© 2015 by The University of Chicago. All rights reserved. DOI: 10.1086/681083

Appendix B from J. G. Kingsolver et al., “Genetic Variation, Simplicity, and Evolutionary Constraints for Function-Valued Traits” (Am. Nat., vol. 185, no. 6, p. 000)

Assessing Variability

We can use either a frequentist or a Bayesian approach to assess the uncertainty of an estimate of **G** or of quantities associated with its principal components analysis (PCA) or simple basis analysis (SBA). Here, we take a frequentist approach, expressing uncertainty in terms of variability and giving approximate standard errors. We will use a subset of the growth curve data set for *Tribolium castaneum* (Irwin and Carter 2013), comprising records at six ages, to illustrate our approach. Of particular interest is quantifying variability in the percentage of variance explained by each principal component (PC) and simple basis (SB), in the loadings of the PCs, and in the loadings for SBs in the nearly null space. We also illustrate how to estimate correlations between SB scores, and we give their associated standard errors.

We derive approximate standard errors using the theory of large-sample normal approximations of the distribution of an estimate. These approximations are based on the asymptotic covariance matrix of the estimated parameters, which is in turn based on the inverse of the Fisher information matrix. For our case of estimating components of **G**, the required asymptotic covariance matrix is readily available from software like SAS, ASreml, and Wombat. In principle, we can find expressions for the standard error of any quantity calculated from an estimate of **G** by using this asymptotic covariance matrix and the delta method (Oehlert 1992). However, we use a method that is more straightforward to apply: we randomly generate 10,000 **G** matrices, with elements following a multivariate normal distribution. The mean of the distribution of the **G**s is equal to $\hat{\mathbf{G}}$, and the variance/covariance matrix depends on the Fisher information (Meyer and Houle 2013). When a generated matrix is nonnegative definite, we modify it by setting any negative eigenvalues equal to 0. For each of these 10,000 randomly constructed and then modified **G** matrices, we calculate the quantity of interest. Our standard error is the standard deviation of these 10,000 calculated quantities.

As an example of a standard error calculation, consider the estimate of the genetic variance explained by $\mathbf{s}_1$, the first vector in the SB. This variance is $\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{s}_1$ (Lin and Allaire 1977); as a proportion of the total genetic variance, this is $\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{s}_1/\mathrm{tr}\,\hat{\mathbf{G}}$, where $\mathrm{tr}\,\hat{\mathbf{G}}$ is the trace of $\hat{\mathbf{G}}$. To calculate the standard error of $\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{s}_1/\mathrm{tr}\,\hat{\mathbf{G}}$, we calculate $\mathbf{s}_1^{\mathrm{T}}\mathbf{G}\mathbf{s}_1/\mathrm{tr}\,\mathbf{G}$ for each of the 10,000 randomly generated **G** matrices. Our standard error is simply the standard deviation of these 10,000 values of $\mathbf{s}_1^{\mathrm{T}}\mathbf{G}\mathbf{s}_1/\mathrm{tr}\,\mathbf{G}$. With our data, we calculated a standard deviation equal to 4.8% as a measure of the variability of 51.2%, the percentage of genetic variance explained by the first SB. A similar calculation yields a standard error of 5.5% for the estimate 77.4%, the percentage of variability explained by the first PC. In addition to summarizing our 10,000 values with standard errors, we can display the values using boxplots. Figure B1 contains boxplots of cumulative percentages of variances explained in PCA and SBA.

To assess the variability in the first PC vectors and in the simplest vectors in the nearly null space, we could calculate componentwise standard errors or construct componentwise boxplots. However, we find it instructive instead to plot some of the 10,000 vectors that result from our analysis of the 10,000 generated **G** matrices (figs. B2, B3). For instance, in the left panel of figure B2, we plot the first PC vectors as a function of age, obtained from the first 100 randomly generated **G**s. We see a clear peak in all 100 vectors, with most peaks occurring at age 6 days.

As can be seen in figure B1 by noting the position of the cutoff of 98%, the dimension of the nearly null space varies among the 10,000 generated **G**s. This variation in dimension can make it challenging to interpret the simplest direction in the nearly null space. To convey the limits on interpretation, we suggest making several plots of SB vectors, as in figure B3. For each of the two nearly null spaces depicted there, we easily see the shape of the simplest basis vector, but we cannot easily decide between the two plots. This is a natural and expected occurrence for this data set—the best we can do is interpret and describe accordingly.

