INFERRING FITNESS LANDSCAPES
Author(s): Ruth G. Shaw and Charles J. Geyer
Source: *Evolution*, Vol. 64, No. 9 (SEPTEMBER 2010), pp. 2510-2520
Published by: Society for the Study of Evolution
Stable URL: http://www.jstor.org/stable/40863328
Accessed: 25-07-2016 23:40 UTC

doi:10.1111/j.1558-5646.2010.01010.x

INFERRING FITNESS LANDSCAPES

Ruth G. Shaw[1,2] and Charles J. Geyer[3,4]

[1] *Department of Ecology, Evolution, and Behavior and Minnesota Center for Community Genetics, University of Minnesota, 100 Ecology Building, 1987 Upper Buford Circle, St. Paul, Minnesota 55108*

[2] *E-mail: rshaw@superb.ecology.umn.edu*

[3] *School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street S.E., Minneapolis, Minnesota 55455*

[4] *E-mail: charlie@stat.umn.edu*

Received October 4, 2009
Accepted March 5, 2010

Since 1983, study of natural selection has relied heavily on multiple regression of fitness on the values for a set of traits via ordinary least squares (OLSs), as proposed by Lande and Arnold, to obtain an estimate of the quadratic relationship between fitness and the traits, the fitness surface. However, well-known statistical problems with this approach can affect inferences about selection. One key concern is that measures of lifetime fitness do not conform to a normal or any other standard sampling distribution, as needed to justify the usual statistical tests. Another is that OLS may yield an estimate of the sign of the fitness function's curvature that is opposite to the truth. We here show that the recently developed aster modeling approach, which explicitly models the components of fitness as the basis for inferences about lifetime fitness, eliminates these problems. We illustrate selection analysis via aster using simulated datasets involving five fitness components expressed in each of four years. We demonstrate that aster analysis yields accurate estimates of the fitness function in cases in which OLS misleads, as well as accurate confidence regions for directional selection gradients. Further, to evaluate selection when many traits are under consideration, we recommend model selection by information criteria and frequentist model averaging.

KEY WORDS: aster models, fitness components, life history, multiple regression, natural selection.

Efforts to characterize natural selection that populations undergo in the wild began within decades of Darwin's enunciation of the process of adaptive evolution (Weldon 1895; Bumpus 1899). Yet, nearly a century later, Endler (1986) noted that surprisingly few estimates of natural selection on quantitative traits had been made. The study of selection in nature expanded rapidly only when Lande and Arnold (1983) presented their method using multiple regression to characterize natural selection and argued that the resulting estimates were those required for predicting evolutionary change in a set of quantitative traits. Since then, an extensive body of research to evaluate selection on many taxa and traits has developed (Kingsolver et al. 2001).

In brief overview, the method proposed by Lande and Arnold (1983) entails ordinary least squares (OLSs) multiple linear regression of relative fitness on the values of a set of traits for a collection of individuals representing a population. They interpreted the resulting estimate of the vector of regression coefficients, in this context termed the selection gradient β, as the strength and direction of selection directly on each of the traits. Similar analyses that also include as predictors the squares of each trait and cross products between them yield estimates of components of a matrix γ, which represent the curvature of the best quadratic approximation of the relationship between fitness and traits, the fitness landscape, and which they, and others since, interpreted as evidencing stabilizing selection if the curvature is negative (disruptive, if positive).

Soon after this paper appeared, concerns were raised that the method could yield misleading inferences about the fitness landscape for several reasons. For example, Mitchell-Olds and Shaw (1987) emphasized that standard hypothesis tests require the assumption that the conditional distribution of fitness given the traits is normal. If fitness is not normally distributed, then the usual hypothesis tests and confidence intervals produced by linear regression software are invalid. In addition, estimates of curvature are not reliably interpreted as indicative of stabilizing (respectively, disruptive) selection (Mitchell-Olds and Shaw 1987; Schluter 1988), which requires not only negative (respectively, positive) curvature, but also that the maximum (respectively, minimum) lies within the range of values in the population. In fact, even the sign of the curvature estimated from the OLS regression can be misleading (Mitchell-Olds and Shaw 1987; Schluter 1988). Further, although the method was originally illustrated with examples of inferring selection by considering a single component of fitness expressed during a brief period of the life span (e.g., survival through one storm), Arnold and Wade (1984a,b) emphasized the importance of fully accounting for differences in fitness that accrue through multiple episodes of selection at different stages of the life cycle (e.g., including fecundity in different years). They grappled with the difficulty of extending the basic method to accomplish this, as did Wade and Kalisz (1989) but acknowledged that they did not resolve it fully. Suggestions of alterations of the original method to alleviate each problem (e.g. Mitchell-Olds and Shaw 1987; Schluter 1988; McGraw and Caswell 1996; Janzen and Stern 1998; van Tienderen 2000) do not, collectively, address all these concerns (Shaw et al. 2008).

We have recently developed a statistical approach, aster modeling (Geyer et al. 2007; Shaw et al. 2008), to address the statistical challenges of making inferences about fitness. Aster jointly analyzes distinct components of life histories (e.g., survival, fecundity) to yield inferences about overall fitness. Geyer et al. (2007) present the formal theory of aster analysis. Shaw et al. (2008) demonstrate that aster can be used to address the range of biological questions for which fitness data are gathered on individuals representing populations; the examples illustrate use of aster in estimation of population growth rate, in comparisons of mean fitness, and in evaluation of the form of selection on multiple traits. Here, we examine the latter problem in greater depth, showing that aster analysis closely approximates the fitness landscape, recovering key features of evolutionary interest, whereas OLS can yield qualitatively misleading estimates of the fitness landscape. To illustrate these points as clearly as possible, we use simulated datasets, so that inferred fitness landscapes can be directly compared with those known for the data at hand. We begin with a case of two correlated traits and then consider estimation of the fitness landscape for a larger set of correlated traits (here, five traits). In this latter situation, we propose that model selection via information criteria (Akaike 1973; Schwarz 1978), implemented using the branch and bound technique (Hand 1981), be used in conjunction with aster modeling.

All statistical analyses discussed in this article are carried out in full in the accompanying technical reports (Geyer and Shaw 2008a,b, 2009, 2010a,b), produced using the R function `Sweave` so all results in them are actually produced by the code shown and hence are fully reproducible.

