Estimating sampling error of evolutionary statistics based on genetic covariance matrices using maximum likelihood

David Houle
Department of Biological Science
Florida State University
Tallahassee, Florida 32308, USA
Email: dhoule@bio.fsu.edu

Karin Meyer
Animal Genetics and Breeding Unit,
University of New England,
Armidale, NSW 2351, AUSTRALIA
Email: kmeyer.agbu@gmail.com

Forthcoming in the Journal of Evolutionary Biology. Doi: 10.1111/jeb.12674

Data available from the Dryad Digital Repository:
http://dx.doi.org/10.5061/dryad.hb37q

Abstract

We explore the estimation of uncertainty in evolutionary parameters using a recently devised approach for resampling entire additive genetic variance-covariance matrices (**G**). Large sample theory shows that maximum likelihood estimates (including restricted maximum likelihood, REML) asymptotically have a multivariate normal distribution, with covariance matrix derived from the inverse of the information matrix, and mean equal to the estimated **G**. This suggests that sampling estimates of **G** from this distribution can be used to assess the variability of estimates of **G**, and of functions of **G**. We refer to this as the REML-MVN method. This has been implemented in the mixed model program Wombat. Estimates of sampling variances from REML-MVN were compared to those from the parametric bootstrap and from a Bayesian Markov chain Monte Carlo (MCMC) approach (implemented in the R package MCMCglmm). We apply each approach to evolvability statistics previously estimated for a large, 20-dimensional data set for Drosophila wings. REML-MVN and MCMC sampling variances are close to those estimated with the parametric bootstrap. Both slightly underestimate the error in the best-estimated aspects of the **G** matrix. REML analysis supports the previous conclusion that the **G** matrix for this population is full-rank. REML-MVN is computationally very efficient, making it an attractive alternative to both data resampling and MCMC approaches to assessing confidence in parameters of evolutionary interest.

Keywords: G matrix, quantitative genetics, evolution, restricted maximum likelihood, evolvability, sampling error

Introduction

The evolutionary properties of sets of phenotypic traits in outbred populations are summarized by the additive genetic variance-covariance matrix, **G** (Lande, 1979). When paired with an estimate of the strength and direction of selection, **G** predicts the rate and direction of evolution. As a result, **G** matrix estimates are essential elements in a wide variety of evolutionary statistics that quantify such features as the ability of a population to respond to directional selection on multiple traits (Lande, 1979, Cheverud, 1996, Hansen & Houle, 2008), the degree of modular structure to variation, and how variation of evolution is spread across phenotypic dimensions (Mezey & Houle, 2005, Hine & Blows, 2006, Kirkpatrick, 2009, Houle & Fierst, 2013). A related set of methods focuses on comparison of the evolutionary potential of different populations (Kirkpatrick, 2009, Cheverud, 1996, Cheverud & Marroig, 2007, Krzanowski, 1979, Houle & Fierst, 2013, Hansen & Houle, 2008, Aguirre et al., 2014, Hine et al., 2009).

While calculating estimates of such statistics is straightforward, assessing the sampling properties of these statistics is much more challenging. The first step is always to identify a set of **G** matrices consistent with sampling variation of the original data. Once this is done, the sampling variation of functions of **G** can then be estimated by applying the function to these sample matrices. For many years, data resampling methods, such as bootstrapping or jackknifing (e.g., Phillips & Arnold, 1999, Mezey & Houle, 2005, Hine et al., 2009) have been the major tool for generating such families of estimates. Since estimation of **G** matrices is generally computationally demanding, data resampling can be prohibitively time-consuming. The rise of numerical Bayesian estimation using Markov chain Monte Carlo (MCMC) methods (Gelman et al., 2013, Hadfield, 2010) and their increasing application to quantitative genetics (Sorensen & Gianola, 2002, O'Hara et al., 2008, Ovaskainen et al., 2008, Aguirre et al., 2014, Stinchcombe et al., 2014) has provided a simpler general route to the assessment of the uncertainty in evolutionary characteristics. In MCMC methods, the estimation of a **G** matrix proceeds by estimating the distribution of **G** matrices consistent with the data. The samples from this posterior distribution are then used to estimate variation in evolutionary statistics (e.g. Aguirre et al. 2014). MCMC approaches can also be computationally demanding, and therefore difficult to apply to data sets with large numbers of parameters and large sample sizes.

Meyer and Houle (2013) recently proposed an alternative method for sampling entire **G** matrices based on Restricted Maximum Likelihood (REML). Provided large sample theory holds, the sampling distribution of the parameters of **G** approaches a multivariate normal distribution with covariance matrix given by the inverse of the information matrix. Values of **G** can be readily sampled from this distribution. This approach has been implemented in the mixed model program Wombat (Meyer, 2010-2015). We call this the REML-MVN method. A similar general approach has been suggested by Mandel (2013). Meyer & Houle (2013) compared estimates of sampling variances from REML-MVN with those based on simulated data drawn from the same distribution, and obtained close agreement. They showed that confidence intervals from REML-MVN were more accurate than those based on the Delta method (Oehlert, 1992) for parameters near their boundaries, such as genetic correlations approaching unity. Kingsolver et al. (2015) used REML-MVN to estimate variation in decompositions of **G** for function-valued traits.