Table B1 gives estimates of the correlation between two genotypes’ SB scores along with standard errors. To define the estimate of the correlation, we use variance/covariance rules as follows. As an example, consider the first and second simplicity scores. Suppose that the loadings defining the first SB vector are in the vector $\mathbf{s}_1 = (\mathbf{s}_1[1], \ldots, \mathbf{s}_1[p])^{\mathrm{T}}$, and the components of an individual’s genotype vector are denoted $\mathbf{g}[1], \ldots, \mathbf{g}[p]$. Then the individual’s first simplicity score is simply the weighted sum of $\mathbf{g}[1], \ldots, \mathbf{g}[p]$, that is, the first simplicity score is equal to $\mathbf{s}_1[1]\mathbf{g}[1] + \cdots + \mathbf{s}_1[p]\mathbf{g}[p]$. From variance/covariance rules, the variance of this score is equal to $\mathbf{s}_1^{\mathrm{T}}\mathbf{G}\mathbf{s}_1$, where **G** is the genetic variance/covariance matrix.

We estimate $\mathbf{s}_1^{\mathrm{T}}\mathbf{G}\mathbf{s}_1$ by substituting $\hat{\mathbf{G}}$ for $\mathbf{G}$. Thus, our estimate of the variance of the first simplicity scores is $\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{s}_1$. Similarly, we can estimate the covariance between scores from, say, the two SB vectors, $\mathbf{s}_1$ and $\mathbf{s}_2$. By variance/covariance rules, the true covariance between the two scores is equal to $\mathbf{s}_1^{\mathrm{T}}\mathbf{G}\mathbf{s}_2$. We estimate this by $\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{s}_2$. We use these variance and covariance estimates to estimate the correlation between the two SB scores: $\hat{\rho}_{12} = \mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{s}_2/(\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{s}_1\, \mathbf{s}_2^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{s}_2)^{1/2}$. The standard error of $\hat{\rho}_{12}$ is simply the standard deviation of the 10,000 correlations calculated from our 10,000 randomly generated **G**s.

The validity of our method relies on standard large sample properties of maximum likelihood estimates and restricted maximum likelihoods, properties that guarantee that the information matrix can be used to assess variability. In particular, the method cannot be expected to perform well if the original sample sizes are small. In addition, if any parameter estimates are near "the boundary" of the parameter space, a larger original sample size may be required for the large sample theory to remain valid. In the case we consider here, our estimate of **G** has eigenvalues close to 0, that is, near the "boundary" of the parameter space for **G**. We suspect that this is not a serious problem; while most of our randomly generated **G**s had negative eigenvalues, these were extremely small. Setting the negative eigenvalues equal to 0 had little impact on our results. While our method of calculating standard errors is supported by asymptotic theory, the user should be aware of these caveats.

In contrast to our frequentist approach to assess uncertainty, in the Bayesian approach one places a prior distribution on **G** and then uses the posterior distribution of **G** given the data to estimate **G** and other quantities involving **G**. The Bayes estimate of a parameter is typically equal to the mean or mode of the parameter with respect to its posterior distribution. The posterior distribution is also used to calculate credible intervals for these parameters, thus providing an assessment of uncertainty in the estimation process. If the posterior distribution is simple in form, means and intervals can be calculated explicitly. When the posterior distribution is of a complicated form, as is often the case, computational methods are required to calculate posterior means and credible intervals. The most popular computational technique is Markov chain Monte Carlo simulation, which generates a large number of **G**s from the posterior distribution. These **G**s can be used directly to calculate posterior means and credible intervals for components of the true **G** and for quantities of interest calculated from the true **G**. For a discussion of Bayes techniques in assessing variability, see Sorenson and Gianola (2002), Hadfield (2010), and Stinchcombe et al. (2014).