Aster Modeling

As Arnold and Wade (1984a) noted, a full understanding of selection requires accounting for multiple episodes of selection over lifetimes of individuals in the population of interest. Variation in survival through successive intervals contributes to variation in overall fitness, as does variation in reproduction. Integrating over all episodes of selection, it is appropriate to consider an individual's lifetime fitness as its total number of offspring in the next generation, when selection is weak and population size is changing slowly (Charlesworth 1980). Otherwise, the rate of offspring production also bears on an individual's fitness. Fitness, as the overall realization of offspring number resulting from multiple distinct life-history events, has a compound distribution, often with multiple modes, including a mode at zero, corresponding to the many individuals that die before reproducing. Consequently, fitness does not generally conform to a normal distribution, nor to any other well-known parametric distribution. This fact undermines the validity of analyses of fitness by standard methods grounded in normal-distribution theory (e.g., OLS regression and analysis of variance).

Aster modeling addresses the statistical problems typical of lifetime fitness. In brief, aster takes advantage of the fact that, in contrast to overall fitness, individual life-history events (components of fitness) tend to fit standard probability distributions in the general class of exponential families (e.g., survival to age x as Bernoulli, fecundity at age x as Poisson). In analyses via aster, an appropriate sampling distribution is specified for each fitness component. Aster analyses also explicitly model a further characteristic of life histories: the dependence of components expressed later in life on ones expressed earlier (e.g., fecundity at age x depends on survival to that age), as represented in the graphical model for the particular life history (see Fig. 1). Aster graphical models allow any fitness component to be influenced by only a single previous component. Consequently, aster modeling accommodates much of the dependence structure of fitness but may not model all of it.

Individual components of fitness can be analyzed via aster, but its particular utility is that it yields estimates of fitness that integrate the underlying fitness components through its use of the unconditional canonical parametrization (Appendix). In this way, given data on all the components of a life history like that of Figure 1, aster produces estimates (with appropriate sampling variance) of the expected number of progeny a newborn contributes to the next generation, taking into account differential survival up to and through the reproductive period. This is accomplished via maximum likelihood, where the likelihood

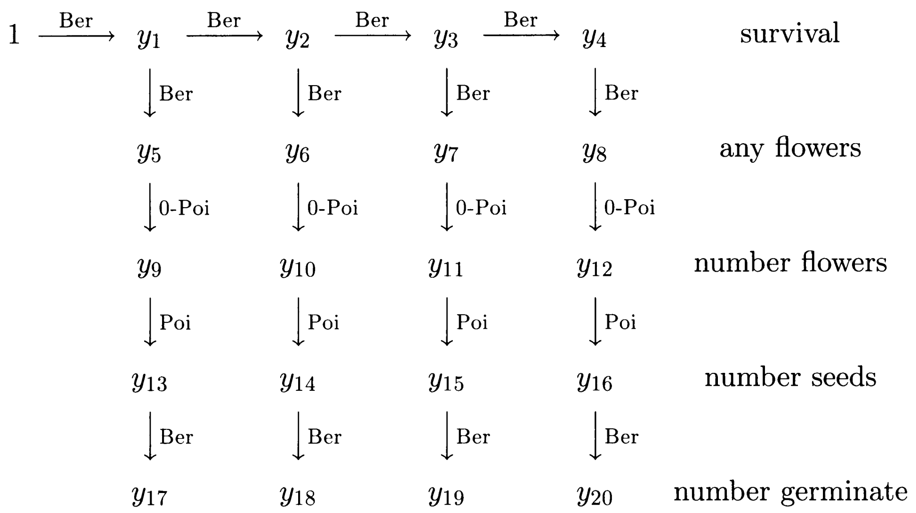

Figure 1. Aster model graphical structure. Subgraph for one individual, simulated data. Labels on arrows denote conditional distributions: Ber, Bernoulli; 0-Poi, zero-truncated Poisson; Poi, Poisson.

function incorporates the dependence of later components of fitness on those expressed earlier (Geyer et al. 2007). This approach offers sound and direct statistical modeling of overall fitness, as well as its relationship with traits. Because the aster approach is developed in a maximum likelihood framework, it provides a straightforward, statistically valid context for hypothesis testing. These aspects suit aster well as a foundation for investigating selection (Shaw et al. 2008).

Challenges of Modeling the Fitness Surface

Information on the direction and strength of selection on traits resides in the fitness surface, the relationship between fitness and all the traits under consideration. The method of Lande and Arnold (1983) employs OLS regression of relative fitness on the values of a set of traits **z** measured on a group of individuals representing a population. If the regression is linear, the resulting vector of regression coefficients contains estimates of the intercept α and the gradient vector **β** in the best linear approximation (BLA) of the fitness landscape with respect to the traits **z**, arrayed as a column vector,

$$g(\mathbf{z}) = \alpha + \mathbf{z}^T\boldsymbol{\beta},$$

where the sign of β_i is interpreted as the direction of selection on trait i, and its magnitude as the strength of selection on that trait.

If quadratic regression is conducted, including the squares of traits and Cross-products between pairs of traits as predictors, the resulting vector of regression coefficients contains estimates of the intercept α, the gradient vector **β**, and the matrix **γ** in the best quadratic approximation (BQA) of the fitness landscape

$$g(\mathbf{z}) = \alpha + \mathbf{z}^T\boldsymbol{\beta} + \frac{1}{2}\mathbf{z}^T\boldsymbol{\gamma}\mathbf{z}.$$

Stinchcombe et al. (2008) have stressed that fitting of this function correctly with typical regression software requires care about the factor of 1/2. Blows and Brooks (2003) have stressed the importance of including off-diagonal components of γ in fitting a general quadratic function.