In this contribution, we demonstrate estimation of evolutionary statistics using REML-MVN for data from a large, high-dimensional data set on wing shape variation in *Drosophila melanogaster* (Mezey & Houle, 2005). Hansen and Houle (2008) previously estimated measures of evolvability for these data. The addition of confidence limits to their analysis allows us to assess the robustness of their conclusions. We compare these error estimates to those estimated using the parametric bootstrap and MCMC.

Sampling G matrices based on REML estimates

The Restricted Maximum Likelihood multivariate normal (REML-MVN) sampling approach relies on the result that the distribution of maximum likelihood estimates asymptotically approaches a multivariate normal distribution as sample size increases. Let $\boldsymbol{\theta}$ denote the vector of parameters to be estimated, e.g. the $k(k+1)/2$ distinct elements of a covariance matrix $\mathbf{G}$. The covariance matrix of the estimates is approximated by the inverse of the information matrix, denoted as $\mathbf{H}(\boldsymbol{\theta})$. If the vector of estimates at convergence is $\hat{\boldsymbol{\theta}}$, then the distribution of $\hat{\boldsymbol{\theta}}$ is $N\left(\hat{\boldsymbol{\theta}},\mathbf{H}\left(\hat{\boldsymbol{\theta}}\right)\right)$.

REML estimates of covariances matrices are constrained to the parameter space, i.e. forced to have non-negative eigenvalues throughout so that they are positive semi-definite. Most REML software enforces this by re-parameterizing to estimate the elements of the Cholesky factors of covariance matrices, the elements of the lower triangular matrix $\mathbf{L}$ for $\mathbf{G} = \mathbf{L}\,\mathbf{L}'$. In addition, positive diagonal elements of $\mathbf{L}$ are ensured by transforming them to logarithmic scale (Meyer & Smith, 1996). On completion of the analysis, a `valid' estimate of $\mathbf{G}$ is obtained by reversing the transformation. Asymptotic normality of $\hat{\boldsymbol{\theta}}$ holds on either scale.

This then presents the possibility of using the multivariate normal sampling approach on two different scales; on the G-scale we can use multivariate normality to directly sample the elements of $\mathbf{G}$ (with vector of estimates $\boldsymbol{\theta}_{\mathbf{G}}$), while on the L-scale we can sample the elements of $\mathbf{L}$ (with vector of estimates $\boldsymbol{\theta}_{\mathbf{L}}$), and use those to construct estimates of $\mathbf{G}$. More formally, we can generate **G** matrix values, denoted $\hat{\mathbf{G}}$, drawn from the sampling distribution of **G**, denoted $\tilde{\mathbf{G}}$, by sampling the elements of $\hat{\mathbf{G}}$, or by sampling the elements of $\hat{\mathbf{L}}$.

Sampling $\boldsymbol{\theta}_G$ directly attempts to approximate the large sample distribution of **G**, similar to what MCMC typically does, albeit for different distributions. There is, however, a key difference between G-sampling and MCMC in that sampling on the G-scale does not guarantee that samples $\hat{\mathbf{G}}$ are positive semi-definite, i.e. we may obtain values outside of the parameter space, especially for matrices with eigenvalues close to the boundary. In contrast, MCMC algorithms typically sample a sum-of-squares and cross-products matrix guaranteed to be positive definite. Sampling on the G-scale will yield a mean of the $\tilde{\mathbf{G}}$ across samples equal to the REML estimate $\hat{\mathbf{G}}$. For linear functions of **G,** sampling errors and confidence intervals derived are equivalent to those obtained from $\mathbf{H}\left(\hat{\boldsymbol{\theta}}_{\mathbf{G}}\right)$. For non-linear functions, we are likely to obtain slightly more appropriate estimates than with the Delta method, as we are not performing a linear approximation.

In contrast, sampling $\boldsymbol{\theta}_L$ mimics what is done during the REML estimation process and thus attempts to approximate the actual distribution of estimates of $\hat{\mathbf{G}}$. This is affected by constraints on the parameter space and, while it ensures positive semi-definite samples $\tilde{\mathbf{G}}$, their mean is thus not necessarily equal to $\hat{\mathbf{G}}$, the difference reflecting bias due to constraints. This bias can be substantial if sample sizes are small and k is reasonably large. Samples of $\tilde{\mathbf{G}}$ or its functions obtained by sampling $\boldsymbol{\theta}_L$ should thus be more comparable to those from the MCMC methods discussed above, which also constrain estimates to the parameter space.

