue, which was applied by Blows et al. (2004) and Rundle et al. (2008) to the related problem of looking for common orientation of the $\mathbf{G}$ and $\boldsymbol{\gamma}$ matrices (Chapter 31). This approach applies equally well to comparing estimates of $\mathbf{G}$ (Petfield et al. 2005), and proceeds as follows. For each matrix, take the first several (up to $k \leq n/2$) leading eigenvectors and construct two projection matrices,

$$\mathbf{B}(i) = (\, \mathbf{e}_1(i) \quad \cdots \quad \mathbf{e}_k(i) \,) \tag{29.9}$$

Here $\mathbf{e}_j(i)$ is the jth leading eigenvector for $\mathbf{G}_i$. These two **subspace projection matrices** can be compared using the matrix

$$\mathbf{S} = \mathbf{B}^T(1)\,\mathbf{B}(2)\,\mathbf{B}^T(2)\,\mathbf{B}(1) \tag{29.10}$$

The eigenvalues of $\mathbf{S}$ describes the angles between the orthogonal axes of $\mathbf{B}(1)$ and $\mathbf{B}(2)$. Specifically, the smallest angle between any two orthogonal axes is given by $\cos^{-1}\sqrt{\lambda_1}$, where λ_1 is the leading eigenvalue of $\mathbf{S}$. Note that the constraint of $k \leq n/2$ arises because if $k > n/2$, then an angle of zero will always be recovered (Krzanowski 1979). Further, the sum of the eigenvalues of $\mathbf{S}$ is the sum of squared cosines between the sets of orthogonal axes in the two projection matrices. If these are completely aligned, this sum equals k, while if there is no shared orientation, this sum is zero. Thus, the sum of eigenvalues in the Krzanowski matrix $\mathbf{S}$ is a measure of the number of shared dimensions (of the k tested) between $\mathbf{B}(1)$ and $\mathbf{B}(2)$. When the matrices being compared are two estimates of $\mathbf{G}$, a randomization test (randomizing families over groups and then computing the sum of the eigenvalues of $\mathbf{S}$) can generate an appropriate threshold for significance.

Such a **subspace comparison** focuses entirely on the space containing variation, but not how much variation is in this space. Thus, the focus is entirely on the orientations within which a set amount of variation occurs, not how the variation is distributed *within* the

subspace. This is a strength, and weakness, of this approach. It is a strength for uunusual comparisons. For example, Blows et al. (2004) suggested using such a comparison to see if **G** and γ (the matrix of quadratic selection coefficients) share a similar space for the majority of their variation. Its weakness is that if variation is very unevenly distributed over axes in the subspace, two matrices may actually contain very little shared variation despite containing a lot of shared orientation space.

Still no Ideal Solution

It is clear that simple tests of similar versus not similar really provide very little useful information, and that comparison of geometric features is clearly the way to proceed. Although Flury is a good start, it is really just a first baby step into this complex issue. It has power issues (too often accepting structure for small sample sizes and too often rejecting structure when significant similarities are present in large sample sizes). The best solution seems to be to use several metrics, such as a general similarity test (e.g., T or MANOVA), a simple geometric metric (random skewers), and a full geometric test (Flury or Krzanowski), as has been done by several authors (Bégin and Roff 2001, 2003; Bégin et al. 2004; Cano et al. 2004; Doroszuk et al. 2008).

The challenge of matrix comparison will only continue to get more interesting as an emerging field is the phylogenetic comparison of **G** — examining changes in **G** over a phylogeny of multiple taxa, such as distant subspecies or recently diverged species (Steppan 1997a, b; Steppan et al. 2002). Formal comparisons over a phylogenetic tree require some appropriate distance metric. Although such a clean single metric for matrix comparison is lacking, measures based on random skewers (such as average angular or absolute distance) could serve as initial place-holders.

COMPARING COVARIANCE MATRICES: DATA

There is a large (and rapidly growing) body of work on comparing **G** matrices, but a little diligence is required when reading this literature as a number of rather distinct questions and comparisons have been lumped together. Many of the studies actually *do not* formally compare *additive* genetic covariances, as their design (as we discuss below) may include non-additive components as well. There is also the issue of what to compare, the full covariance matrix or the set of heritabilities and the correlation matrix. Kohn and Atchley (1988) found that while the correlation matrices for skull traits between mice and rat were significantly similar, their covariance matrices were not, largely because of the increase in total variance for a few traits. Covariance and correlation matrices share the same sets of eigenvectors, and hence have identical orientation, but their eigenvalues differ dramatically. An unresolved issue is which comparison (covariance or correlation) has more power to detect common aspects of structure. Covariance matrices are susceptible to scaling issues while correlation measures likely have higher standardized sampling variances (as we divide a covariance estimate by two variance estimates to obtain the correlation).

We mentioned diversity of comparisons. The empirical literature on **G** includes comparisons over different taxonomic scales — between different geographic populations the same species (Arnold 1981), between different subspecies (Lofsvold 1986), species (Paulsen 1996), and different genera (Kohn and Atchley 1988). Such comparisons are in line with the question of time scale for stability. However, the **G** comparison literature also includes contrasts among different groups from the same population, for example between sexes (Holloway et al 1993), different major morphs (such as large and small-winged forms in insects, Bégin et al. 2004), and different environments/treatments (Platenkamp and Shaw 1992). It has also been applied to compare matrices with different major alleles (Stinchcombe et al. 2009), be-

tween asexual and sexual populations following clonal selection (Pfrender and Lynch 2000), among parthenogenetic populations (Doroszuk et al. 2008), and covariance matrices from mutation-accumulation lines (Camara and Pigliucci 1999, Estes and Phillips 2006). A partial list of comparisons is given in Table 29.1.

While most papers refer to comparing **G**, strictly speaking this is the additive genetic covariance matrix. Estimates using father-offspring regression or paternal half sibs return estimates of **G**, while designs that use full-sib families (e.g., Arnold 1981, Brodie 1993) return estimates where additive and dominance effects are confounded (not to mention potential maternal effects). Estimates using clones or RILs (e.g., Platenkamp and Shaw 1992, Doroszuk et al. 2008) return the full *genotypic* covariance matrix. Indeed, the majority of studies on the stability of "**G**" are actually comparing matrices with the potential of additional contributions from non-additive genetic variance and shared maternal effects. A subtle feature when replicated genotypes (clones or RILs) are used is that line-means can be used to directly estimate **G** via a product-moment approach. This results in greatly increased power (over variance-component based approaches) but is also means that Flury can be over-discriminating for such cases. Which "**G**" is then of interest? If our concern is long-term prediction of selection response or retrospective analysis of past selection, then our focal target is the additive-genetic covariance matrix. Other genetic-variance matrices may provided some insight (for example, if they are stable, it is more likely that the additive-genetic subset of their covariances are also stable), but caution is still required in their interpretation.

Table 29.1. Summary (not exhaustive) of results from comparison of **G** matrices (mostly) from natural populations. If type of comparison is not mentioned, it is between populations within the listed species. In some cases the "populations" being compared are different groups (e.g., sex) or treatments (e.g., environment) from the same population. This table is only meant to showcase the diversity of studies, rather than provide an in-depth analysis of the general conclusions of each cited reference. For this, readers should consult the original papers.

Thamnophis elegans (garter snake): Chemoreceptive traits. Arnold (1981)
No apparent heterogeneity of estimates between populations
No formal statistics.

Peromyscus (3 species and subspecies): Morphological traits. Lofsvold (1986)
No apparent heterogeneity of estimates between populations
Matrix correlation using Pearson and Spearman
Reanalysis by Kohn and Atchley (1988) found no heterogeneity using Matrix correlation with K_c

Rats (*Rattus norvegicus*) vs. mice (*Mus musculus*): Pelvic traits. Kohn and Atchley (1988)
Genetic correlation matrices similar between species, covariance matrices not similar
Matrix correlation

Gammarus minus (Cave-dwelling amphipod): Eye and antenna size. Fong (1989)
Five of ten pair-wise comparisons between five populations showed significant similarity
Matrix correlation

Holcus lanatus (grass): Morphological traits. Billington et al (1988), Shaw and Billington (1991)
No significant differences in variance (upon reanalysis by Shaw and Billington)
Variance-component likelihood ratio test

Daphina pulex (water-flea): Life history traits. Spitze et al. (1991)
No significant differences between two pops in any of the six statistics used
Matrix correlation, T-type statistic, leading eigenvalue, determinant: tested by bootstrapping.

Anthoxanthum odoratum (grass): Growth and reproduction traits. Platenkamp and Shaw (1992)
Clones over two environments, no significant differences in variance components
Variance-component likelihood ratio test

Thamnophis ordinoides (garter snake): antipredator traits. Brodie (1993)

No heterogeneity of estimates between populations
Pairwise comparison of elements using t-tests (via delete-one family jackknife).

Adalia bipunctata (two-spot ladybird beetle): Chemical defense traits. Holloway et al. (1993)
No heterogeneity of estimates between sexes
Barlett's likelihood ratio test.

Mimulus guttatus/M. micranthus, 2 pops/each (Monkeyflower): floral traits. Carr and Fenster (1994)
No heterogeneity of estimates within populations of the same species or between species.
Element-by-element comparison (using t tests), matrix regression

Precis coenia and *P. evarete* (Buckeye butterflies): Wing pattern traits. Paulsen (1996)
No significant differences under Barlett's test, significant difference under pairwise element test
Barlett's likelihood ratio test, pairwise element tests using bootstrap confidence intervals.