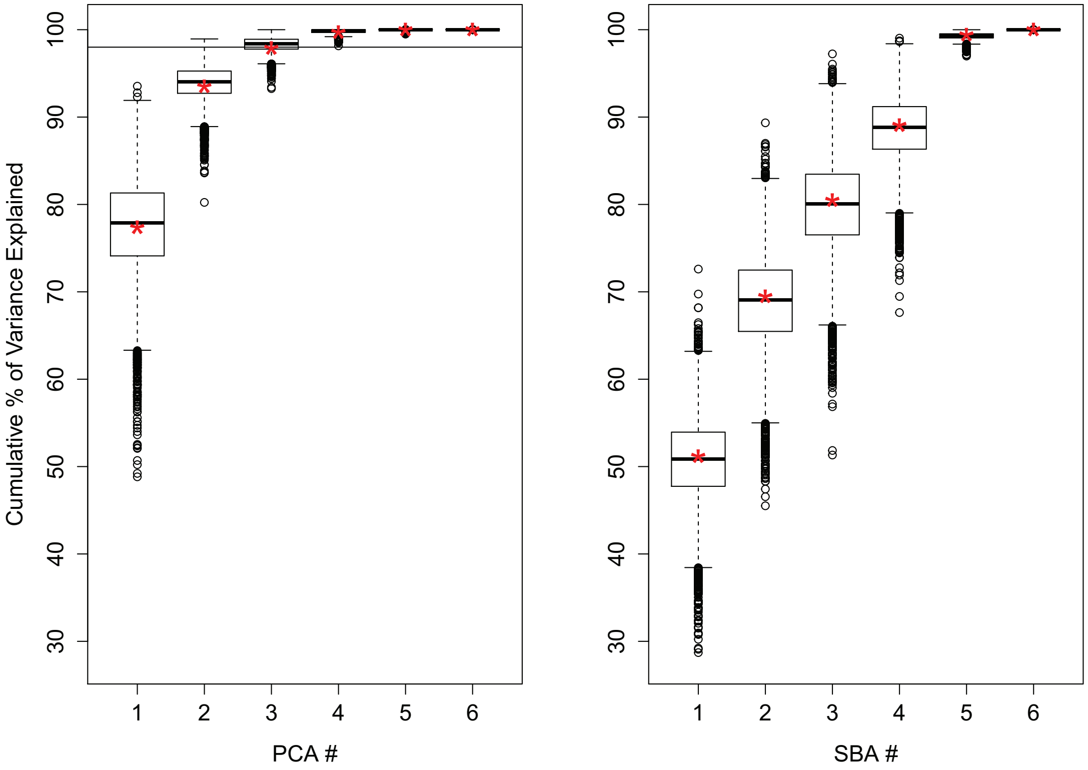

Figure B1: The left panel contains boxplots showing the cumulative percentage of variance explained by principal components basis vectors obtained from the 10,000 simulated **G** matrices. Red asterisks indicate the cumulative percentage using the original estimate of **G**. The horizontal line at 98% indicates the cutoff of percentage of variance explained that is used to define the nearly null space. The right panel shows the cumulative percentage of variance explained by the full simple basis. Again, red stars indicate the cumulative percentage using the original estimate of **G**. While the values in some of the boxplots have a wide range (e.g., from around 48% to 98% for PCA1), the variability of the values is better measured by the height of the boxplots (the interquartile range) or the standard deviations, which must be calculated separately (see the main text). PCA = principal components analysis; SBA = simple basis analysis.

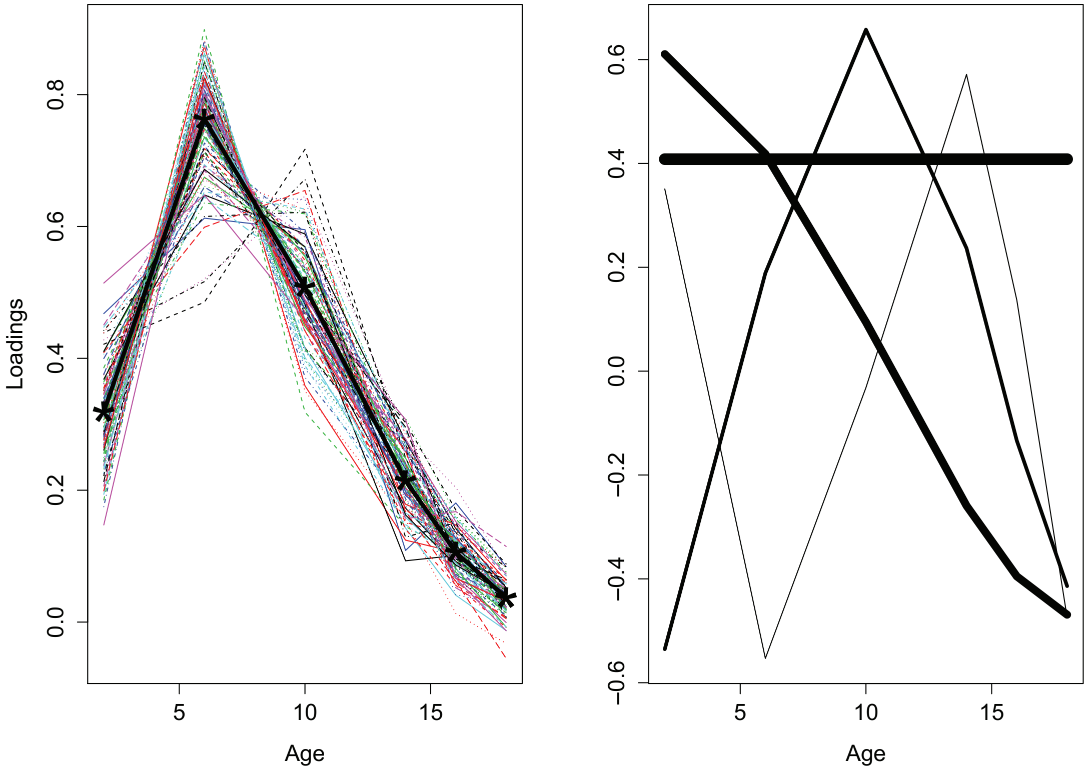

Figure B2: The left panel contains the loadings of the first principal component as a function of age (in days) for 100 of the 10,000 simulated **G** matrices. The heavy black line and asterisks show the loadings for the **G** estimated from the data. The right panel contains the loadings of the first four simple basis vectors, with line thickness ordered by simplicity: the loadings of the simplest basis vector are displayed with the thickest line.