On either the **L** or **G** scale, samples from the distribution $\tilde{\mathbf{G}}$ are obtained as

$$\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \mathbf{L}_{\mathbf{H}}\mathbf{d}$$

where $\mathbf{L}_H$ is the Cholesky factor of the inverse of the information matrix, and **d** is a vector of standard normal deviates $d_i \sim N(0,1)$. The vector $\tilde{\boldsymbol{\theta}}$ is then reshaped into a sample matrix $\tilde{\mathbf{G}}$ for analysis. This approach has been implemented in the freely available mixed-model package Wombat (Meyer, 2010-2015). Using simulated data, Meyer and Houle (2013) demonstrated excellent agreement between empirical estimates of sampling variation and the L-scale REML-MVN estimates, a point we return to in the Discussion.

Methods

We estimated the **G** matrix based on wing measurements of a wild-collected population of *D. melanogaster* from Wabasso, Florida USA (Mezey & Houle, 2005). Mezey and Houle generated 170 half-sib and 790 full-sib families and measured 17,323 wings from parents and offspring. The phenotypic data were the x,y coordinates of 12 vein intersections measured with WINGMACHINE, a semi-automated system that records scale information and detects vein positions from digital wing images (Houle et al., 2003). The 24 coordinates obtained from each wing were geometrically aligned to the mean shape using Procrustes least-squares superimposition (Rohlf & Slice, 1990), which removes centroid size as a scaling factor. Although the superimposed data are still in the form of 12 pairs of coordinates, 4 degrees of freedom are used for superimposition, so the resulting **G** matrix has a maximum rank or dimensionality of 20. Mezey & Houle (2005) estimated **G** piecewise using a method-of-moments mixed model analyses of each pair of traits. Hansen and Houle (2008) used the average of Mezey & Houle's male and female **G** matrices, shown in Table S1. We will refer to this as the H&H08 **G**.

To estimate sampling error using REML-MVN, we re-estimated **G** using REML implemented in Wombat (Meyer, 2010-2015). Before the new analyses, the original Wabasso data were geometrically aligned with a much larger set of 83,000 wings, including specimens from 117 dipteran species, our spontaneous mutation data (Houle and Fierst 2013), and 184 Drosophila Genome Reference Project (Mackay et al., 2012) inbred lines. This enables as yet unpublished comparisons of the Wabasso **G** matrix to these data sets. We refer to the original superimposition used in previous publications (Mezey & Houle, 2005, Hansen & Houle, 2008) as the 'Wabasso' superimposition, and the new one as the 'combined' superimposition. Before analysis, we scored wing data on the first 20 eigenvectors of the phenotypic variance-covariance matrix from the pooled male and female Wabasso population data. We fit sex as a fixed effect to obtain a direct estimate of the pooled-sex **G** matrix. Estimation of **G** was carried out for both full- and reduced-rank models (Kirkpatrick & Meyer, 2004, Meyer & Kirkpatrick, 2005, 2008), and we selected the best-fitting model on the basis of Akaike's information criterion corrected for small sample size (AICc). REML-MVN estimates of sampling variances were then obtained drawing 100,000 samples of **G** on both the G- and L-scale.

MCMC analyses were carried out in the R package MCMCglmm (Hadfield, 2010). To investigate convergence, we initiated runs using parameters that were functions of the sex-adjusted phenotypic covariance matrix. All runs used a degree of belief of 20.002, slightly more than the dimensions of each matrix, and parameter expansion with a half-Cauchy prior with a scale parameter of $\sqrt{1000}$. These values combine to establish the priors as minimally informative. With parameter expansion, convergence was rapid, and burn-ins of just 100 iterations were necessary. Thinning to 60 iterations reduced autocorrelations between samples to 0.1 or less. Without parameter expansion, runs with different priors needed approximately 5,000 iterations of burn-in to achieve a stationary distribution, and runs with starting parameters far from the REML estimates often did not converge.

To provide a meaningful baseline against which to compare the parameter means and variances we carried out a parametric bootstrap analysis. This involved resampling data from a multivariate normal distribution on the pedigree of the Wabasso experiment, using the REML estimates of **G** and residual variances as population parameters. A full REML analysis was then carried out for each of 1000 simulated data sets, and estimates of sampling variances were obtained as empirical variances across replicates. Both resampling and analysis were carried out in Wombat.

We used the mean wing shapes of seven other drosophilid species (listed in Tables 2 and 3) to choose interesting directions in which to investigate evolvability (Hansen & Houle, 2008). The mean of each species was based on approximately 200 wings obtained from lab-reared flies. We recalculated the directions from *D. melanogaster* based on the same specimens used in H&H08, but using the combined superimposition, instead of a species-data only superimposition. This resulted in slightly different estimates of phenotypic distance and direction from those shown in H&H08.