Clarkia dudleyana (annual plant, Onagraceae): Morphological traits. Podolsky et al. (1997)
No significant differences in variance components
Variance-compoment likelihood ratio test

Medicago truncatula (Alfalfa relative): 24 Morphological traits. Bonnin et al et al. (1997)
Seven of 24 h^2 estimates significantly different between 2 populations.
Element-by-element comparison (using jackknife and bootstrap tests)

Allonemboius socius & *A. fasciatus* (Crickets): Femur & ovipositor length. Roff and Mousseau (1999)
Significant heterogeneity in both variances & the covariance between species
Element-by-element variance heterogeneity and t tests

Allonemboius socius & *A. fasciatus* (Crickets): Male calling traits. Roff et al. (1999)
No significant difference within populations, significant differences between populations in $\mathbf{G}$,
but not the correlation matrix. Species differences in $\mathbf{G}$ consistent with proportional changes.
Element-by-element comparison using T test, regression

Arabidopsis: Morphological traits over three EMS-mutation lines. Camara and Pigliucci (1999)
CPC(1) over the three lines. Flury hierarchy

Impatiens capensis (jewel-weed), 2 pops (of RILS) over 2 envs: Morphological traits. Donohue et al. (2000)
Significant differences (REML), unrelated to CPC(2) for comparisons across environments
Significant differences (REML), all Full CPC for comparisons across populations
Variance-component likelihood ratio test, Flury hierarchy

Drosophila melanogaster: Female life history traits in three populations. Service (2000)
No significant differences in variance components
Variance-component likelihood ratio test

Scabiosa canescens & *S. columbaria* (perennial herbs): 6 pops/species. Waldmann & Andersson (2000)
Species matrices CPC(1), Full CPC for pops of *canescens*, CPC(1) for pops of *columbaria*
Flury hierarchy

Daphina (water flea): Life history traits in asexual vs. sexual pops. Pfrender and Lynch (2000)
Full CPC between population of asexual clones before and after season of selection
CPC(1) between season-ending pop of asexual clones and their sexual offspring
Flury hierarchy

Gryllus firmus & *G. pennsylvanicus* (Crickets): Morphological traits. Bégin and Roff (2001)
G. pennsylvanicus matrices equal across two environments (Flury and T)
G. firmus equal (T), CPC (Flury) across two environments
3/4 between-species comparison equal, one CPC (Flury). 2/4 equal, 2/4 not equal (T)
T test and Flury hierarchy

G. firmus, pennsylvanicus, veletis (Crickets): Morphological traits. Bégin and Roff (2003)
firmus - pennsylvanicus: Equality under all three methods.
veletis - pennsylvanicus: Equality under CPC; distinct under T, MANOVA
veletis - firmus: Full CPC; distinct under T, MANOVA

Gryllus firmus over different wing morphs/environments: Morphological traits. Bégin et al. (2004)
Shortwing-32 degree $\mathbf{G}$ matrix different from others, all other combs. of morphs/temps similar

T test, Flury hierarchy, Jackknife-MANOVA

Rana temporaria (Frog) 2 pops over 3 environments: Morphological traits. Cano et al. (2004)
All populations and environments: jointly unrelated (Flury, MANOVA)
Env. treatments within same source populations: Full CPC or equal, no difference (MANOVA)
Pops within the same treatment: two Full CPC, one unrelated. Significant MANOVA difference.
Flury hierarchy, Jackknife-MANOVA

Lobelia siphilitica & *L. cardinalis* (Perennial wildflowers): Physiological traits. Caruso et al. (2006)
CPC(2), but PCs 1 & 2 account for 87% of total variation. Flury hierarchy

Impatiens capensis (jewel weed): Morphological traits . Stinchcombe and Schmitt (2006)
Full CPC for RILSs compared over litter vs. bare soil. Flury hierarchy

Caenorhabditis elegans: Life history traits. Estes and Phillips (2006)
Three sets of 50 lines/treatment over different N_e and accumulation times.
Full CPC (1/5), Proportionality (3/5) equality (1/5) among five sets of N_e comparisons.
Flury hierarchy

Acrobeloides nanus (Pathenogentic soil nematode): 3 Life history traits. Doroszuk et al. (2008)
Four replicate pops in two different (20 year) natural selection treatments
Flury: Equality within, unrelated across, treatments. Skewers: Similarity within, none between.
Flury hierarchy, random skewers

Conclusions

What general conclusions can be made about the stability of **G**? Dramatic changes in **G** are indeed seen over short time scales (usually typified by very strong selection) and also over changes in the environment. The later is a critical, and somewhat under-appreciated, observation. If most traits in nature are under stabilizing selection and then move to being directionally-selected by a change in the environment, there is no guarantee that **G** will not also change, and perhaps dramatically, as the environmental shifts.

Despite the dramatic shifts in **G** that have been seen in some studies, one general conclusion from the empirical studies (Table 29.1) is that while **G** itself is highly unlikely to be identical (or even proportional) between populations, it does seem to conserve some orientation for significant periods of time. This fits in with the theory, as we have seem that the effect of drift is to cause **G** to "wobble"or "breath", expanding and contracting in its eigenstructure and wiggling around its eigenvectors. Largely through the simulations of Jones et al. we see that a number of features (correlations in the selection and mutational patterns) can enhance the orientational stability (stability of eigenvectors) in the face of drift. While this pattern is consistent with the data, some caution is also in order. Assessment of stability in orientation is typically done through Flury, which is known to be under-powered for tests of common principal components. Thus, even when applied to largely random matrices we would see more CPC than expected given the significance level. This caution aside, the empirical observations of at least some common structure likely reflects both some true underlying similarities plus some help from lower power. The eigenvalues, on the other hand, are not at all stable. Biologically, the shared common orientation means that the *independent* trait combinations do not change dramatically over many populations, but the genetic variances associated with them do. Lofsvold (1986), in one of the pioneering papers on **G** matrix comparison, said it best: the issue is not whether **G** matrices between populations are different, but rather the time scale over which they diverge. Orientation seems to be preserved for longer time scales, while eigenvalues (genetic variances of independent traits) are much more liable.

ESTIMATING THE DIMENSION OF A COVARIANCE MATRIX

As we have seen, the appears to be a trend for independent trait combinations (eigenvectors) to be somewhat preserved over time, while the genetic variances associated with them (eigenvalues) are quite labile. If one thinks of the eigenvectors as the combinations of traits upon which selection can act, their associated eigenvalues reflect (among another things) the strengths of selection and mutation on those combinations. A critical unresolved issue in biology is whether an observed low variance for a particular trait combination reflects a past history of strong selection or simply the inability of the developmental system to produce sufficient variation in that trait (or perhaps both).

As we saw in Chapter 30, much of the focus on genetic constraints can the phrased in terms of the distribution of the eigenvalues for **G**. If, as is often seen (e.g., Kirkpatrick and Lofsvold 1992, Blows and Higgie 2003, Kirkpatrick 2009), **G** has most of its variation accounted for by just a few eigenvalues, questions arise as to the actual values of the smaller (**minor**) eigenvalues. Are they effectively zero? If so, how many are? When zero eigenvalues are present, there exist **evolutionary forbidden trajectories** (Kirkpatrick and Lofsvold 1992), trait combinations that cannot evolve. Of special interest to us is the **rank** of a covariance matrix, the number of independent trait combinations with at least some variation. This is given by the number of non-zero eigenvalues, and there is much interest in estimating the rank of both **G** and **P**. While the presence of zero eigenvalues marks an absolute constraint, very small eigenvalues can also impose significant constraints. When very small eigenvalues exist, the matrix is said to be **ill-conditioned**, and response is extremely slow along the direction(s) associated with these eigenvalue(s). Indeed, response might be so small as to be swamped by the effects of genetic drift. Thus, the distribution (**spectrum**) of eigenvalues for a covariance matrix is of fundamental importance. If the majority of trait combinations have very little standing variation, evolution is severely constrained in terms of how it can operate. Selection can be quite effective in changing one or two component traits, but if the strength of selection is roughly equally distributed over a large number of traits, the chance of significant usable variation along the direction favored by selection becomes increasingly unlikely, especially if **G** is ill-conditioned or has reduced rank. From an evolutionary response perspective, this means that while certain components will response to change, others will not (or only respond extremely slowly). Thus, the eigenvalue distribution has deep evolutionary implications. If only one or two trait combinations show the vast majority of variation, most directions for potential selection response are severely constrained. Information on these constraints lies in the lower tail of the eigenvalue spectrum. As we will see, unfortunately, there are significant bias when attempting to make inferences about minor eigenvalues.

The closely-related topic of **reduced rank estimators** for **G** is also examined here. As quantitative geneticists consider ever more traits, the dreaded "curse of dimensionality" becomes increasingly onerous. If we have q traits and add an additional one, we have $q + 1$ addition parameters to estimate: the genetic variance of our new trait and its genetic covariance with all q previous traits. Further, since **G** matrices typically have much of their variation in the first few dimensions, we quickly run out of power to accurately estimate these new parameters unless we correspondingly increase sample size. One approach to warding off this curse is with a reduced rank estimator, wherein we attempt to estimate the subset of **G** that has the most support (and also the most variation). This is related to the issue of the dimensionality of **G**, as methods for obtaining reduce-rank estimators also contain information on the rank of **G**.

These topics can be rather technical, and our review will be a bit brief. Meyer and Kirkpatrick (2008) provide an excellent introduction for readers wishing a deeper excursion into the theoretical statistics literature on these subjects.

Leading Eigenvalues are Overestimated, Minor Eigenvalues Underestimated

For covariance matrices in general, the leading eigenvalues tend to be overestimated and the

smaller eigenvalues underestimated. This arises because eigenvalues for a sample covariance matrix are chosen to partition the observed variation into orthogonal axes of maximal variation. This results in an overfitting for the largest eigenvectors (as we try to maximally account for variation) and hence an underfitting of the residual variance (minor eigenvectors). In the case of variance-component based matrices, the underestimation very often result in negative estimates, and hence not proper covariance matrices. For product-moment covariance matrices, Lawley (1956) obtained a large-sample approximation for the expected value of the eigenvalues for the special case where the data follows a multivariate-normal distribution. In this case, the sample covariance matrix follows a Wishart distribution (Appendix 3). Assuming an estimated k dimensional (product-moment) covariance matrix from a sample of size N, Lawley showed that if $\lambda_1 > \lambda_2 > \cdots \lambda_k$ (i.e. all eigenvalues are distinct), then the expected value for the ith eigenvalue is given by

$$E[\,\widehat{\lambda}_i\,] \simeq \lambda_i \left(1 + \frac{f_i}{N}\right), \quad \text{where} \quad f_i = \sum_{j \neq i}^{k} \left(\frac{\lambda_i}{\lambda_j} - 1\right)^{-1} \tag{29.11}$$

This expression ignores terms of order N^{-2} and higher. Equation 29.11 shows how the estimation error is partitioned into two components. The first, f_i, is an inflation factor that depends on the spread of the eigenvalues and where λ_i ranks with respect to the others. Eigenvalues that are very close together ($\lambda_i/\lambda_j \simeq 1$) show the greatest inflation, which is positive for $\lambda_i > \lambda_j$ and negative when the inequality is reserved. Sample size N enters by simply reducing the inflation factor for each eigenvalue by $1/N$.