Lande and Arnold (1983) gave two additional interpretations of **β** and **γ**, under the assumption that the phenotype vector **z** is jointly multivariate normal with mean zero. Then **β** is also the average gradient vector of the fitness landscape and **γ** the average Hessian (second derivative) matrix, these averages being with respect to the assumed multivariate normal distribution of **z**. Also, **β** appears in the expression for the change of trait mean due to selection, and **γ** appears in the expression for the change in genetic variance due to selection, when a standard quantitative genetic model is assumed (Bulmer 1972; Turelli 1988). Lande and Arnold (1983), in effect, defined four different betas and three different gammas: one set are the coefficients appearing in the BLA or BQA (they may have different betas), another set are the expected gradient vector and Hessian matrix of the fitness landscape, a third set relate to changes in trait mean and variance under selection (see Geyer and Shaw 2008b, Section 1.2 for details). When **z** is multivariate normal with mean zero, these distinct definitions of beta and gamma are equal, but not otherwise, with one exception: the beta that appears in the BLA is equal to $\boldsymbol{\beta} = \text{var}(\mathbf{z})^{-1}\text{cov}(w, \mathbf{z})$, which is related to changes in trait mean under selection even when **z** is not multivariate normal. Otherwise, the conditions for their equality involve first and second derivatives of the probability density function and first, second, third, and fourth moments of the assumed multivariate normal distribution of **z**. Consequently, these different betas and gammas need not match closely unless the distribution of **z** is very close to multivariate normal. Because statistical methodology for transformation to multivariate normality is rather crude (Andrews et al. 1971; Riani 2004), this assumption is not generally achieved in practice. Further details are provided in Geyer and Shaw (2008b).

We focus on the interpretation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as parts of the BLA or BQA. The OLS estimates of these parameters are best linear unbiased estimates (BLUE) of the BLA or BQA. These OLS estimates can, however, be severely biased estimates of the true fitness landscape.

A simple example demonstrates that the BQA of the fitness surface may seriously misrepresent the true fitness surface, and can therefore yield a misleading picture of the nature of selection. Suppose that a single trait is under stabilizing selection and, setting aside the statistical considerations of sampling and uncertainty, that the true relationship between fitness and that trait is known. The BQA of the fitness function approximates the true relationship well when the peak of the true fitness landscape coincides with the mean phenotype value (Fig. 2A).

The BQA approximation becomes poorer the more the fitness peak deviates from the trait mean (Fig. 2B,C). When the peak of the true fitness landscape is two standard deviations away from the mean of **z**, the BQA no longer has a peak, and the curvature of the BQA is positive, opposite to that of the true fitness function near its peak and therefore misleading about the form of selection on the trait, as shown in Figure 2C. Consequently, one might erroneously conclude from the BQA that selection is disruptive, whereas, in fact, selection is stabilizing, as well as directional. This problem, which has long been recognized for the case of analyzing selection via a single component of fitness (Mitchell-Olds and Shaw 1987; Schluter 1988), means that the curvature of the fitness function may be inferred erroneously as an artifact of biased statistical methodology.

A further problem with the Lande–Arnold method is that the conditional probability distribution of fitness given the phenotype vector **z** is usually not homoscedastic and normal. In fact, the distribution of fitness often has a large mode at zero, representing individuals that die before reproducing. As a result, it is far from normal, nor can it be transformed to normality; see, for example, the distribution of fitness for data simulated from the graph in Figure 1 (Fig. 3).

For this reason Lande and Arnold (1983) did not make any assumptions about the distribution of fitness in their mathematical development. Normality is not required for the OLS estimates of the BLA or BQA to be the BLUE (Lindgren 1993, p. 510). However, all other familiar properties of OLS estimators do require homoscedastic normality, hence this assumption is required for validity of all P-values and standard errors in the examples in Lande and Arnold (1983) (Mitchell-Olds and Shaw 1987).

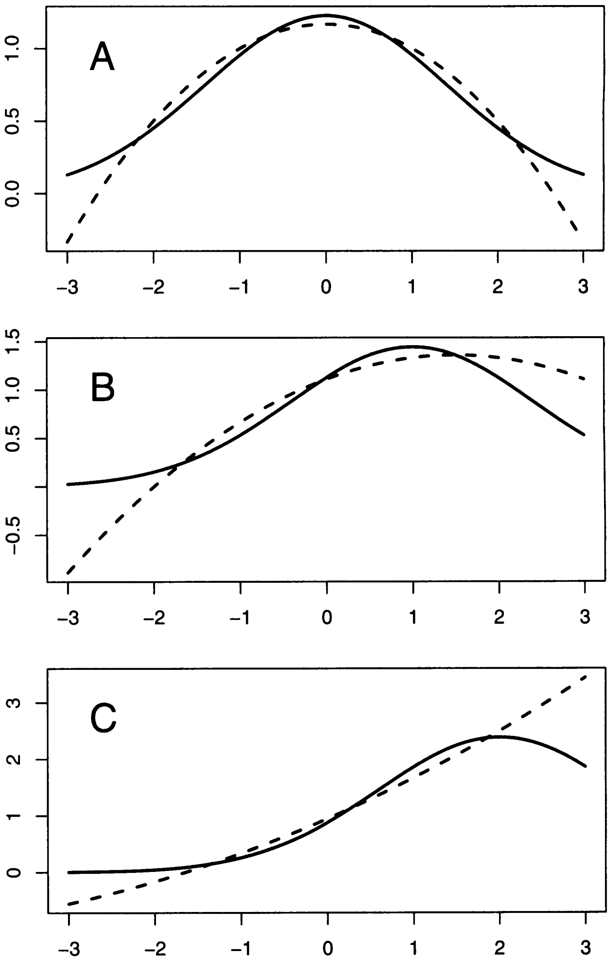

Figure 2. Fitness landscape (solid) and its best quadratic approximation (dashed). Vertical axis is relative fitness (fitness divided by mean fitness). Horizontal axis is phenotype in units of standard deviations from the mean. Source: Geyer and Shaw (2008b).

Estimating the Fitness Surface

The aster approach for statistical modeling of lifetime fitness yields direct estimates of the fitness surface. We illustrate this using data simulated for a hypothetical population, in which individuals express two selected traits, simulated as jointly normal to meet the assumptions required for the interpretations given in Lande and Arnold (1983). Selection occurs in 20 bouts over four years (Fig. 1): for each year, each individual's survival, three aspects of its reproduction (e.g., for a plant, whether it flowered, the number of flowers it produced and the number of seeds it produced; or for an animal, whether it mated, the number of matings it had, and the number of eggs it laid), and the number of offspring that germinated (respectively, hatched). We use as the measure of overall fitness for each individual the total number of offspring derived from each parental individual (i.e., the sum of y_{17} through y_{20}). (If data were available only through seed