Evolvability, *e*, is the predicted response to unit strength selection in the direction of the selection gradient, **β**, in the absence of stabilizing selection. It is calculated as the projection of the response vector to a unit-length **β** on **β**

$$e\left(\boldsymbol{\beta}\right) \equiv \boldsymbol{\beta}'\mathbf{G}\boldsymbol{\beta}.$$

Conditional evolvability, *c*, is the response to unit strength selection when stabilizing selection around the selected direction is infinitely strong. Conditional evolvability is

$$c(\boldsymbol{\beta}) = \left(\boldsymbol{\beta}'\mathbf{G}^{-1}\boldsymbol{\beta}\right)^{-1} \boldsymbol{\beta}'\boldsymbol{\beta},$$

and gives the response in direction $\boldsymbol{\beta}$ to a unit-length $\boldsymbol{\beta}$ when the response is constrained to be in direction $\boldsymbol{\beta}$. The actual response to selection in direction $\boldsymbol{\beta}$ will be between $e(\boldsymbol{\beta})$ and $c(\boldsymbol{\beta})$, falling closer to $e(\boldsymbol{\beta})$ when stabilizing selection in other directions is weak. Autonomy, a, is the ratio c/e, and captures the proportion of variation that allows response in the direction of a selection gradient. These measures of evolvability are informative when the units in which traits are measured are the same (as in our wing shape data), or the traits have been standardized in the same manner.

When the direction of selection is not predictable, one can ask about the average evolvability of a population averaged over all possible directions. Hansen and Houle (2008) showed that the expected evolvability, $\bar{e}$, is the average eigenvalue of the **G** matrix. No exact solution is available for the expected conditional evolvability, $\bar{c}$, or the expected autonomy, $\bar{a}$, but good approximations have been derived in Hansen & Houle (2008, 2009). The corrected formulas for these are repeated in Appendix 1.

Results

Reanalysis of Mezey& Houle's (2005) data on wing shape in the Wabasso population of *Drosophila melanogaster* shows that the best estimate is a **G** of rank 20 (full-rank). The full model is superior by 38 AIC-penalized log-likelihood units to the simplified rank 19 model in both the Wabasso and combined superimpositions. Mezey & Houle's (2005) conclusion that there were at least 18 dimensions of genetic variation in these data was conservative. The REML estimate of **G**, back-projected into the original 24 dimensions is shown in Table S2.

Table 1 shows the values of a set of evolvability statistics (Hansen & Houle, 2008, see Methods for definitions) and their sampling errors from parametric bootstrapping, MCMC estimation and the REML-MVN method. In addition estimates for the **G** estimated by Hansen & Houle (2008) are also shown for comparison. Overall, the sampling standard deviations are quite small relative to their means, resulting in sampling coefficients of variation for the evolvability statistics of 5% or less, with the exception of the minimum eigenvalue, e_{min}, which has a CV greater than 10% by all methods. The minimum eigenvalue is the most difficult to estimate as it is the variance closest to a boundary value of 0. G-scale estimates are not constrained to have a non-negative e_{min}, so the fact that the G-scale estimates of e_{min} are still many standard deviations greater than 0 supports the finding of a full-rank **G** matrix. The sampling distributions of all statistics were approximately normal (results not shown).

The parametric bootstrap estimates are a suitable baseline to compare the other methods with, as that method enforces multivariate normal data, and makes no large-sample assumption. The mean REML and MCMC estimates are all within a small fraction of the sampling standard deviation of the parametric bootstrap value, suggesting that there is little bias in the mean estimates of the parameters. On the other hand, the H&H08 estimates of $\bar{e}$ and e_{max} are more than 4 standard deviations higher than the REML estimates. Conversely, the H&H08 $\bar{c}$ and e_{min} are about 2 standard deviations lower than the REML estimates. The larger eigenvalues in the H&H08 estimate are biased upwards, while the smaller eigenvalues are biased downwards. Systematic over-dispersion of sample eigenvalues is a well-known outcome for estimates that are not constrained to the parameter space (Hill & Thompson, 1978).

Closer examination shows that the estimates of mean and sampling variation may show subtle biases. Even though the parametric bootstrap was initiated with the REML estimate, the estimates recovered from the bootstrap do not match the ‘best’ REML’ estimate precisely. In particular, the three statistics that depend on the inverse of **G** and therefore on the smallest eigenvalues (e_{min}, $\overline{c}$, $\overline{a}$), are all more than a standard deviation lower in the bootstrap sample. This may indicate departures of the data from multivariate normality in the original data. The same three statistics have slightly higher means in the L-scale sample than in the G-scale sample, which is consistent with the L-scale constraint towards positive-definite matrices. For these data, sampling on the G-scale**,** $\boldsymbol{\theta}_{G,}$ did not yield any samples which were not positive definite, and no values of e_{min} based on sampling the elements of its Cholesky factor, $\boldsymbol{\theta}_L$ approached the arbitrary constrained value of 0.0001 in Wombat. This leaves the precise cause of the discrepancy somewhat unclear.