Example 29.5. As an example of the implications of Equation 29.11, suppose we are examining the eigenvalues of an estimated 6 x 6 phenotypic covariance matrix $\mathbf{P}$ with true eigenvalues 10, 3, 2, 1, 0.5, 0.1. The (approximate) expected values for sample sizes of 20, 50, and 100 are

λ	f_i	N=20	N=50	N=100
10	0.852	10.852	10.170	10.085
3	1.306	3.392	3.078	3.039
2	-2.864	1.427	1.885	1.943
1	-3.500	0.650	0.930	0.965
0.5	-5.330	0.234	0.447	0.473
0.1	-5.450	0.046	0.089	0.095

Note that the two leading eigenvalues tend to be overestimated ($f_i > 0$) while the smaller eigenvalues tend to be underestimated ($f_i < 0$). While the effect may not appear to be dramatic, recall that these are *mean* values, and that there is considerable spread around these in any particular realization. Further, these results are for product-moment based covariance matrices. As we see below, the effect is expected to potentially be even more dramatic with between-group covariance matrices.

Bootstrap-based Confidence Intervals for Eigenvalues and Rank

The bootstrap can be a powerful approach for obtaining approximate confidence intervals and standard errors for complex distributions. Thus, one obvious test for the rank of a matrix is to use the bootstrap to determine how many of the estimated eigenvalues are significantly greater than zero. Specifically, a bootstrapped $\mathbf{G}$ is created by sampling families with replacement, its eigenvalues determined and assigned. An eigenvalue is declared to be significantly

greater than zero at the $(1-\alpha)$ level when α (or less) of the bootstrap samples assign it a value of zero (or less). Such an approach, however, must be used with care (Hine and Blows 2006). The direction of eigenvectors can shift significantly (even dramatically) among bootstrap samples, and correct assignment of an eigenvalue to its associated eigenvector from the original sample can be a bit problematic. Thus, in each bootstrap sample the temptation is to assign estimates of eigenvalues by their order in the sample (**order assignment**), rather than being consistent about assigning eigenvalues to similar eigenvectors over samples. Example 29.6 shows that rank-based assignment results in an underestimation of the true variance associated with each eigenvalue, making their bootstrap confidence intervals too narrow. This in term results in rank being significantly overestimated if it is based on the number of such support intervals that exclude zero. Hine and Blows (2006) verfied this point with computer simulations .

Example 29.6. Suppose we have bootstrap samples from a 3×3 covariance matrix, and we are trying to construct confidence intervals for the eigenvalues $\lambda_1, \lambda_2, \lambda_3$. Suppose the first four samples return the following eigenvalues:

	True Values				Inferred values				True Var	Inferred Var
λ_1	4	3	4	6	4	4	4	6	1.58	1.00
λ_2	2	4	3	2	2	3	3	3	0.92	0.25
λ_3	1	2	1	3	1	2	1	2	0.92	0.33

The four columns under true values correspond to the four bootstrap samples with the correct assignment of the bootstrap sample eigenvalue with its corresponding eigenvalue in the original sample matrix. It should be emphasized that due to sampling error in the orientation of the eigenvectors such a correct assignment can be problematic, at best. For example, in sample 2, the value corresponding to λ_1 in the original sample is now 3, while the value corresponding to λ_2 is larger, being four. Thus, the eigenvectors have switched rank between this particular bootstrap sample and the original. However, if we simply order the estimates, with the first corresponding the estimate of λ_1 and so forth, this gives the inferred values listed above. Such a rank-based assignment results in the values being more clustered, and hence with a smaller variance, than actually occurs. As we have seen, another source of error is the overestimation of leading eigenvalues and the underestimation of minor eigenvalues. The net result of these two sources of bias for minor eigenvalues (which is largely where rank is determined) is unclear. The underestimation of their value places their estimators closer to zero (creating a bias for underestimating rank), while underestimated variance creates a bias for overestimating rank.

An alternative for constructing bootstrap-based CIs is to use **projection assignment** to obtain the eigenvalues (Wagner et al. 2008). Since the eigenvectors from a bootstrap matrix may be hard to associate with those from the original sample, the projection assignment of eigenvalue i from a bootstrapped sampled matrix $\widehat{\mathbf{G}}_B$ is just

$$\widehat{\lambda}_i = || \, \widehat{\mathbf{G}}_B \, \mathbf{e}_i \, || \tag{29.12}$$

where $\mathbf{e}_i$ is the (unit length) eigenvector for eigenvalue i in the original sample estimate $\widehat{\mathbf{G}}$. Equation 29.12 is the amount of variation in the bootstrap estimate along the direction of eigenvector i in for the original estimate, as $\widehat{\mathbf{G}}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, which has length λ_i (since $\mathbf{e}_i$ has unit lenght).

A final caution is that if one can indeed construct the appropriate confidence intervals, there is a multiple comparison issue when moving from an eigenvalue CI to a statement

about rank. Rank is given by largest k such that the confidence intervals (CIs) for λ_i to λ_k all exclude zero. If these CIs are independent, if each individual confidence interval has a $1-\alpha$ chance of excluding zero, the *joint* probability that all k do is $(1-\alpha)^k \simeq 1-\alpha/k$. Thus for a $(1-\alpha)$ one-sided CI for rank, all k CIs should have a $1-\alpha/k$ chance of excluding zero. With lack of independence of CIs (which is likely, given their estimates use the same data), the $(1-\alpha)$ CI for rank is given by the largest k-dimensional $1-\alpha$ confidence ellipsoid (or hypervolume) that excludes zero.

A more appropriate bootstrap estimate for rank was offered by Mezey and Houle (2005), who simply used the empirical distribution of rank (number of non-zero eigenvalues) in a bootstrap sample of $\mathbf{G}$ (Example 31.7). Note that there has been some confusion (Hine and Blows 2006) with the analysis of Mezery and Houle, as they also used bootstrapping to construct confidence intervals on the eigenvalues, and some of their wording was ambiguous as to whether they used the number of non-zero confidence intervals or the distribution of rank in their bootstrap samples for their bootstrapped estimate of rank. The result was apparently conflicting simulation results showing that bootstrap estimates of rank work poorly (Hine and Blows 2006) or work fairly well (Mezey and Houle 2005). This discrepancy arises from these authors examining different estimators.

Example 29.7. In an attempt to estimate the rank of a large $\mathbf{G}$ matrix of similar traits, Mezey and Houle (2005) measured the locations of 12 wing vein intersections in *Drosophila melanogaster*. Shape was recovered from these 24 (x, y) coordinates using geometric morphometric techniques that remove size, and result in the loss of four degrees of freedom, so the that the maximal rank would be 20. Bootstrap samples of $\mathbf{G}$ were obtained by sampling (with replacement) from families within common treatment blocks. The rank of each sample was taken as the number of non-zero eigenvalues for that realization of $\mathbf{G}$, and the empirical distribution of rank generated using 1023 bootstrap samples. The bootstrap results are shown in the table below. Original refers to the rank seen in the initial sample. Median, 5% and 1% correspond to the median (50%), (lowest) 5% and 1% values for the observed ranks in the bootstrap sample.

Matrix	Original	Median	5%	1%
$\mathbf{G}$ (Males)	20	19	17	17
$\mathbf{G}$ (Females)	21	20	18	17

The original rank is full for both the male and female-specific $\mathbf{G}$ matrices. The 95% and 99% confidence intervals are that rank is ≥ 17 in males, while the 99% and 95% confidence intervals for female are a rank of ≥ 17 and ≥ 18, respectively. Most estimates in the literature for matrices with a significantly fewer number of traits are much less than full rank, so these results are a little surprising. Mezey and Houle also observed a very unusual distribution of eigenvalues, with a gradual decline rather that the more precipitous decline that is typically seen. Two factors could help explain these results. First, the sample sizes used to estimate $\mathbf{G}$ were much larger than normal. Second, the analysis was of shape with size variation removed. For morphological traits, the dominant eigenvalue is a measure of size variation and this typically accounts for half (or more) of the variation. Conversely, Mezey and Houle observed each eigenvalue was roughly 0.34 of its predecessor, resulting in a smooth decline seen down to very small eigenvalues. As a point of reference, the amount of decline they observed over their 20 eigenvaluess was comparable to what Kirkpatrick and Lofsvold (1992) saw over four.

While a simple bootstrapping using the observed rank seems a straightforward and

reasonable approach for constructing confidence limits on rank, two caveats are important to keep in mind. First, smaller eigenvalues tend to be underestimated in samples, and this is also expected occur in each bootstrap sample, biasing their values downward. Second, while the bootstrap is a powerful tool, and often the only one available for complex distributions, it does not necessarily give appropriate length confidence intervals, and thus care must be taken in the interpretation of results.

Canonical Decomposition of the Estimated Covariance Matrix

Here we develop (in outline form) the canonical decomposition for an estimate of **G** under the simplest case of estimating variance components from a balanced one-away ANOVA (for example, using half- or full-sib families). As we will shortly see, such a decomposition will prove quite useful, providing the basis for a test of the rank of **G** as well as **reduced-rank** estimator of **G**. Recall (Chapter 30) that the canonical decomposition for a $p \times p$ symmetric matrix **A** is given by

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$$

where

$$\mathbf{U} = (\mathbf{e}_1, \ldots, \mathbf{e}_p), \quad \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_p)$$

where $\mathbf{e}_i$ and λ_i are ith eigenvector and its associated eigenvalue of **A**.