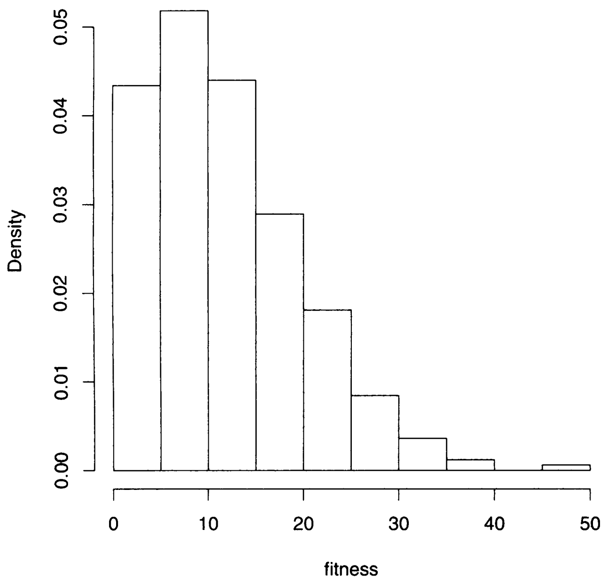

Figure 3. Histogram of the nonzero part (66.4%) of the fitness distribution for simulated data having the graph Figure 1; 33.6% of individuals have fitness zero. Data: variable `y` in the `ladata` data frame of the `sim` data in the `aster` package.

production, then the summed seed counts could be used as the measure of fitness. We here wish to illustrate aster's capability of accounting for selection taking into account all available information, including into the next generation, if it is available.) We estimate the fitness function over the range of bivariate phenotypes as the expected number of offspring that result in the next generation from an individual of a given phenotype in the previous generation. Fitness also varies with the timing of reproduction in age-structured populations (Charlesworth 1980), with earlier produced progeny contributing more to their parents' fitness than ones produced later. Work in progress will incorporate consideration of the timing of reproduction in fitness inference.

The aster analysis recovers the simulated direction of selection: fitness increasing with the value of both traits over most of the range (Fig. 4A, B; see also Section 5 below). It also recovers the negative curvature of the fitness function, as simulated for both traits, indicated by the contours of the estimated fitness function. The aster estimate does not perfectly match the true fitness function from which the data were simulated. It would match more closely for larger sample sizes (maximum likelihood provides consistent estimation). The estimated position of the fitness optimum is also near to the true optimum, falling near the edge of the observed data for trait 1, but well within the range for trait 2 (Fig. 4A), (cf. the true fitness optimum, indicated by the dot in Fig. 4B). Consequently, true stabilizing selection occurs only on trait 2. Figure 5 shows confidence regions for the maximum of the fitness landscape.

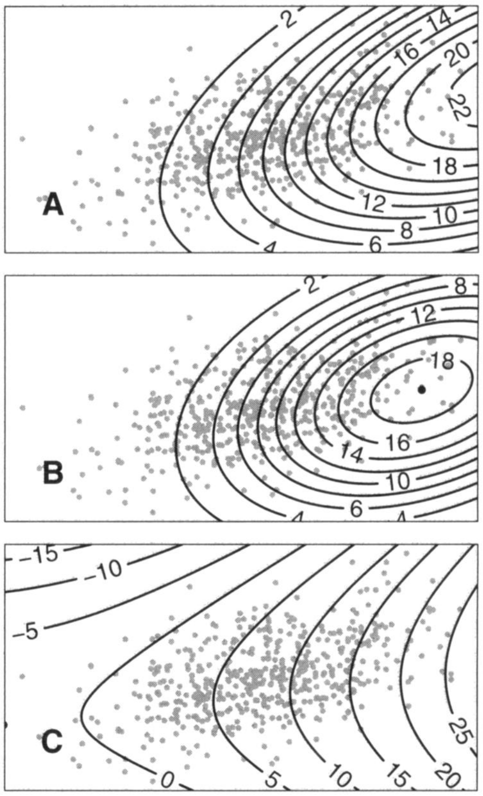

Figure 4. Scatterplot of two phenotypic variables (trait 1 on abscissa, trait 2 on ordinate). Gray dots indicate phenotypes of individuals, lines are contours of the fitness landscape and black dot the location of the maximum or stationary point of the fitness landscape. (A) fitness landscape estimated by the aster model. (B) simulation truth fitness landscape. (C) fitness landscape estimated by the Lande–Arnold method. Source: Geyer and Shaw (2008a).

Geyer and Shaw (2010a, Section 3.1 and 3.2) describe tests of the hypothesis that the fitness landscape has a maximum, that is, that selection is stabilizing in all directions. When this is true, the parameters describing the fitness landscape fall within 2^{-k} of the total parameter space, where k is the number of traits (i.e., a maximum exists only if $\gamma_{ii} < 0$ for all i). Consequently, it is appropriate to adjust the nominal P-value based on the chi-square approximation for the distribution of the deviance of the test of the hypothesis that the fitness landscape is quadratic versus that it is linear on the canonical parameter scale. An approximate adjustment is to divide the nominal P-value by 2^p. This approximation is not rigorous, but the parametric bootstrap (Geyer and Shaw 2010a, Section 3.2) is and shows that the approximation is conservative, at least in this example.

In sum, the aster analysis yields a close approximation of the key attributes of the true fitness landscape, its shape and the point at which fitness is maximized. In contrast, the BQA of the fitness landscape estimated by OLS misrepresents the shape of the true fitness function, particularly with respect to selection on trait 1,

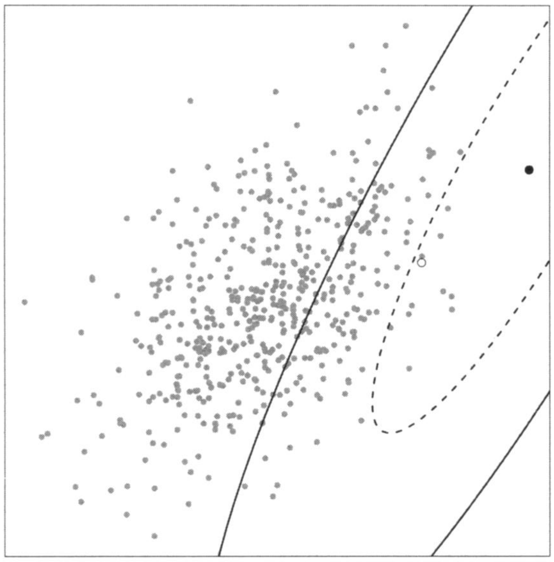

Figure 5. Scatterplot of two phenotypic variables (gray dots) with location of MLE of maximum of the fitness landscape (solid dot), maximum of the simulation truth fitness landscape (hollow dot), boundary of asymptotic 50% confidence region for the maximum (dashed curve), and boundary of the asymptotic 95% confidence region (solid curve). Source: Geyer and Shaw (2010a).