To get a broader sense for the similarity of the estimates, we calculated the mean and standard deviation of a range eigenvalues, with the results shown in Figure 1. On the log scale all four sets of mean estimates are quite similar, with differences only becoming apparent in the smallest eigenvalues. Sampling standard deviations are systematically lower in the REML estimates compared to the bootstrap; MCMC standard deviations are even lower. This may suggest a small bias in the REML-MVN error estimates, as they are asymptotic, lower bound values. While the Wabasso data set comprises a large number of records, a 20-variate, full rank REML analysis requires estimation of 420 covariance components. Larger estimates from the parametric bootstrap may thus indicate that the sample size is not quite sufficient for large sample theory to hold. This pattern is sometimes reversed for the smallest eigenvalues and the statistics that depend on $\mathbf{G}^{-1}$. This may be due to the fact that the REML constraints on the parameter space will tend to truncate the smallest eigenvalues (Amemiya, 1985). An alternative explanation for these exceptions is sampling error, as the precision of the error estimates for these statistics is relatively low.

Schluter (1996) found that among-species and among-population variation tended to lie close to the first eigenvector of **G**, $\boldsymbol{g_{max}}$. Hansen and Houle (2008- H&H08) reasoned that if **G** shapes among-species differences, then the differences among species should be in those aspects of variation that have the highest evolvabilities, even if those are very different from $\boldsymbol{g_{max}}$. To choose interesting directions of selection to investigate, Hansen and Houle (2008) took *Drosophila melanogaster* as the focal species and predicted the ability of *D. melanogaster* to evolve towards the phenotype of seven other species that span the traditional genus *Drosophila* and one closely related outgroup (*Scaptodrosophila latifasciaeformis*). The results are shown in Table 2 for evolvability and Table 3 for conditional evolvability.

As originally found with the H&H08 **G**, evolvabilities and conditional evolvabilities in the directions of these species are all in the more variable parts of the phenotype space. As a result, most of the estimates in H&H08 are substantial overestimates, consistent with the bias in the higher eigenvalues of **G** noted above.

Estimates of sampling error for the evolvabilities estimated with each method are again broadly similar, consistent with the results noted above. The estimates are fairly precise, with sampling coefficients of variation slightly less than 5% for the evolvabilities, and 6 to 15% for the conditional evolvabilities. These errors are sufficiently small that almost all differences in evolvabilities between species are statistically significant.

Discussion

It has long been known that the additive genetic variance-covariance **G** is a useful tool for making predictions about evolution, and for interpreting the pattern of diversification among taxa (Lande, 1979). Until recently, efforts to utilize these results have been hampered by the difficulty of assessing the sampling variation of **G** and of the complex and often non-linear statistics that are functions of **G**. Bayesian estimation using a Markov-chain Monte Carlo algorithm (MCMC) has recently been applied to such problems (e.g., O'Hara et al., 2008, Hadfield, 2010, Aguirre et al., 2014, Stinchcombe et al., 2014), but application of MCMC methods can be computationally intensive for large problems.

As an alternative, we have applied our recently implemented REML-MVN method Meyer & Houle, 2013) of estimating the sampling variation in restricted maximum likelihood (REML) estimates of additive genetic variance-covariance matrices. As our example, we used data on wing shape in *Drosophila melanogaster* from a very large experiment (Mezey & Houle 2005). We focused on sampling variation in the evolvability statistics proposed in Hansen & Houle (2008).

Our goal in this contribution has been first to demonstrate the REML-MVN approach for a single-well-estimated data set. Comparison of parameter estimates and their sampling error based shows that REML-MVN estimates are quite similar to those derived from the parametric bootstrapping and MCMC in mean and variance. We can use the parametric bootstrap as the baseline for comparison, as those results depend on simulated data that corresponds to the assumptions of the analysis. The similarity of all three sets of results validates the accuracy both the parameter estimates and their sampling errors from the REML-MVN and MCMC approaches. This validation of the REML-MVN approach is also supported by the results for simulated data reported by Meyer & Houle (2013).

Looking more closely, there are small quantitative departures between bootstrap, REML-MVN and MCMC estimates. Discrepancies could in principle be explained either by flaws in the methods, in their application, or by departures of the data from the assumed multivariate normal distribution. In the case of REML-MVN, these departures potentially reflect insufficiently sampled aspects of **G** for which large sample results do not hold.

Given these results, the REML-MVN approach is attractive because it is usually computationally much more efficient than either MCMC, or bootstrap approaches. For the data reanalyzed here, convergence in Wombat (Meyer, 2007, Meyer, 2010-2015) from a poor initial estimate of **G** (equal to half the phenotypic variance-covariance matrix) takes 9.5 hours on an AMD Opteron 4180 processor with speed of 2793 MHz. Generation of 100,000 REML-MVN samples then requires only seconds of processor time. Using the R package MCMCglmm (Hadfield, 2010) the same problem takes about 6.5 hours to produce 1000 iterations. Thinning to every 60 generations, production of the 1,000 samples used in this analysis took over 400 hours of processor time. The greater the number of variables, and the closer the initial estimates are to the final estimate, the greater the run time advantage of REML-MVN over MCMC.