Consider a balanced one-way design where s groups (for example, half- or full-sib families) are measured, each with n observations. As we saw in LW Chapter 18, the univariate estimate of the between-group variance given by $\widehat{\sigma}_B^2 = (M_B - M_W)/n$. Here M_B denotes the between-group mean squares (squared deviations of the means of families from the grand mean) and M_W the within-group mean squares (squared deviations of observations within a family from the family mean). Since the variance between groups equals the covariance within groups (LW Chapter 18), the variance between family means equals the covariance among family members. Assuming no nonadditive genetic effects and no shared environmental/maternal effects, $\sigma_B^2 = r\sigma_A^2$, where r is the genetic relatedness among the family members ($r = 0.25$ for half sibs and 0.5 for full-sibs). Hence,

$$\widehat{\sigma}_A^2 = \frac{1}{r}\,\widehat{\sigma}_B^2 = \frac{1}{r}\,\frac{M_B - M_W}{n}$$

Similarly, with a vector of observations per individual, the estimated genetic covariance matrix becomes

$$\widehat{\mathbf{G}} = \frac{1}{r}\,\frac{\mathbf{M}_B - \mathbf{M}_W}{n} \tag{29.13}$$

where M_x is replaced by $\mathbf{M}_x$, symmetric matrices of sums of squares and crossproducts. Thus, the eigenvectors and eigenvalues of the estimated covariance matrix correspond to the eigenvectors and (scaled) eigenvalues of the matrix $\mathbf{M}_B - \mathbf{M}_W$. A number of authors (Amemiya 1985, Hine and Blows 2006, Meyer and Kirkpatrick 2008) have shown that we can write the canonical decomposition of this matrix as

$$\mathbf{M}_B - \mathbf{M}_W = \mathbf{T}\,(\boldsymbol{\Lambda}_Q - \mathbf{I})\,\mathbf{T}^T \tag{29.14a}$$

Here $\boldsymbol{\Lambda}_Q$ is a diagonal matrix containing the eigenvalues of the matrix

$$\mathbf{Q} = \left(\mathbf{M}_W^{-1/2}\right)\mathbf{M}_B\left(\mathbf{M}_W^{-1/2}\right)^T \tag{29.14b}$$

Note that $\boldsymbol{\Lambda}_Q - \mathbf{I}$ is also a diagonal matrix, and contains the eigenvalues for the matrix $(\mathbf{M}_B - \mathbf{M}_W)$, and hence (appropriately scaled) for $\widehat{\mathbf{G}}$/ Likewise,

$$\mathbf{T} = \mathbf{W}^{1/2}\mathbf{U}_Q \tag{29.14c}$$

where $\mathbf{U}_Q = (\,\mathbf{e}_1, \cdots, \mathbf{e}_p\,)$ is the matrix of eigenvalues for $\mathbf{Q}$. Several useful applications immediately follow from Equation 29.14, and we discuss these in turn.

Amemiya's Rank Test

One immediate observation from Equation 29.14a is that the ith eigenvalue of $\widehat{\mathbf{G}}$ is given by $\lambda_{Q,i} - 1$ and is only non-negative if the eigenvalue for $\mathbf{Q}$ is one or greater. Any eigenvalue of $\mathbf{Q}$ less than one creates a negative eigenvalue for $\widehat{\mathbf{G}}$. Thus, one test for the rank of $\widehat{\mathbf{G}}$ is how many eigenvalues of $\mathbf{Q}$ are significantly greater than one. Such a test was proposed by Amemiya (1985) and used by Hine and Blows (2006), who extended this to the design of half-sibs nested within full-sibs. Amemiya's test proceeds as follows. Order the eigenvalues of $\mathbf{Q}$ as

$$\lambda_1 > \lambda_2 \cdots > \lambda_k \geq 1 > \lambda_{k+1} > \cdots > \lambda_p$$

An upper bound on the supported dimension of $\widehat{\mathbf{G}}$ is thus k (as there are only k eigenvalues ≥ 1). Suppose we wish to test that there is statistical support of $m \leq k$ dimensions. To test against the null that $m \leq b$, compute the test statistic

$$Y = (M+N) \sum_{i=b+1}^{k} \ln\left(\frac{M\lambda_i + N}{M+N}\right) - M \sum_{i=b+1}^{k} \ln(\lambda_i) \tag{29.15}$$

where $M = s-1$ and $N = s(n-1)$ are the degrees of freedom for the between and within levels. This is a likelihood ratio tests under assumptions of normally-distribution breeding values. Its the asymptotic distribution does not follow a chi-square distribution (Anderson and Amemiya 1991), but critical values of Y are tabulated by Amemiya et al. (1990) and Kuriki (1993). One first starts with a test of $b = k-1$ and proceeds to successively smaller values until significance is reached. Rejection the null that $m \leq b$ implies evidence for $b+1$ dimensions of support for $\mathbf{G}$. A small simulation study by Hine and Blows (2006) found that Amemiya's method works well for traits with higher heritabilities, but consistently underestimates the correct dimension for low heritability traits. This is perhaps not surprising in that critical values for Y are based on large-sample asymptotics and low heritability and/or small size (i.e., number of families) equates to low power. Meyer and Kirkpatrick (2008) obtained similar results.

Reduced-Rank Estimates of G

It will prove useful to partition the canonical decomposition (Equation 31.12a) into two parts. To do so, we write

$$\boldsymbol{\Lambda}_Q = \begin{pmatrix} \boldsymbol{\Lambda}_k & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_l \end{pmatrix} \tag{29.16a}$$

where $\boldsymbol{\Lambda}_k$ is a diagonal matrix of the first k eigenvalues of $\mathbf{Q}$, while $\boldsymbol{\Lambda}_l$ is an $p-k$ diagonal matrix of the remaining eigenvalues,

$$\boldsymbol{\Lambda}_k = \text{diag}(\lambda_1, \cdots, \lambda_k), \quad \text{and} \quad \boldsymbol{\Lambda}_l = \text{diag}(\lambda_{k+1}, \cdots, \lambda_p)$$

One obvious partition is to take k as the rank of the smallest eigenvalue > 1, corresponding to all the eigenvalues in $\widehat{\mathbf{G}}$ greater than zero, while $\boldsymbol{\Lambda}_l$ contains eigenvalues ≤ 1 (zero and negative eigenvalues in $\widehat{\mathbf{G}}$). We can also partition the matrix $\mathbf{T}$ containing the eigenvectors of $\mathbf{Q}$ into components corresponding to each of these pieces,

$$\mathbf{T} = (\,(\mathbf{e}_1 \quad \cdots \quad \mathbf{e}_k) \quad (\mathbf{e}_{k+1} \quad \cdots \quad \mathbf{e}_n)\,) = (\,\mathbf{T}_k \quad \mathbf{T}_l\,) \tag{29.16b}$$

$\mathbf{T}_k$ is a $p \times k$ matrix containing the first k eigenvector of $\mathbf{Q}$, while $\mathbf{T}_l$ is a $p \times (p-k)$ matrix containing the last $(p-k)$ eigenvectors of $\mathbf{Q}$.

Using these, a little matrix multiplication completes our partition,

$$\begin{aligned}\mathbf{T}\left(\boldsymbol{\Lambda}_Q - \mathbf{I}\right)\mathbf{T}^T &= \begin{pmatrix}\mathbf{T}_k & \mathbf{T}_l\end{pmatrix}\left(\begin{pmatrix}\boldsymbol{\Lambda}_k & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_l\end{pmatrix} - \begin{pmatrix}\mathbf{I}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{p-k}\end{pmatrix}\right)\begin{pmatrix}\mathbf{T}_k & \mathbf{T}_l\end{pmatrix}^T \\ &= \mathbf{T}_k\left(\boldsymbol{\Lambda}_k - \mathbf{I}_k\right)\mathbf{T}_k^T + \mathbf{T}_l\left(\boldsymbol{\Lambda}_l - \mathbf{I}_{p-k}\right)\mathbf{T}_l^T \end{aligned} \tag{29.17}$$

Equations 29.13and 29.17 imply that we can write the estimate $\mathbf{G}$ matrix as

$$\widehat{\mathbf{G}} = \widehat{\mathbf{G}}_R + \widehat{\mathbf{G}}_N \tag{29.18a}$$

where

$$\widehat{\mathbf{G}}_R = \frac{1}{nr}\mathbf{T}_k\left(\boldsymbol{\Lambda}_k - \mathbf{I}_k\right)\mathbf{T}_k^T = \frac{1}{nr}\sum_{i=1}^{k}\left(\lambda_{Q,i} - 1\right)\mathbf{e}_i^T\mathbf{e}_i \tag{29.18b}$$

and

$$\widehat{\mathbf{G}}_N = \frac{1}{nr}\mathbf{T}_l\left(\boldsymbol{\Lambda}_l - \mathbf{I}_{p-k}\right)\mathbf{T}_l^T = \frac{1}{nr}\sum_{i=k+1}^{p}\left(\lambda_{Q,i} - 1\right)\mathbf{e}_i^T\mathbf{e}_i \tag{29.18c}$$

If k corresponds to the smallest eigenvalue of $\mathbf{Q}$ greater than one, then $\widehat{\mathbf{G}}_R$ is a strictly positive-definite matrix, while $\widehat{\mathbf{G}}_N$ contains all of the negative eigenvalues of $\widehat{\mathbf{G}}$. Dropping the second term is equivalent to setting all negative eigenvalues to zero (equivalently setting to one any eigenvalues of $\mathbf{Q}$ that are less than one). Within a balanced one-way MANOVA, Amemiya (1985) showed that the estimate given by Equation 29.18b is equivalent to the REML estimate constrained to be positive-define (Klotz and Putter 1969).