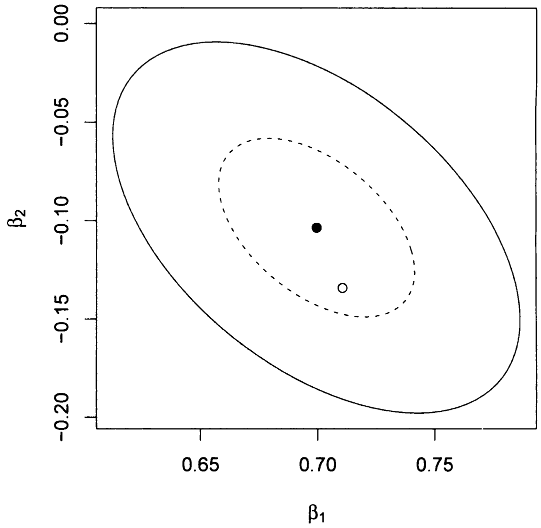

Figure 6. Confidence regions for the selection gradient β. Solid curve: boundary of the asymptotic 95% confidence region for β, dashed curve: boundary of asymptotic 50% confidence region, solid dot MLE of β, hollow dot simulation truth value of β. Source: Geyer and Shaw (2010b).

for which the true optimum differs greatly from the trait mean. The BQA does capture the general increase in fitness with the trait over the observed range, but the OLS estimate is a saddle, suggestive of disruptive selection on trait 1 (Fig. 4C). A further problematic aspect of the fitness landscape estimated by OLS is that there is a sizable region over which the fitness is negative.

Estimating the Directional Selection Gradient

For many purposes, analysis of absolute fitness, as above, is most biologically informative. For the purpose of predicting response to selection $\mathbf{G}\boldsymbol{\beta}$, where $\mathbf{G}$ is the variance–covariance matrix of breeding values (G matrix) estimated using quantitative genetics, $\boldsymbol{\beta}$ must be the slope of the BLA of the relative fitness landscape, termed by Lande and Arnold (1983) the directional selection gradient. Thus, in this section only, we use relative fitness (individual fitness divided by mean fitness) as the response variable, and we estimate the slope of the BLA of the relative fitness landscape using the aster model.

Applying the OLS regression method of Lande and Arnold (1983) to the simulated data analyzed in Section 4 above, we get the estimate $\beta_1 = 0.6997$ and $\beta_2 = -0.1036$ (Geyer and Shaw 2010b provide details of all computations for this section). To estimate $\boldsymbol{\beta}$ via the aster approach, we apply OLS, using as the response the *expected* relative fitness estimated from the aster model, instead of observed relative fitness. Because of the properties of maximum likelihood in exponential family models, the two procedures yield identical point estimates of $\boldsymbol{\beta}$. However, the sampling distribution of the estimate derived from the aster model is very different from the sampling distribution based on the usual assumptions for OLS (that the response is homoscedastic normal given the predictors). It is clear that these data grossly violate the homoscedastic normality assumptions (Fig. 3); this invalidates tests and confidence intervals using standard OLS. We can perform valid hypothesis tests and obtain correct confidence intervals using the aster model.

For example, consider a two-tailed test with null hypothesis $\beta_2 = 0$. OLS software gives a nominal P-value $P = 0.015$ whereas the correct P-value derived from the aster model is $P = 0.007$. So using the P-value derived from OLS software is, in this case, about twice what it should be. Figure 6 shows a confidence region for the true unknown $\boldsymbol{\beta}$. Because the 95% confidence region does not touch either coordinate axis, we can reject both null hypotheses $\beta_1 = 0$ and $\beta_2 = 0$ at the 0.05 level, even accounting for having done two tests.

Selection on Many Traits

Lande and Arnold (1983) emphasized that the whole phenotype, encompassing numerous traits, is subject to selection simultaneously. Accordingly, understanding selection requires accounting for variation in fitness with respect to many traits at once. In practice, when many correlated predictors are included in an analysis, the statistical power to detect relationships with individual traits can be severely compromised. In the extreme, $\boldsymbol{\beta}$ is undefined because substantial correlation between traits makes the matrix of phenotypic variances and covariances singular. To evaluate selection when many traits are being considered or their correlations are strong, Lande and Arnold (1983) suggested assessment of selection with respect to one or a few principal components (PCs) of the trait variation. This approach addresses the computational problem, but there is no reason to expect that fitness is related to the multivariate axes of greatest variation (Mitchell-Olds and Shaw 1987; Schluter and Nychka 1994). In fact, use of PCs may obscure the nature of selection, most obviously if the linear combination of traits under selection is perpendicular to the major PCs. Cox (1968) and Cook (2007) more generally critique the use of PCs as predictors in regression models.

Statisticians have devoted considerable study to this problem; Burnham and Anderson (2002) is a good introduction to the subject. Statistical and computational developments in the past 35 years have established a direct approach to selecting among statistical models that can be used to identify the traits to which fitness is related. When the number of models is modest, less than a few thousand, the data can be fit to all the possible statistical models. Fitness may be allowed to vary in relation to each trait either linearly, quadratically, or not at all. With only these three possibilities for each of k traits, there are 3^k models. For all pairs of traits included as quadratic in a given model, the cross-products must also be included to allow for correlational selection. Higher order terms could also be considered, but even restricting attention to quadratic functions, the number of models increases rapidly with the number of traits; when $k = 7$, there are $3^7 = 2187$ models, and when $k = 10$, there are $3^{10} = 59{,}049$ models. The practical limit on the number of traits that can be handled is roughly 7–10.

Once all possible models have been fit, it is necessary to compare them. A long established method for comparing two models is the likelihood ratio, but its use is restricted to the case when the smaller model is derived from the larger by fixing some parameter values, typically at zero, that is, the smaller model is nested within the larger. The likelihood ratio should not be used to compare nonnested models, nor should it be used for more than a few comparisons.