A second advantage of a maximum likelihood approach is that it can be used to test whether fitting a complex model over a simpler one is supported by the data (Meyer & Kirkpatrick, 2005, Meyer & Kirkpatrick, 2008). Such tests are important to perform when there is some doubt about whether a complex model can be supported by the data, given that both standard MCMC and the L-scale REML-MVN approach produce estimates constrained to be of full rank.

While our results, plus the simulations reported in Meyer & Houle (2013), validate the use of REML-MVN in some cases, this does not mean that REML-MVN will perform well for all data sets. Therefore, we suggest that REML-MVN estimates of sampling error should continue to be validated with estimates from a second approach. Parametric bootstrapping based on the REML estimates obtained is probably the least computationally intensive of the alternatives, given that if the model is strongly supported by the data, convergence with a new simulated data set should be relatively rapid. Restricted maximum likelihood does well for multivariate normal data, but is unsuitable when the data follows other distributions, whereas Bayesian methods readily accommodate such cases. REML-MVN depends on large-sample approximations that are inappropriate for data sets where the amount of information in the data is small relative to the number of parameters estimated. For such cases MCMC is likely to perform better. Alternative approaches, based for example on the profile likelihood for individual parameters, might also be more appropriate than REML-MVN when large sample properties do not hold.

The REML reanalysis of these data confirmed Mezey & Houle's (2005) conclusion that the **G** matrix for this data set is full-rank. Models with lower dimensionality fit at least 38 Akaike information criterion units less well than the full 20-dimensional model. Hine & Blows (2006) suggested that the bootstrapping method employed by Mezey & Houle (2005) was biased towards high dimensionality, but Hine & Blows simulated only one of the two bootstrapping approaches of Mezey & Houle. On the other hand, these new analyses do show that the original estimates obtained by Mezey & Houle (2005), using a method of moments analysis, were biased. Results that depend on the best-estimated parts of the **G** with large additive genetic variances, such as the maximum evolvability and the average evolvability were overestimated by Mezey &

Houle (2005) by up to 17%. On the other hand, the less well-estimated aspects of the matrix that have the least genetic variance were underestimated by up to 8%. This pattern of bias is expected for unconstrained estimates of covariance matrices (Hill & Thompson, 1978).

In conclusion, resampling **G** matrices using the restricted maximum likelihood, multivariate normal approach can generate accurate assessments of sampling variation in evolutionary statistics. The relatively short run time of this method makes it an attractive alternative to both data resampling and Bayesian estimation using a Markov chain Monte Carlo approach.

Acknowledgements

We thank two anonymous reviewers for their comments. This work was supported by U.S. National Science Foundation grant DEB-0950002, and Meat and Livestock Australia grant B.BFG.0050, and the working group “Darwinian Morphometrics: Cross-topology registration of shape” at the National Institute for Mathematical and Biological Synthesis, an institute sponsored by the National Science Foundation, the US Department of Homeland Security, and the US Department of Agriculture through NSF Awards (EF-083285 and DBI-1300426) with additional support from The University of Tennessee, Knoxville.

References

Aguirre, J. D., Hine, E., McGuigan, K. & Blows, M. W. 2014. Comparing G: multivariate analysis of genetic variation in multiple populations. *Heredity* **112**: 21-29.

Amemiya, Y. 1985. What should be done when an estimated between-group covariance matrix is not nonnegative definite? *The American Statistician* **39**: 112-117.

Cheverud, J. M. 1996. Quantitative genetic analysis of cranial morphology in the cotton-top (*Saguinus oedipus*) and saddle-back (*S. fuscicollis*) tamarins. *Journal of Evolutionary Biology* **9**: 5-42.

Cheverud, J. M. & Marroig, G. 2007. Comparing covariance matrices: random skewers method compared to the common principal components model. *Genetics and Molecular Biology* **30**: 461-469.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. 2013. *Bayesian data analysis*. CRC press.

Hadfield, J. D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of statistical software* **33**: 1-22.

Hansen, T. F. & Houle, D. 2008. Measuring and comparing evolvability and constraint in multivariate characters. *Journal of Evolutionary Biology* **21**: 1201-1219.

Hansen, T. F. & Houle, D. 2009. Corrigendum. *Journal of Evolutionary Biology* **22**: 913-915.

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

Hine, E. & Blows, M. W. 2006. Determining the effective dimensionality of the genetic variance-covariance matrix. *Genetics* **173**: 1135-1144.

Hine, E., Chenoweth, S. F., Rundle, H. D. & Blows, M. W. 2009. Characterizing the evolution of genetic variance using genetic covariance tensors. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**: 1567-1578.

Houle, D. & Fierst, J. 2013. Properties of spontaneous mutational variance and covariance for wing size and shape in *Drosophila melanogaster*. *Evolution* **67**: 1116-1130.

Houle, D., Mezey, J., Galpern, P. & Carter, A. 2003. Automated measurement of Drosophila wings. *BMC Evolutionary Biology* **3**: 25.