$\widehat{\mathbf{G}}_R$ is a reduced rank estimate of $\mathbf{G}$, and contains that space that whose orientations are associated with positive eigenvalues. Two important modifications of this estimate both consider only the first $m \leq k$ eigenvalues. One is to set the dimension not by the *observed* number of eigenvalues greater than one, but rather by taking m to be the number *statistically supported* as being greater than one, for example by using Amemiya's test. The second approach is where the dimensionality is to some degree preset by the investigator, for example, using only the first four PCs of $\mathbf{Q}$, even if more are supported. The motivation for this last approach is that these may explain the vast majority of the variation in $\mathbf{G}$, even if other eigenvalues of $\mathbf{Q}$ are still significantly greater than one.

The dimension chosen in Equation 29.18b is critical. Suppose we decided to fix this at m, where for the true $\mathbf{G}$, $\lambda_{m+1} > 0$, i.e., its true rank exceeds m. What happens in large sample sizes? $\mathbf{G}_r$ will not be a consistent estimator, as it will *not* converge to $\mathbf{G}$ as sample size increases (Remadi and Amemiya 1994, Meyer and Kirkpatrick 2008). In particular, Meyer and Kirkpatrick (2008) show that

$$E[\mathbf{G}_r] = \mathbf{G} - \frac{1}{r}\sum_{i=k+1}^{p} E[\,\lambda_i\,\mathbf{e}_i\mathbf{e}_i^T\,] \tag{29.19}$$

where λ are eigenvalues of $\mathbf{G}$ and $\mathbf{e}$ are eigenvector of $\mathbf{Q}$. The bias is inversely related to the degree of genetic relatedness (worse for half- than full-sibs) and it does not decrease with increasing sample size. This bias arises because we have *fixed* the dimension. If we let the dimension be set by statistical support (such as Amemiya 's test for rank), then the estimator is consistent, as the supported dimensions will increase with sample size. While Equation

31.18b ensures that our estimate of $\mathbf{G}$ is positive-definite, it offers no such guarantee about $\mathbf{E}$ (Meyer and Kirkpatrick 2008).

Factor-Analytic Approaches for Building Reduced-Rank Estimates

Amemiya's Reduced-rank estimator (Equation 29.17b) is appealing given the common observation that much of variation in $\mathbf{G}$ is typically restricted to a few dimensions, but it has its limitations. First, this approach is restricted to particular designs (such as balanced ANOVAs) that allow the canonical decomposition of $\widehat{\mathbf{G}}$ can be performed. Second, we still have to first estimate $\mathbf{G}$, which (for a p-dimensional matrix) requires estimation of p variances and $p(p-1)/2$ covariances, for a total of $p(p+1)/2$ parameters.

Kirkpatrick and Meyer (2004) and Meyer and Kirkpatrick (2005) dealt with both of the issues by proposing a very general class of reduced-rank estimates of $\mathbf{G}$, where rather than estimating $\mathbf{G}$, we instead directly estimate the first k PCs of $\mathbf{G}$ instead. As we will see, this idea generalizes to most designs and involves estimating many less parameters. With the first k PCs in hand, we can construct a reduced-rank estimate that has the important property of being guaranteed to be positive semi-definite. This general approach has been called **factor-analytic** modeling, as we try to account for the observed variation in the data with a few factors (in this case, by directly estimating first few PCs of $\mathbf{G}$).

The motivation for this approach follows from the spectral decomposition of a square matrix $\mathbf{A}$ into its eigenvalues and eigenvectors,

$$\mathbf{A} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T + \cdots + \lambda_n \mathbf{e}_n \mathbf{e}_n^T$$

here $\mathbf{e}_i$ is the (unit-lenght) eigenvector for $\mathbf{A}$ and λ_i be its associated eigenvalue (Equations 30,44, A4.9b). Taking $\mathbf{f}_i = \sqrt{\lambda_i}\mathbf{e}_i$, we can write

$$\mathbf{A} = \sum_{i=1}^{p} \mathbf{f}_i \mathbf{f}_i^T \tag{31.18a}$$

Note that the variation along direction given by $\mathbf{f}_i$ is just the square of its length, as

$$|| \mathbf{f}_i ||^2 = \mathbf{f}_i^T \mathbf{f}_i = \lambda_i \, \mathbf{e}_i^T \mathbf{e}_i = \lambda_i \, || \mathbf{e}_i ||^2 = \lambda_i \tag{29.18b}$$

as the $\mathbf{e}_i$ have unit length.

Kirkpatrick and Meyer proposed to directly estimate the first k of the $\mathbf{f}_i$, and take

$$\widehat{\mathbf{G}}_R = \sum_{i=1}^{k} \widehat{\mathbf{f}}_i \widehat{\mathbf{f}}_i^T \tag{29.19}$$

as their estimator of $\mathbf{G}$. From Equation 29.18b, the additive genetic variance along the i-th PC is the square of the length of $\mathbf{f}_i$. By construction $\widehat{\mathbf{G}}_R$ is guaranteed to be positive semi-definite, as all the $\mathbf{f}_i$ have positive eigenvalues. However, it is also singular and hence a reduced-rank estimate as $p-k$ eigenvalues are set to zero. Note that this approach also provides a test for dimensionality, as one can perform likelihood-ratio tests for whether the fit is significantly improved with $k+1$ versus k PCs.

As an example of how to apply this approach, consider the simplest animal model,

$$\mathbf{z}_i = \boldsymbol{\mu} + \mathbf{a}_i + \mathbf{e}_i \tag{31.20a}$$

where $\mathbf{z}_i$ is the vector of trait values for individual i, $\boldsymbol{\mu}$ the mean vector for these traits, $\mathbf{a}_i$ the vector of breeding values, and $\mathbf{e}_i$ the vector of residual errors. We can decompose the

vector of breeding values for the observed i traits into a vector of breeding values for the trait combinations given by the PCs,

$$\mathbf{a}_i = \sum_{j=1}^{p} \alpha_{i,j}\mathbf{f}_j \tag{29.20b}$$

where $\alpha_{i,j}$ is the breeding value for jth PC in individual i, giving

$$\mathbf{z}_i = \boldsymbol{\mu} + \sum_{j=1}^{p} \alpha_{i,j}\mathbf{f}_j + \mathbf{e}_i \tag{31.21}$$

For example, for a model assuming only one PC,

$$\mathbf{z}_i = \boldsymbol{\mu} + \alpha_{i,1}\mathbf{f}_1 + \mathbf{e}_i \tag{31.22a}$$

For n individuals, one estimates $\boldsymbol{\mu}$, $\mathbf{f}_1$ and the n $\alpha_{i,1}$, correspond to the breeding values along the direction given by $\mathbf{f}_1$. If adding the first PC results in a significant improvement in the model, a second PC can be added (which results in a re-estimation of the first PC as well),

$$\mathbf{z}_i = \boldsymbol{\mu} + \alpha_{i,1}\mathbf{f}_1 + \alpha_{i,2}\mathbf{f}_2 + \mathbf{e}_i \tag{31.22b}$$

One continues in this fashion until fitting of an additional PCs does not significantly improve fit (which can be tested using a standard likelihood-ratio test). The stopping value k also provides an estimate of the rank as well.

This approach can also be used as a matrix-comparison method (akin to Flury). Suppose we wish to ask if two matrices share the same first PC. Let $\mathbf{z}_{j,i}$ denote a vector of observations from individual i in group j. We first fit a model allowing for different group means, but a similar first PC,

$$\mathbf{z}_{j,i} = \boldsymbol{\mu}_j + \alpha_{j,i,1}\mathbf{f}_1 + \mathbf{e}_{j,i} \tag{31.23a}$$

Next, we fit a model that allows the first factor to vary over groups,

$$\mathbf{z}_{j,i} = \boldsymbol{\mu}_j + \alpha_{j,i,1}\mathbf{f}_{j,1} + \mathbf{e}_{j,i} \tag{31.23b}$$

and a likelihood ratio test for an improved fit is done. If this is not significant, we accept PC1 as being in common and then move to PCs (note this is very similar to Flury). Also note that while we have framed this in terms of two groups, this approach allows us to compare an arbitrary number of groups.

Kirkpatrick and Meyer's approach is extremely appealing for several reasons: it easily extends to a very large number of designs, returns a reduced-rank estimate of **G** which does not require fully estimating all of the components of **G**, provides a likelihood-based test for rank, and guarantees a positive-definite estimate of **G** (and **E** as well with REML under the animal model, Meyer and Kirkpatrick 2008). However, Meyer and Kirkpatrick (2007, 2008) caution that some care is needed when implementing this approach. We have already seen one source of error, which also arises with Equation 31.18b, namely bias due to rank reduction (Equation 29.19). A second source of error is that, as one adds more PCs, the values of the initial PCs can change. Note that is not the case with our reduced estimated based on Equation 29.18b. If we change the dimension from k to $k+m$, the estimate still contains the original estimates of $\lambda_{Q,i}$ and $\mathbf{e}_i$ as for $1 \leq i \leq k$, these do not change. Conversely, under a factor-analysis model, the values associated with each PC can change as addition PCs are fitted, resulting in a bias due to what Meyer and Kirkpatrick call **subset selection**, where the

wrong PCs are included. They found in their simulations that the last PC (or factor) added tends to be underestimated and that the bias associated with it declines rapidly as addition PCs are fitted. In the extreme case, the first factor fitted may turn out to be something other than PC1, such as PC2 or PC 3 (Meyer and Kirkpatrick 2008, M. Blows pers comm. 2009). Incorrect assignment of early PCs is especially problematic if the eigenvalues of $\mathbf{G}$ and $\mathbf{E}$ align. Meyer and Kirkpatrick strongly caution care when using a stopping rule for number of PCs to include, and offer two suggestions. First, instead of a step-up approach (start at one, build to two, etc.), that a step-down approach might be better (start at $m+1$ then test fit of m, etc.). They also suggest monitoring the behavior of the trace of the reduced rank estimate at each iteration. Roughly speaking, for PC $m+1$, we should have

$$\text{tr}\left(\widehat{\mathbf{G}}_{r,m+1}\right) - \text{tr}\left(\widehat{\mathbf{G}}_{r,m}\right) = \lambda_{m+1}$$

larger differences indicates that the subset of PCs chosen is not sufficiently large.