Maximum likelihood estimates are asymptotically unbiased, but maximized log likelihood is not. Model selection procedures correct for this bias of the maximized log likelihood by applying a penalty. Let Q_m denote minus twice the maximized log likelihood for the mth model. The Akaike (1973) information criterion (AIC), is $Q_m + 2p_m$, where p_m is the number of parameters for the mth model. It is an asymptotically unbiased estimate of $E(Q_m)$, where the expectation is with respect to the true unknown distribution of the data. An alternative criterion is the Bayes information criterion (BIC), an asymptotic approximation to twice the log Bayes factor, used by Bayesians for model selection (Schwarz 1978), which is $Q_m + \log(n)p_m$, where n is the sample size. Yet another is AIC corrected for sample size (AICc) (Sugiura 1978), which is $Q_m + 2p_m \cdot n/(n - p_m - 1)$.

Any one of these criteria is reasonable, as are others in the statistics literature. They are designed for different situations. When one of the models under consideration is correct, BIC chooses the correct model with probability approaching one as n goes to infinity; this is called consistent model selection. Neither AIC nor AICc have this property. Burnham and Anderson (2002, Section 1.2.5) are particularly emphatic about the biological unrealism of "true" models with only a few parameters. AIC is designed for the situation in which it may be that none of the models under consideration is correct (the truth is more complicated than any of them). In this situation, consistent model selection is impossible, and the justification for BIC does not hold, so AIC or AICc is preferred. If, in fact, fitness depends only on a few traits, then BIC is expected to be the best guide for model selection, whereas, if fitness depends on many traits, then AIC or AICc would be a better guide. Of course, the number of traits on which fitness depends is never known.

Once a criterion is chosen, the model selection procedure entails fitting all possible models, and selecting the model that minimizes the criterion. The branch and bound algorithm (B&B) speeds up the computing by finding the model with the smallest value of the information criterion without actually evaluating each model (Furnival and Wilson 1974; Hand 1981). B&B allows selection among somewhat larger classes of models, but is not magic. If 10 traits is the limit without B&B, then B&B may increase it to 12. Information criteria and B&B are both rigorously justified methods that are regularly employed, for example, in phylogenetic inference when many taxa are under consideration (Felsenstein 2004).

Model selection is not guaranteed to find the model that includes all (and only) the traits that contribute to fitness; in fact, when many models are under consideration the probability that it will is small. Thus, the goal of model selection or model averaging (see below) is not to find the true model; this is usually futile. The goal is to account for the response variable accurately even without knowing the true model. When the true model includes many traits, that model, with estimated parameters, may account for fitness less well than an alternative one obtained via model selection or model averaging. This issue, well recognized in the quantitative genetic efforts of inferring G matrices (Hill and Thompson 1978) and QTL mapping (Bernardo 2001) results from the trade-off between bias and variance; in models with more parameters, the estimates of parameters are more accurate (less biased) but also less precise. Under the true model, the variance of expected values of fitness may be very large, whereas much smaller models yield smaller mean square error (variance plus bias squared).

There may be too many possible models to fit them all, or even to find the one minimizing a criterion using B&B. An approach developed for this situation in the context of linear and generalized linear models is the LASSO (Hastie et al. 2009, Sections 3.4 and 4.4); something similar could be done with aster models, but this is a very active area of research. Because it is unclear how best to proceed when there are too many possible models to fit them all, we make no specific proposal for this situation.

To illustrate, we used B&B to obtain models relating fitness to traits for data simulated on 350 individuals and using a graphical model similar to that of Figure 1, but differing in the number of fitness components (three: survival, reproduction, and the number of seeds/eggs produced) and the number of years (ten) (Geyer and Shaw 2009, give details of all computations for this section). We considered two scenarios. In both, 10 traits were simulated, with all pairs of traits jointly normal and correlated with $r = 0.5$. In the first, the simulation truth fitness landscape depended on only two traits, with directional and stabilizing selection on both. In our second scenario, we simulated more realistic data, in which fitness depended on all 10 phenotypic traits, but not equally, each one after the first two having only half the importance of the preceding one. Again, selection was directional and stabilizing for all 10 traits. We limited the models to consideration of only the first five traits, and then to the first seven traits, corresponding to the situation, undoubtedly common, in which not all traits under selection have been measured.

In the first scenario, when model selection was done among all $3^5 = 243$ models with five traits, B&B found six models having AICc within $2 \log(20) = 5.99$ of the minimum, this cutoff corresponding approximately to a posterior probability of 0.05 (Geyer and Shaw 2009, Section 8); the simulation truth model ranked third best. B&B does not evaluate all models, but exhaustive search found nine models having AICc within 2 log(20) of the minimum, and the simulation truth model was fourth best among these. The top 10 models according to AICc (those within 2 log(20) of the minimum plus one more) are shown in Table 1.

When model selection was done in the same scenario among all $3^7 = 2187$ models with seven traits, B&B found 32 models having AICc within 2 log(20) of the minimum; the simulation truth model ranked seventh best among them (not shown). Again, B&B does not evaluate all models; we did not do an exhaustive search. Thus, even with a seemingly modest number of moderately correlated traits and a sizable dataset, the analysis may not recover the model under which the data were generated; increasing the number of traits exacerbates this situation. Conversely, if the sample size is increased while keeping the number of traits fixed, and, as in the first scenario, the true model is among the models under consideration, then, for sufficiently large sample sizes, BIC is guaranteed to find the correct model. We do note that all of the models within 2 log(20) of the best one correctly detected the nonlinear relationship between fitness and each of the traits that actually influence fitness (in the simulation truth), both when models considered five traits and when they considered seven traits.

Table 1. Model selection criteria. Ten best models in Scenario 1 with 5 phenotypic variables. Rank is the rank according to AICc. Q, L, and - indicate that the fitness landscape is quadratic, linear, or constant, respectively, in the indicated phenotypic variable in that model. The true model is given in bold.