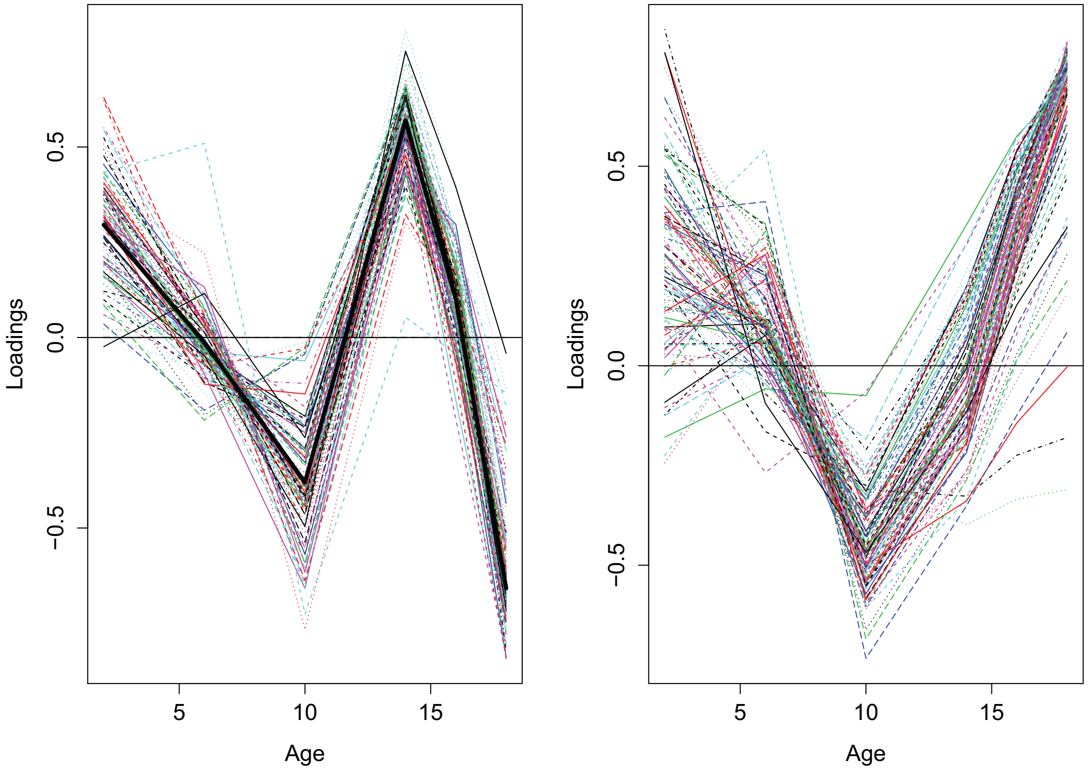

Figure B3: Loadings, as a function of age in days, of the simplest vector in the nearly null space, based on the 10,000 simulated **G** matrices. The dimensions of the nearly null spaces were determined so that the model space explained at least 98% of the variance. The left panel displays loadings of vectors from 100 of the 3,328 simulated **G** matrices that produced two-dimensional nearly null spaces. The right panel displays loadings of vectors from 100 of the 6,628 simulated **G** matrices that produced three-dimensional nearly null spaces. The heavy black line in the left panel shows the loadings for the simplest vector in the nearly null space for the **G** estimated from the data. We do not display the vectors for the remaining 44 of the 10,000 **G** matrices. These vectors all correspond to nearly null spaces of dimension 4.

Table B1: Estimated correlations among the six simplicity scores, with standard errors in parentheses

	SB1	SB2	SB3	SB4	SB5	SB6
SB1	1.00	.86 (.05)	.81 (.07)	−.45 (.16)	−.20 (.20)	.13 (.27)
SB2	...	1.00	.70 (.12)	−.59 (.15)	−.28 (.21)	.11 (.29)
SB3	...	...	1.00	−.61 (.19)	−.46 (.22)	−.22 (.30)
SB4	...	...	...	1.00	.89 (.08)	.35 (.31)
SB5	...	...	...	...	1.00	.57 (.27)
SB6	...	...	...	...	...	1.00

Note: SB = simple basis.

Literature Cited Only in Appendix B

Lin, C. Y., and F. R. Allaire. 1977. Heritability of a linear combination of traits. Theoretical and Applied Genetics 51:1–3.

Oehlert, G. W. 1992. A note on the delta method. American Statistician 46:27–29.

Sorenson, D., and D. Gianola. 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer, New York.