Dimensionality of G: Data

Unlike the vast enterprise of $\mathbf{G}$ matrix comparison, to date only a handful of $\mathbf{G}$ matrices have had formal estimates of their dimensionality, in part due to the lack appropriate statistical tools. Untill recently, there also was much greater interest in comparison of $\mathbf{G}$, but little general interest in its dimension. While the theroetical importance of the dimensionality of $\mathbf{G}$ is not a new concept (Dickerson 1955, Lande 1979, Pease and Bull 1988, Kirkpatrick et al. 1990), that it might be something biologically important, rather than being statistically trivial, was a bit longer in developing (reviewed in Blows 2007).

The first formal statistical tests of the dimensionality of a $\mathbf{G}$ matrix appears to be Kirkpatrick and Lofsvold (1992). They used a **parametric bootstrapping** approach suggested by Kirkpatrick et al. (1990), wherein one uses parameter estimates to generate random variables of interest. In their case, they used the estimated standard errors (under normality assumptions) for the individual components (variances and covariances) of $\mathbf{G}$ to generate random draws for each of these component, which together formed a single bootstrap sample of $\mathbf{G}$. With these bootstrap estimates of $\mathbf{G}$ in hand, eigenvalues were computed and the distribution of an eigenvalue across the collection of samples used to construct empirical confidence intervals (CIs) for each eigenvalue. Rank given by the number eigenvalues whose bootstrapped CIs excluded zero. While we have seen that rank-assigned eigenvalue methods tend to overestimate the dimensionality, simulation studies by Mezey and Houle (2005) found that this parametric form appeared to significantly *underestimate* rank. Using a $\mathbf{G}$ with rank 20 and the eignevalue distribution observed in their wing data, this approach returned generally fewer than 10 eigenvalues whose confidence intervals excluded zero, a dramatic underestimation of rank. While an important benchmark paper, it is likely that the supported number of dimensions was conservative.

Statistical support on rank is a tricky issue and perhaps the strongest statement we can make is that the *actual* rank is very likely to be larger than the *supported* rank. First, rank is largely determined by how many minor eigenvalues have support above zero, but observed minor eigenvalues are typically *underestimates* of their true value. Hence, the rank of a sample matrix is a biased underestimate of the true rank. This also holds for bootstrap samples, as the process of obtaining the eignevalues for a matrix tends to over-inflate the variance accounted for the by largest factors and under-inflate the impact of the minor eigenvalues. Thus, using the empirical distribution of rank in a collection of bootstrap samples also likely results in an underestimate. Finally, power is an also issue. Eigenvalues of small, but positive, effect have low power of being supported. Taking all of these concerns into account, one might think that estimates of the statistically-support rank should always be viewed as *underestimates* of the true rank. Two factors balance this somewhat. With a matrix of less than full rank,

measurement error can introduce support onto dimensions of zero genetic variation, adding error variance in place of zero genetic variance. Second (and likely related), Remadi and Amemiya (1994) and Meyer and Kirkpatrick (2008) found an excess of $\lambda_Q > 1$ values when **G** was less than full rank.

Table 29.2 summaries the current rank estimates of **G**. The results are mixed, with some **G** appearing to have statistical support for nearly full rank (Mezey and Houle 2005, Meyer 2005), while others had support for less than half of the potential maximum rank (Kirkpatrick and Lofsvold 1992, Hine and Blows 2006, McGuigan and Blows 2007).

Table 29.2. Statistically-supported estimates of the rank of **G**. CIs refers to confidence intervals, FA to factor-analytic approaches, AIC to Akaike's information criteria as the model-fitting statistic.

Kirkpatrick and Lofsvold 1992	Growth traits in 4 species. Parametric bootstrap CIs on λ
White leg-horn chickens	Support for 2 of 4 dimensions.
Chios sheep	Support for 3 of 5 dimensions.
Hybrid sheep	Support for 2 of 5 dimensions.
Mice	Support for 4 of 9 dimensions.
Mezey and Houle 2005	
Drosophila melanogaster	Support for 17 (females) and 18 (males) dimensions
size-corrected wing traits	(out of 20) using boostrap distribution of ranks.
Meyer 2005	
Angus cattle	Best-fitting FA model (AIC) full rank
8 Meat quaillty traits	
Mezey and Houle 2005	
Drosophila serrata	Support for 2 dimensions under FA, Amemiya's test;
8 cuticular hydrocarbons	4 under bootstrap eigenvalue CIs.
Meyer 2007	
Angus cattle	Best-fitting FA model (AIC) 8 Factors out of 14.
14 carcass traits	possible dimensions.
McGuigan and Blows 2007	
Drosophila bunnanda	Support for 2 (females) & 5 (males) of 10 possible
wing traits	dimensions using factor-analytic modeling.

Given the common observation that the first few PCs often account for the vast majority of variation in **G**, why all the fuss about its true dimension? At the heart of this apparently simple question is a deep biological issue: do traits historically change by selection along axes of maximal variation or does a history of past selection remove variation in important directions? We have commented at length (Chapter 30) on the observation of evolution along lines of genetic lest resistance (along large eigenvalues of **G**), but this is also what is expected under drift. We have also seen (Chapter 30) that for cases where both $\boldsymbol{\beta}$ and **G** can be measured, there is often very little genetic variation along the direction $\boldsymbol{\beta}$ of apparent natural selection. Together, these observations suggest that one view of evolution is that much of the change along lines of genetic least resistance is largely neutral, while natural selection is required to do some very heavy lifting in order to make further progress along the direction of past selection, as most of the variation has been removed. If this view is correct, then the minor eigenvalues of a **G** matrix may turn out to be some of its *most* important, as these describe the amounts of usable variation that a natural population has to work with for adaptation.

Eigenvalue-based Measures of Effective Dimensionality

Our final thoughts return us once again to the eigenvalue spectrum. Thus far, we have been focusing on rank as our sole measure of the dimensionality of a matrix. While rank is indeed the critical measure of the number of dimensions of variation with positive support, it ignores any information as to the actual values of the λ_i, simply binning them as positive or otherwise. As an extreme example, consider two $\mathbf{G}$ matrices both of rank 10. $\mathbf{G}_1$ has $\lambda_1 = 100, \lambda_2 = 10, \lambda_2 = \cdots = \lambda_{10} = 0.125$, while $\mathbf{G}_2$ has $\lambda_1 = \cdots = \lambda_{10} = 11$. Both matrices have the same rank and the same eigenvalue mean $\overline{\lambda} = 11$, but clearly there are far more constraints associated with $\mathbf{G}_1$. Measures of the *dispersion* of the eigenvalues attempt to quantity this, as the more spread among the λ_i, the more potential constraints.

One early measure was **Cheverud's index of integration** (Cheverud et al 1983), which is based on the eigenvalues of the $n \times n$ *correlation* matrix $\mathbf{R}$,

$$I = 1 - \left(\prod_i^n \lambda_i \right)^{1/n} = 1 - \det(\mathbf{R})^{1/n} \tag{29.24}$$

The second step in Equation 29.24 follows by recalling that the determinant of a matrix equals the product of its eigenvalues. If all traits are uncorrelated, $\det(\mathbf{R}) = \det(\mathbf{I}) = 1$ and $I = 0$. Conversely, if $\mathbf{R}$ is singular, $\det(\mathbf{R}) = 0$ and $I = 1$. While I attempts to capture how integrated traits are (the higher the amount of correlations, the smaller $\det[\,\mathbf{R}\,]$), when the matrix is less than full rank it provides no additional information.

Wagner et al. (2008), building on the results of Wagner (1984), proposed an measure $n_{d,W}$ of the effective dimension of a correlation matrix by considering the variance of its eigenvalues. The motivation is that if $\mathbf{R} = \mathbf{I}$, all eigenvalues are one, and hence their variance is zero. In such cases the dimension is clearly n. At the other extreme, if there is only a single positive eigenvalue, then $\sigma^2(\lambda) = n - 1$ and the dimension is one (as all of the traits act as a giant single supertrait of constrained combinations). Thus, they suggest

$$n_{d,W} = n - \sigma^2(\lambda) \tag{29.25}$$

The idea of using the variance of eigenvalues appears to the intuitive notion that the more dispersed the eigenvalues, the more constraints.

A third measure of effective dimension was suggested by Kirkpatrick (2009). Here, one examines the eigenvalues associated with a *mean-standardized* covariance matrix $\mathbf{G}_m$ — each trait is divided by its mean. Such a covariance matrix can be decomposed as $\mathbf{G}_m = \mathbf{E}_v \mathbf{R} \mathbf{E}_v$, where (as above) $\mathbf{R}$ is the correlation matrix, while $\mathbf{E}_v$ is a diagonal matrix whose entries correspond to the additive genetic coefficients of variation (Houle's evolvabilities, Chapter 30). Kirkpatricks' index is given by

$$n_{d,K} = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_1} = \frac{\text{trace}(\mathbf{G})}{\lambda_1} \tag{29.26}$$

note that $1/n_{d,K} = \lambda_1/\text{tr}(\mathbf{G})$ is the fraction of variation accounted by the first eigenvalue of $\mathbf{G}_m$. Kirkpatrick defined two other summary statistics based on the iegenvalues of $\mathbf{G}_m$. The **maximum evolvability**, $\sqrt{\lambda_1}$, is the additive-genetic coefficient of variation for the trait combination with the maximal variance, while the total genetic variance of $\mathbf{G}_m$ is given by the sum of its eigenvalues,

$$\sum_{i=1}^{n} \lambda_i = \lambda_1 \, n_{d,K} \tag{29.27}$$

Using these metrics, Kirkpatrick noted that $n_{d,K}$ was between 1.1 and 1.9 over the 9 data sets he examined (those that could be mean-standardized). Thus, for mean-standardized traits, most of the genetic variation is accounted for by the leading eigenvalue ($n_{d,K} = 2$ when λ_1 accounts for 50% of the total variation). While effective dimension was very constrained over these data sets, maximal evolvability and total genetic variance were not – the former ranged from 0.033 to 1.2, the later from 0.002 to 1.7. Kirkpatrick noted that such a small value of $n_{d,K}$ does indeed imply a significant constraint. Average selection response was estimated using a procedure similar to random skewers, and Kirkpatrick noted that as the number of traits n increases with $n_{d,K}$ kept small (under 2), that the expected response substantially declines with n.