Rank	z_1	z_2	z_3	z_4	z_5	AICc	BIC
1	Q	Q	L	-	-	1826.61	1864.54
2	Q	Q	L	L	-	1828.46	1870.11
3	Q	Q	L	-	L	1828.74	1870.40
4	**Q**	**Q**	**-**	**-**	**-**	**1829.39**	**1863.58**
5	Q	Q	-	L	-	1830.01	1867.94
6	Q	Q	L	L	L	1830.59	1875.96
7	Q	Q	-	-	L	1831.25	1869.18
8	Q	Q	Q	-	-	1831.36	1880.43
9	Q	Q	-	L	L	1832.08	1873.74
10	Q	Q	Q	L	-	1833.14	1885.90

In our second, more biologically realistic, scenario, the true model, involving all 10 traits, is not among the models under consideration (not shown). This mimics reality; it is never possible to measure every trait that influences fitness. In this scenario, when B&B was used to compare all models with five traits, it found three models with AICc within 2 log(20) of the minimum; all were quadratic in the first four traits, which had the strongest effects. Exhaustive search found no more models within 2 log(20) of the minimum. When all models with seven traits were considered, B&B found six models with AICc within 2 log(20) of the minimum, all quadratic in the first four traits and trait six, but differing in whether traits five and seven were constant, linear, or quadratic.

Given that model selection finds the correct model often only with low probability, it is useful to average over the models, weighting them according to their goodness, that is, to employ frequentist model averaging (FMA, Burnham and Anderson 2002). The weights, $\exp(-\frac{1}{2}\text{BIC}(m))$, where BIC (m) is BIC for model m, are asymptotically proportional to posterior probabilities. Hence, FMA with these weights approximates Bayesian Model Averaging (BMA) (Hoeting et al. 1999). By analogy, we use weights $\exp(-\frac{1}{2}\text{AICc}(m))$, where AICc$(m)$ is corrected AIC for model m, when using AICc as our model selection criterion. In the spirit of what Madigan and Raftery (1994) called Occam's window we set weights below a certain threshold to be zero, our chosen threshold being information criterion 2 log(20) above the minimum, where the weight is 0.05 of the maximum weight.

In our first scenario, FMA with either AICc or BIC resulted in lower root mean square error averaged over the whole estimated fitness surface than choice of a single model. Yet here, where the simulation truth model had two predictors and nine parameters, characterization of fitness using the true model performs better than either model selection or model averaging because the true model has fewer parameters than the models selected or averaged. In our second scenario, FMA with BIC yielded a better model and FMA with AICc a slightly worse one than choosing a single model. In this case, where the simulation truth model had 10 predictors and 69 parameters, fitness prediction with all 10 traits performs worse than either model selection or model averaging because the true model has many more parameters, each estimated with uncertainty, than the models selected or averaged.

Discussion

Accurate estimates of fitness functions are central to understanding adaptive evolution. In particular, stabilizing selection has generally been considered the predominant mode of selection on quantitative traits (Crow and Kimura 1970) and consequently figures heavily in quantitative genetic theory. Yet in a review of studies of selection in nature, Kingsolver et al. (2001) found no predominance of stabilizing selection. The rarity of direct evidence for this mode of selection could indicate a need for thorough reevaluation of evolution of quantitative traits. However, the available body of evidence could be systematically misleading, as we discuss below.

Apart from the sheer arduousness of acquiring sufficient data, empirical determination of selection surfaces has been problematic for fundamentally statistical reasons. To characterize phenotypic selection fully, it is necessary to obtain selection estimates that take into account the multiple bouts of selection over the full life cycle (or even into the next generation). Yet the distribution of lifetime fitness in a population never conforms to a single well-known statistical distribution like the normal, binomial, or Poisson. Rather, because mortality and reproduction are episodic, fitness is typically a compound distribution, for which no standard statistical methods are available. In contrast, components of fitness typically do conform to standard probability distributions and so are amenable to generalized linear modeling. From the simple distributions of the underlying components of fitness, aster builds the compound distribution of fitness to accomplish statistically sound estimation and hypothesis testing.

Aster yields accurate estimates of fitness functions, as we demonstrate here. Our analyses recover the known fitness relationships from which the data were simulated (compare Figs. 4A and 4B). In contrast, OLS (Lande and Arnold 1983) can produce a misleading estimate of the shape of the fitness function. It can be expected to do so whenever the population's trait mean differs substantially from the trait value that confers highest (or lowest) fitness (Fig. 4C). This problem traces to the inadequacy of the best quadratic approximation (BQA) for fitting the shape of fitness surfaces, even simple ones (Fig. 2C), a problem that has long been known (Mitchell-Olds and Shaw 1987; Schluter 1988), yet is often overlooked. Aster does not produce such artifacts because it models the relationship between fitness and traits via the underlying unconditional canonical parameter of the appropriate compound distribution of fitness (Appendix). More complicated fitness surfaces, for example, having multiple peaks, could be fitted using aster either by including higher order terms in the model, or by employing regression splines, as proposed by Schluter (1988). We suggest that OLS' artifactual estimates of curvature when the optimum differs from the population mean could largely account for the findings of Kingsolver et al.'s (2001) review of selection studies; the numbers of cases with positive curvature and negative curvature were roughly equal.

We emphasize that, even once the sequential components of fitness are appropriately modeled, significant curvature of the fitness function need not imply that stabilizing/optimizing or disruptive selection is operating. For example, strictly directional selection with increments in fitness declining as the trait value becomes more extreme also has negative curvature (Fig. 4A). Thus, once the sign of the curvature is established, evidencing stabilizing selection requires demonstrating that the maximum of the fitness function lies within the range of phenotypes in the populations (Mitchell-Olds and Shaw 1987), or, in the case of disruptive selection, the minimum. Aster can also yield estimates of the position of maximum (or minimum) points, as our example shows, as well as their statistical uncertainty (Fig. 5); the accompanying technical reports (Geyer and Shaw 2008a,b, 2009, 2010a,b) fully describe details of estimation and inference for fitness landscapes.

Beyond evaluation of the form of selection on a single or few, moderately correlated traits, consideration of numerous, intercorrelated traits greatly compounds the challenge. Increasing the number of traits as predictors demands ever larger sample sizes to maintain reasonable statistical power and precision and consequently calls for careful consideration. When many correlated traits are under consideration, Lande and Arnold (1983) recommended use of Principal Components Analysis to reduce the number of traits to few mutually orthogonal composite traits. We argue against this approach because there is no reason to expect that selection operates on these linear combinations of trait values, or on any combinations close to them. We illustrate a statistically well-established approach to characterizing selection when many traits are being considered. Often, as in our example, it will not be possible to identify a single best relationship between fitness and the traits, ruling out all other models. It is then appropriate to acknowledge the uncertainties about the relationship between fitness and traits. Even in this situation, frequentist model averaging

(FMA) yields a single estimate of the fitness surface by averaging over the estimates from best set of models. Regardless of the degree of statistical confidence in the resulting estimate of the fitness landscape, it should be considered a hypothesis about which traits are under direct selection, subject to further investigation, for example, by manipulating trait distributions (Mitchell-Olds and Shaw 1987).