Kingsolver, J. G., Heckman, N., Zhang, J., Carter, P. A., Knies, J. L., Stinchcombe, J. R. & Meyer, K. 2015. Genetic Variation, Simplicity, and Evolutionary Constraints for Function-Valued Traits. *The American Naturalist* **185**: E166-E181.

Kirkpatrick, M. 2009. Patterns of quantitative genetic variation in multiple dimensions. *Genetica* **136**: 271-284.

Kirkpatrick, M. & Meyer, K. 2004. Direct estimation of genetic principal components: Simplified analysis of complex phenotypes. *Genetics* **168**: 2295-2306.

Krzanowski, W. J. 1979. Between groups comparison of principal components. *Journal of the American Statistical Association* **74**: 703-707.

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

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D. H., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barron, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., Harris, Z., Javaid, M., Jayaseelan, J. C., Jhangiani, S. N., Jordan, K. W., Lara, F., Lawrence, F., Lee, S. L., Librado, P., Linheiro, R. S., Lyman, R. F., Mackey, A.

J., Munidasa, M., Muzny, D. M., Nazareth, L., Newsham, I., Perales, L., Pu, L. L., Qu, C., Ramia, M., Reid, J. G., Rollmann, S. M., Rozas, J., Saada, N., Turlapati, L., Worley, K. C., Wu, Y. Q., Yamamoto, A., Zhu, Y. M., Bergman, C. M., Thornton, K. R., Mittelman, D. & Gibbs, R. A. 2012. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**: 173-178.

Mandel, M. 2013. Simulation-based confidence intervals for functions with complicated derivatives. *The American Statistician* **67**: 76-81.

Meyer, K. 2007. Wombat--A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *Journal of Zhejiang University (Science B)* **8**: 815-821.

Meyer, K. 2010-2015. Wombat: A program for mixed model analyses by restricted maximum likelihood. Animal Genetics and Breeding Unit, University of New England, Armidale, NSW, Australia http://agbu.une.edu.au/~kmeyer/wombat.html.

Meyer, K. & Houle, D. (2013) Sampling based approximation of confidence intervals for functions of genetic covariance matrices. In: *Proceedings of the Association for Advances in Animal Breeding*, Vol. 20. pp. 523-527. http://www.aaabg.org/aaabghome/AAABG20papers/meyer20523.pdf

Meyer, K. & Kirkpatrick, M. 2005. Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. *Genetics, Selection and Evolution* **37**: 1-30.

Meyer, K. & Kirkpatrick, M. 2008. Perils of Parsimony: Properties of Reduced-Rank Estimates of Genetic Covariance Matrices. *Genetics* **180**: 1153-1166.

Meyer, K. & Smith, S. P. 1996. Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. *Genetics Selection Evolution* **28**: 23-49.

Mezey, J. G. & Houle, D. 2005. The dimensionality of genetic variation for wing shape in *Drosophila melanogaster*. *Evolution* **59**: 1027-1038.

O'Hara, R. B., Cano, J. M., Ovaskainen, O., Teplitsky, C. & Alho, J. S. 2008. Bayesian approaches in evolutionary quantitative genetics. *Journal of Evolutionary Biology* **21**: 949-957.

Oehlert, G. W. 1992. A note on the Delta method. *American Statistician* **46**: 27-29.

Ovaskainen, O., Cano, J. M. & Merila, J. 2008. A Bayesian framework for comparative quantitative genetics. *Proceedings of the Royal Society B-Biological Sciences* **275**: 669-678.

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

Rohlf, F. J. & Slice, D. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology* **39**: 40-59.

Schluter, D. 1996. Adaptive radiation along genetic lines of least resistance. *Evolution* **50**: 1766-1774.

Sorensen, D. & Gianola, D. 2002. *Likelihood, Bayesian and MCMC methods in quantitative genetics*. Springer.

Stinchcombe, J. R., Simonsen, A. K. & Blows, M. W. 2014. Estimating uncertainty in multivariate responses to selection. *Evolution* **68**: 1188-1196.

Table 1.Overall evolvability statistics. Evolvabilities and conditional evolvabilities have units of 10^6 centroid size. Bootstrap, REML resamples and MCMC posterior distributions are each calculated from 1,000 samples.

	Mean					Standard deviation				
	$\bar{e}$	e_{max}	e_{min}	$\bar{c}$	$\bar{a}$	$\bar{e}$	e_{max}	e_{min}	$\bar{c}$	$\bar{a}$
H&H08	14.61	83.04	0.09	1.00	0.069					
REML	13.071	70.870	0.129	1.076	0.0947					
Parametric bootstrap	13.081	71.652	0.109	1.000	0.0883	0.247	3.247	0.016	0.049	0.0045
REML-MVN, G-scale	13.083	71.527	0.109	1.001	0.0883	0.222	2.834	0.018	0.055	0.0049
REML-MVN, L-scale	13.121	71.418	0.122	1.067	0.0937	0.227	2.822	0.017	0.049	0.0044
MCMC	13.259	72.168	0.110	1.022	0.0888	0.211	2.558	0.015	0.050	0.0044