Literature Cited

Ackermann, R. R., and J. M. Cheverud. 2000. Phenotypic covariance structure in Tamarins (Genus *Saguinus*): a comparison of variation patterns using matrix correlation and common principal component analysis. *Amer. J. Phys. Ath.* 111: 489–501. [29]

Akaike, H. 1973. Information theory and the maximum likelihood principle. *In* B. N. Petrov and F. Csaki (eds) *2nd International Symposium on Information Theory*, pp. 267–281. Akademiaia Kiado, Budapest, Hungary. [29]

Amemiya, A. 1985. What should be done when an estiamted between-group covariance matrix is not nonegative definite? }si Am. Stat. 39: 112–117. [29]

Amemiya, T., T. W. Anderson, and P. A. W. Lewis. 1990. Percentage points for a test of rank in multivariate components of variance. *Biometrika* 77: 637–641. [29]

Anderson, T. W., and Y. Amemiya. 1991. Testing dimensionality in the multivariate analysis of variance. *Stat. Prob. Lett.* 12: 445–463. [29]

Arnold, S. J. 1981. Behavioral variation in natural populations. I. Phenotypic, genetic and environmental correlations between chemoreceptive responses to prey in the garter snake, *Thamnophis elegans. Evolution* 35: 489–509. [29]

Arnold, S. J., and P. C. Phillips 1999. Hierarchical comparison of genetic variance-covariance matrices. II. Coastal-inland divergence in the garter snake, *Thamnophis elegans. Evolution* 53: 1516–1527. [29]

Atchley, W. R., J. J. Rutledge, and D. E. Cowley. 1981. Genetic components of size and shape. II. Multivariate covariance patterns in the rat and mouse skull. *Evolution* 35: 1037–1055. [29]

Bégin, M., and D. A. Roff. 2001. An analysis of **G** matrix variation in two closely related cricket species, *Grylus firmus* and *G. pennsylvanicus*. *J. Evol. Biol.* 14: 1–13. [29]

Bégin, M., and D. A. Roff. 2003. The constancy of the **G** matrix through species divergence and the effects of quantitative genetic constraints on phenotypic evolution: a case study in crickets. *Evolution* 57: 1107–1120. [29]

Bégin, M., D. A. Roff, and V. Debat. 2004. The effect of temperature and wing morphology on quantitative genetic variation in the cricket *Gryllus firmus*, with an appendix examining the statistical proerties of the Jackknife-MANOVA method of matrix comparison. *J. Evol. Biol.* 17: 1255–1267. [29]

Bhargava, A. K., and D. Disch. 1982. Exact probabilities of obtaining estimated non-positive definite between-group covariance matrices. *J. Stat. Comp. Sim.* 15: 27–32. [29]

Billington, H. L., A. M. Mortimer, and T. McNeilly. 1988. Divergence and genetic structure in adjacent grass populations. I. Quantitative genetics. *Evolution* 42: 1267–1277. [29]

Blows. M. W. 2007. A tale of two matrices: multivariate approaches in evolutionary biology. *J. Evol .Biol.* 20: 1–8. [29]

Blows, M. W., S. F. Chenoweth, and E. Hine. 2004. Orientation of the genetic variance-covariance matrix and fitness surface for multiple male sexually selected traits. *Amer. Nat.* 163: 329–340. [29]

Blows, M. W., and M. Higgie. 2003. Genetic constraints on the evolution of mate recognition under natural selection. *Amer. Natl.* 161; 240–253. [29]

Bonnin, I., J. M. Prosperia, and I. Olivieri. 1997. Comparison of quantitative genetic parameters between two natural populations of a selfing plant species, *Medicago truncatula* Gaertn. *Theor. Appl. Genet* 94: 641–651. [29]

Brodie, E. D. III. 1993. Homogeneity of the genetic variance-covariance matrix for antipredator traits in two natural populations of the garter snake *Thamnophis ordinoides*. *Evolution* 47: 844–854. [29]

Camara, M. D., and M. Pigluicci. 1999. Mutational contributions to genetic variance-covariance matrices: an experimental approach using induced mutations in *Arabidopsis thaliana. Evolution* 53: 1692–1703. [29]

Cano, J. M., A. Laurila, J. Palo, and J. Meril: a. 2004. Population differentiation in **G** matrix structure due to natural selection in *Rana temporaria*. *Evolution* 58: 2013–2020. [29]

Carr, D. E., and C. B. Fenster. 1994. Levels of genetic variation and covariation for *Mimulus* (Scrophulariaceae) floral traits. *Heredity* 72: 606–618. [29]

Caruso, C. M., H. Maherali, A. Mikulyuk, K. Carlson, and R. B. Jackson. 2005. Genetic variance and covariance for physiological traits in *Lobelia:* are there constraints on adapative evolution? *Evolution* 59: 826–837. [29]

Cheverud, J. M. 1996. Quantitative genetic analysis of cranial morphology in the cotton-top (*Saguinus oedipus*) and saddle-back (*S. fuscicollis*) tamarins. *J. Evol. Biol.* 9: 5–42. [29]

Cheverud, J., J. Rutledge, and W. Atchley. 1983. Quantitative Genetics of Development: Genetic Correlations Among Age-Specific Trait Values and the Evolution of Ontogeny. *Evolution* 37: 895–905. [29]

Cheverud, J. M., and G. Marroig. 2007. Comparing covariance matrices: random skewers method compared to the common principal components model. *Genet. Mol. Biol.* 30: 461–469. [29]

Cheverud, J. M. , G. P. Wagner, and M. M. Dow. 1989. Methods for the comparative analysis of variation patterns. *Syst. Zool.* 38: 201–213. [29]

Cowley, D. E., and W. R. Atchley. 1990. Development and quantitative genetics of correlation structure among body parts of *Drosophila melanogaster*. *Amer. Natl.* 135: 242–268. [29]

Cowley, D. E., and W. R. Atchley. 1992. Comparison of quantitative genetic parameters. *Evolution* 46: 1965–1967. [29]

Dickerson, G. E. 1955. Genetic slippage in response to selection for multiple objectives. *Cold Spring Harb Symp Quant Biol.* 20: 213–224. [29]

Dietz, E. J. 1983. Permutation tests for association between two distance matrices. *Syst. Zool.* 32: 21–26. [29]

Donoghue, K., D. Messiqua, E. H. Pyle, M. S. Heschel, and J. Schmitt. 2000. Density dependence and population differentiation of genetic architecture in *Impatiens capensis* in natural environments. *Evolution* 54: 1969–1981. [29]

Doroszuk, A., M. W. Mojewodzic, G. Gort, and J. E. Kammenga. 2008. Rapid divegence of genetic variance-covariance matrix within a natural population. *Amer. Nat.* 171: 281–304. [29]

Estes, S., and P. C. Phillips. 2006. Variation in pleiotropy and the mutational underpinnings of the **G**-matrix. *Evolution* 60: 2655–2660. [29]

Flury, B. 1987. A hierarch of relationships between covariance matrices, *In* A. K. Gupta (ed) *Advances in multivariate statistical analysis*, pp. 31–43. Reidel, Boston. [29]

Flury, B. 1988. *Common principal components and related multivariate models.* Wiley, New York. [29]

Fong, D. W. 1989. Morphological evolution of the amphipod *Gammarus minus* in caves: quantitative genetic analysis. *Am. Midl. Nat.* 121: 361–378. [29]

Goodnight, C. J., and J. M. Schwartz. 1997. A bootstrap comparison of genetic covariance matrices. *Biometrics* 53: 1026–1039. [29]

Griswold, C. K., B. Logsdon, and R. Gomulkiewicz. 2007. Neutral evolution of multiple quantatitive characters: a genealogical approach. *Genetics* 176: 455–466. [29]

Hansen, T. F., and D. Houle. 2008. Measuring and comparing evolvability and constraint in multivariate characters. *J. Evol. Bio.* 21: 1201–1219. [29]

Hill, W. G. and R. Thompson. 1978. Probabilities of non-positive definite between group or genetic covariance matrices. *Biometrics* 34: 429–439. [29]

Hine, E., and M. W. Blows. 2006. Determining the effective dimensionality of the genetic variance-covariance matrix. }si Genetics 173: 1135–1144. [29]

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometricka* 75: 800-802. [29]

Holloway, G. J., P. W. de Jong, and M. Ottenheim. 1993. The genetics and cost of chemical defense in the two-spot ladybird (*Adalia bipunctata* L.). *Evolution* 47: 1229–1239. [29]

Holm, S. 1979. A simple sequential rejection multiple test procedure. *Scand. J. Stat.* 6: 65–70. [13]

Hommel, G. 1989. A comparison of two modified Bonferonii procedures. *Biometrika* 76: 624–625. [29]

Houle, D., J. Mezey, and P. Galperin. 2002. Interpretation of the results of common principal components analysis. *Evolution* 56: 433–440. [29]

Hubert, L. J. 1983. Inference procedures for the evaluation and comparison of proximity matrices. *In* J. Felsenstein (ed.), *Numerical taxonomy*, pp. 209–228. Springer-Verlag, Berlin. [29]

Kirkpatrick M. 2009. Patterns of quantitative genetic variation in multiple dimensions. *Genetica* In press.

Kirkpatrick, M., and D. Lofsvold. 1992. Measuring selection and constraint in the evolution of growth. *Evolution* 46: 954–971. [29]

Kirkpatrick, M., D. Lofsvold, and M. Bulmer. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* 124: 979–993. [31]

Kirkpatrick, M., and K. Meyer. 2004. Direct estimation of genetic principal components: simplified analysis of complex phenotypes. *Genetics* 168: 2295–2306. [29]

Klingenberg, C. P., B. E. Neuenschwander, and B. D. Flury. 1996. Ontogeny and individual variation: analysis of patterned covariance matrices with common principal components. *Syst. Biol.* 45: 135–150. [29]

Klotz, J., and J. Putter. 1969. Maximum likelihood estimation of multivariate covariance components for the balanced one-way layout. *Ann. Math. Stat.* 40: 1100–1105. [29]

Kohn, L . A., and W. R. Atchley. 1988. How similar are genetic correlation structures? Data from mice and rats. *Evolution* 42: 467–481. [29]

Kuriki, S. 1993. One-sided test for equality of two covariance matrices. *Biometrika* 43: 128–136. [29]

Krzanowski, W. J. 1979. Between-group comparisons of principal components. *J. Amer. Stat. Assoc.* 74: 703–707. [29]

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. *Evolution* 33: 402–416. [29]

Lande, R. 1980. Genetic variation and phenotypic evolution during allopatric speciation. *Amer. Natl.* 116: 463–479. [29]

Lawley, D. N. 1956. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika* 43: 128–136. [29]

Lofsvold, D. 1986. Quantitative genetics of morpholoigcal differentiation in *Peromyscus.* I. Tests of homogeneity of genetic covariance structure among species and subspecies. *Evolution* 40: 559–573. [29]

López-Fanjul, C., A. Fernández, and M. A. Toro. 2004. Epistasis and the temporal change in the additive variance-covariance matrix induced by drift. *Evolution* 58: 1655–1663. [29]

Manly, B. F. J. 1991. *Randomization and Monte Carlo methods in biology.* Chapman and Hall, London. [29]

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209–220. [29]

Marroig, G., and J. M. Cheverud. 2001. A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology, and ontogeny during cranial evolution of new world monkeys. *Evolution* 55: 2576–2600. [29]

McGuigan, K., and M. W. Blows. 2007. The phenotypic and genetic covariance structure of Drosphilid wings. *Evolution* 61: 902–911. [29]

Meyer, K. 2005. Genetic principal components for live ultrasound scan traits of Angus cattle. *Anim. Sci.* 81: 337–345. [29]

Meyer, K. 2005. Multivariate analyses of carcass traits for Angus cattle fitting reduced rank and factor analytic models. *J. Anim. Bred. Genet.* 1241: 50–64. [29]

Meyer, K., and M. Kirkpatrick. 2005. Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. *Gen. Sel. Evol* 37: 1–30. [29]

Meyer, K., and M. Kirkpatrick. 2007. A note on bias in reduced rank estimates of covariance matrices. *Proc. Assoc. Adv.. Anim. Breed. Genet.* 17: 154–157. [29]

Meyer, K., and M. Kirkpatrick. 2008. Perils of parsimony: properties of reduced-rank estimates of genetic covariance matrices. *Genetics* 180: 1153–1166. [29]

Mezey, J. G., and D. Houle. 2005. The dimensionality of genetic variation for wing shape in *Drosophila melanogaster*. *Evolution* 59: 1027–1038. [29]

Miller, R. G. 1974. The jackknife – a review. *Biometrika* 61: 1–15. [29]

Paulsen, S. M. 1996. Quantitative genetics of the wing color pattern in the buckeye butterfly (*Precis coenia* and *Precis evarete*): evidence against the constancy of **G**. *Evolution* 50: 1585–1597. [29]

Petfield, D., S. F. Chenoweth, H. D. Rundle, and M. W. Blows. 2005. Genetic variance in female condition predicts indirect genetic variance in male sexual display traits *PNAS* 102: 6045–6050. [29]

Pease, C. M., and J. J. Bull. 1988. A critique of methods for measuring life history trade-off. *J. Evol. Biol.* 1: 293–303. [29]

Pfrender, M. W., and M. Lynch. 2000. Quantitative genetic variation in *Daphina*: temporal changes in genetic architecture. *Evolution* 54: 1502–1509. [29]

Phillips, P. C., and S. J. Arnold. 1999. Hierarchical comparison of genetic variance-covariance matrices. I. Using the Flury hierarchy. *Evolution* 53: 1506–1515. [29]

Phillips, P. C., M. C. Whitlock, and K. Fowler. 2001. Inbreeding changes the shape of the genetic covariance matrix in *Drosophila melanogaster*. *Genetics* 158: 1137–1145. [29]

Pielou, E. C. 1984. Probing multivariate data with random skewers: a preliminary to direct gradient analysis. sio Oikis 42: 161–165. [29]

Platenkamp, G. A. J., and R. G. Shaw. 1992. Enviromental and genetic constraints on adapative population differentiation in *Anthoxanthum odoratum*. *Evolution* 46: 341–352. [29]

Podolsky, R. H., R. G. Shaw, and F. H. Shaw. 1997. Population structure of morphological traits in *Clarkia dudleyana*. II. constancy of within-population genetic variance. *Evolution* 51: 1785–1796. [29]

Remadi, S., and Y. Amemiya. 1994. Asymptotic properties of the estimators for multivariate components of variance. *J. Multivar. Anal.* 49: 110–131. [29]

Roff, D. 1997. *Evolutionary quantitative genetics.* Chapman and Hall. [29]

Roff, D. 2000. The evolution of the G matrix: selection or drift? *Heredity* 84: 135–142. [29]

Roff, D. 2002. Comparing **G** matrices: a MANOVA approach. *Evolution* 56: 1286–1291. [29]

Roff, D. A., and T. A. Mousseau. 1999. Does natural selection alter genetic architecture? An evaluation of the quantitative genetic variation among populations of *Allonemobius socius* and *A. fasciatus*. *J. Evol. Biol.* 12: 361–369. [29]

Roff, D. A., T. A. Mousseau and D. J. Howard. 1999. Variation in genetic architecture of calling song among populations of *Allonemobius socius*, *A. fasciatus*, and a hybrid population: drift or selection? *Evolution* 53: 216–224. [29]

Roff, D. A., and R. Preziosi. 1994. The estimation of the genetic correlation: the use of the jackknife. *Heredity* 73: 544–548. [29]

Rundle, H. D., S. E. Chenoweth, and M. W. Blows. 2008. Comparing complex fitness surfaces: among-population variation in mututal sexual selection in *Drosophla serrata. Amer. Natl.* 171: 443–454. [29]

Schluter, D. 1996. Adaptive radiation along genetic lines of least resistance. *Evolution* 50: 1766–1774. [29]

Service, P. M. 2000. The genetic structure of female life history in *D. melanogaster*: comparisons among populations. *Genet. Res. Camb.* 75: 153–166. [29]

Shaw, R. G. 1991. The comparison of quantitative genetic parameters between populations. *Evolution* 45: 143–151. [29]

Shaw, R. G. 1992. Comparison of quantitative genetic parameters: Reply to Cowley and Atchley. *Evolution* 46: 1967–1969. [29]

Shaw, R. G., and H. L. Billington. 1991. Comparison of variance componente between two populations of Holcus lanatus: a reanalysis. *Evolution* 45: 1287-1289. [29]

Simes, J. R. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73: 751–754. [29]

Simons, A. M., and D. A. Roff. 1994. The effect of environmental variability on the heritabilities of traits of a field cricket. *Evolution* 48: 1637–1649. [29]

Spitze, K., J. Burnson, and M. Lynch. 1991. The covariance structure of life-history characters in *Daphina pulex. Evolution* 45: 1081–1090. [29]

Steppan, S. J. 1997a. Phylogenetic analysis of phenotypic covariance structure. I. contrasting results from matrix correlation and common principal component analysis. *Evolution* 51: 571–586. [29]

Steppan, S. J. 1997b. Phylogenetic analysis of phenotypic covariance structure. I.I. Reconstructing matrix evolution. *Evolution* 51: 587–594.. [29]

Steppan, S. J., P. C. Phillips, and D. Houle. 2002. Comparative quantitative genetics: evolution of the **G** matrix. *Trends Ecol. Evol.* 17: 320–327. [29]

Stinchcombe, J. R., and J. Schmitt. 2006. Ecosystem enginners as selective agents: the effects of leaf litter on emergence time and early growth in *Impatiens capensis. Ecol. Let.* 9: 258–270. [29]

Stinchcombe, J. R., C. Weinig, K. D. Health, M. T. Brock, and J. Schmitt. 2009. Polymorphic genes of major effect: consequences for variation, selection, and evolution in *Arabidopsis thaliana.* **SUBMITTED** [29]

Tabachnick, B. G., and L. S. Fidell. 2006. *Using multivariate statistics*, 5th Edition. Allyn and Bacon, Boston. [29]

Tukey, J. W. 1958. Bias and confidence in not quite large samples. *Ann. of Math. Stati.* 29: 614. [29]

Tukey, J. W. 1986. Sunset salvo. *Amer. Stat* 40: 72–76. [29]

Turelli, M. 1988. Phenotypic evolution, constant covariances, and the maintenance of additive variance. *Evolution* 42: 1342–1347. [29]

Wagner, G. P. 1984. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: evidence for a nonrandom organization of quantitative character variation. *J. Math. Biol.* 21: 77–95. [29]

Wagner G.P., J. P. Kenney-Hunt, M. Pavlicev, J. R. Peck, D. Waxmman, and J. M. Cheverud. 2008. Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature* 452: 470–474. [29]

Waldmann, P., and S. Andersson. 2000. Comparison of genetic (co)variance matrices within and between *Scabiosa canescens* and S. columbaria. *J. Evol. Biol.* 13: 826–835. [29]

Whitlock, M. C., P. C. Phillips, and K. Fowler. 2002. Persistence of changes in the genetic covariance matrix after a bottleneck. *Evolution* 56: 1968–1975. [29]

Willis, J. H., J. A. Coyne, and M. Kirkpatrick. 1991. Can one predict the evolution of quantitative characters without genetics? *Evolution* 45: 441–444. [29]

Wright, S. 1951. The genetic structure of populations. }si Ann. Eogen. 15: 323–354. [29]

Zhang, J., and D. D. Boos. 1992. Bootstrap critical values for testing homogeneity of covariance matrices. *J. Am. Stat. Assn.* 87: 425–429. [29]

Zhang, J., and D. D. Boos. 1993. Testing hypothesis about covariance matrices using bootstrap methods. *Comm. Stat. Theory Meth.* 22: 723–739. [29]