Table 2. Evolvabilities in the direction of species divergence, $e(\boldsymbol{\beta})$, in units of centroid size × 10^6. Phenotypic distances from *D. melanogaster* wings to other Drosophilid flies are in centroid size units.

	distance to D. *melanogaster*	Best estimate			Mean				Standard deviation			
Species		H&H08	REML	MCMC	bootstrap	REML L-scale	REML G-scale	MCMC	bootstrap	REML L-scale	REML G-scale	MCMC
D. simulans	0.011	34.4	22.52	22.22	22.50	22.55	22.59	23.08	1.11	1.00	0.98	0.92
D. ananassae	0.082	66.7	41.44	41.85	41.43	41.50	41.54	42.11	1.92	1.70	1.67	1.45
D. pseudoobscura	0.041	64.9	38.44	38.50	38.47	38.46	38.40	38.99	1.79	1.64	1.57	1.59
D. willistoni	0.056	55.1	47.5	48.40	47.60	47.50	47.75	48.35	2.26	2.03	2.07	1.81
D. virilis	0.057	46.6	30.96	31.31	31.00	30.84	31.00	31.26	1.40	1.28	1.20	1.20
D. grimshawi	0.172	55.2	41.78	41.95	41.82	41.66	41.89	42.20	1.94	1.70	1.64	1.55
S. latifasiaeformis	0.114	56.9	48.63	49.03	48.68	48.65	48.84	49.21	2.29	1.95	1.96	1.65

Table 3. Conditional evolvabilities in the direction of species divergence, $c(\boldsymbol{\beta})$, in units of centroid size × 10^6. Samples described in Table 2.

	Best estimate			Mean				Standard deviation			
Species	H&H08	REML	MCMC	bootstrap	REML L-scale	REML G-scale	MCMC	bootstrap	REML L-scale	REML G-scale	MCMC
D. simulans	2.7	1.69	1.50	1.57	1.66	1.58	1.50	0.17	0.17	0.18	0.16
D. ananassae	13.7	13.75	13.11	13.09	13.51	13.11	13.11	1.04	0.96	0.99	0.84
D. pseudoobscura	12.7	6.69	6.51	6.28	6.58	6.30	6.51	0.56	0.54	0.59	0.57
D. willistoni	10.7	10.88	10.68	10.48	10.68	10.46	10.68	0.68	0.65	0.64	0.60
D. virilis	10.5	4.68	4.58	4.48	4.60	4.50	4.58	0.30	0.28	0.30	0.28
D. grimshawi	17.4	7.5	7.65	7.20	7.36	7.21	7.65	0.46	0.43	0.46	0.45
S. latifasiaeformis	24.9	9.53	8.24	8.75	9.37	8.75	8.24	1.15	1.19	1.24	1.08

Figure 1. Mean (A) and standard deviation (B) of $\log_{10}$ eigenvalue estimates from the parametric bootstrap, REML-MVN on the L- and G-scales, and MCMC.

Appendix 1

The original approximations for the expected conditional evolvability, $\bar{c}$, and autonomy, $\bar{a}$, over all directions in phenotype space in Hansen & Houle (2008) were incorrect, and were corrected in Hansen & Houle (2009). For clarity, we repeat the corrected equations here.

The approximations depend on the following quantities: *k* is the dimension of matrix, $\mathrm{E}[\lambda]$ and $\mathrm{E}\left[1/\lambda\right]$ are the means of the eigenvalues and of the inverse eigenvalue, respectively, $\mathrm{H}\left[\lambda\right]=1/\mathrm{E}\left[1/\lambda\right]$ is the harmonic mean eigenvalue; $\mathrm{I}\left[\lambda\right]=Var\left(\lambda\right)\Big/\left(\mathrm{E}\left[\lambda\right]^2\right)$ is the variance of the eigenvalues, standardized by the square of the mean eigenvalue; $\mathrm{I}\left[1/\lambda\right]=Var\left(1/\lambda\right)\Big/\left(\mathrm{E}\left[1/\lambda\right]^2\right)$ is the variance of the inverse of the eigenvalues standardized by the square of the mean inverse eigenvalue.

The expected value of $\bar{c}$ is approximately

$$\bar{c} \approx \mathrm{H}\left[\lambda\right]\left(1+\frac{2\mathrm{I}\left[1/\lambda\right]}{k+2}\right) .$$

The expected value of $\bar{a}$ is approximately

$$\bar{a} \approx \frac{\mathrm{H}\left[\lambda\right]}{\mathrm{E}\left[\lambda\right]}\left(1+2\frac{\mathrm{I}\left[\lambda\right]+\mathrm{I}\left[1/\lambda\right]-1+\mathrm{H}\left[\lambda\right]/\mathrm{E}\left[\lambda\right]+2\ \mathrm{I}\left[\lambda\right]\ \mathrm{I}\left[1/\lambda\right]/\left(k+2\right)}{k+2}\right) .$$