Travis (1989), noting the paucity of evidence for optimizing/stabilizing selection, identified aspects of empirical methodology common in studies of stabilizing selection that may account for failures to detect this mode of selection when it is in fact occurring. He mused: "Most of the issues discussed here [concerning the limited empirical support for optimizing selection] were outlined by Alan Robertson (1968) over 20 years ago. It is a sobering reflection on our discipline that we remain so far from satisfactory resolutions." Twenty years still farther on, the situation remains sobering. Valid inference about the nature of phenotypic selection will be advanced by studies that jointly and rigorously analyze the key components of fitness.

ACKNOWLEDGMENTS

We thank M. Turelli and J. Kingsolver for discussions, encouragement, and comments on the manuscript and A. Dykstra, S. Keller, C. Martin, K. Mercer, G. Quiram, J. Stanton-Geddes, and D. Warren, who also commented helpfully, clarifying many points. Comments of two anonymous reviewers and Jason Wolf further improved the manuscript.

LITERATURE CITED

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory 267–281.

Andrews, D. F., R. Gnanadesikan, and J. L. Warner. 1971. Transformations of multivariate data. Biometrics 27:825–840.

Arnold, S. J., and M. J. Wade. 1984a. On the measurement of natural and sexual selection: theory. Evolution 38:709–719.

———. 1984b. On the measurement of natural and sexual selection: applications. Evolution 38:720–734.

Barndorff-Nielsen, O. E. 1978. Information and Exponential Families. Wiley, Chichester, U.K.

Bernardo, R. J. 2001. What if we knew all the genes for a quantitative trait in hybrid crops? Crop Sci. 41:1–4.

Blows, M. W., and R. Brooks. 2003. Measuring nonlinear selection. Am. Nat. 162:815–820.

Bulmer, M. G. 1972. The genetic variability of polygenic characters under optimizing selection, mutation, and drift. Genet. Res. 19:17–25.

Bumpus, H. C. 1899. The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. Biol. Lectures, Woods Hole Marine Biol. Station 6:209–226.

Burnham, K. P., and D. R. Anderson. 2002. Model Selection and multimodel inference: a practical information-theoretic approach, 2nd ed. Springer-Verlag, New York.

Charlesworth, B. 1980. Evolution in age-structured populations. Cambridge Univ. Press, Cambridge.

Cook, R. D. 2007. Fisher lecture: Dimension reduction in regression (with discussion). Stat. Sci. 22:1–43.

Cox, D. R. 1968. Notes on some aspects of regression analysis. J. R. Stat. Soc., Ser. A 131:265–279.

Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory. Harper & Row, New York.

Endler, J. A. 1986. Natural Selection in the Wild. Princeton Univ. Press, Princeton, NJ.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer, Sunderland, MA.

Furnival, G. M., and R. W. Wilson, Jr. 1974. Regressions by leaps and bounds. Technometrics 16:499–511.

Geyer, C. J., and R. G. Shaw. 2008a. Supporting data analysis for a talk to be given at Evolution 2008 Univ. of Minnesota, June 20–24. Technical Report No. 669. Sch. of Statistics, Univ. of Minnesota. http://purl.umn.edu/56204.

———. 2008b. Commentary on Lande-Arnold analysis. Technical Report No. 670. Sch. of Statistics, Univ. of Minnesota. http://purl.umn.edu/56218.

———. 2009. Model selection in estimation of fitness landscapes. Technical Report No. 671, revised. Sch. of Statistics, Univ. of Minnesota. http://purl.umn.edu/56219.

———. 2010a. Hypothesis tests and confidence intervals involving fitness landscapes fit by Aster models. Technical Report No. 674, revised. Sch. of Statistics, Univ. of Minnesota. http://purl.umn.edu/56328.

———. 2010b. Aster Models and Lande-Arnold Beta. Technical Report No. 675 revised. School of Statistics, Univ. of Minnesota. http://purl.umn.edu/56394.

Geyer, C. J., S. Wagenius, and R. G. Shaw. 2007. Aster models for life history analysis. Biometrika 94:415–426.

Hand, D. J. 1981. Branch and bound in statistical data analysis. The Statistician 30:1–13.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning, 2nd ed. Springer-Verlag, New York.

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

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial (with discussion). Stat. Sci. 19:382–417. Corrected version available at http://www.stat.washington.edu/www/research/online/1999/hoeting.pdf.

Janzen, F. J., and H. S. Stern. 1998. Logistic regression for empirical studies of multivariate selection. Evolution 52:1564–1571.

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gilbert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations. Am. Nat. 157:245–261.

Lande, R., and S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.

Lindgren, B. W. 1993. Statistical theory, 4th ed. Chapman & Hall, New York.

Madigan, D., and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Am. Stat. Assoc. 89:1535–1546.

McGraw, J. B., and H. Caswell. 1996. Estimation of individual fitness from life-history data. Am. Nat. 147:47–64.

Mitchell-Olds, T., and R. G. Shaw. 1987. Regression analysis of natural selection: statistical inference and biological interpretation. Evolution 41:1149–1161.

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at: http://www.r-project.org.

Riani, M. 2004. Robust multivariate transformations to normality: constructed variables and likelihood ratio tests. Statistical Methods & Applications 13:179–196.

Rockafellar, R. T., and R. J.-B. Wets. 2004. Variational analysis, corr. 2nd printing. Springer-Verlag, Berlin.

Schluter, D. 1988. Estimating the form of natural selection on a quantitative trait. Evolution 42:849–861.

Schluter, D., and D. Nychka. 1994. Exploring fitness surfaces. Am. Nat. 143:597–616.

Schwarz, G. 1978. Estimating the dimension of a model. Annals. Stat. 6:461–464.

Shaw, R. G., C. J. Geyer, S. Wagenius, H. H. Hangelbroek, and J. R. Etterson. 2008. Unifying life history analysis for inference of fitness and population growth. Am. Nat. 172:E35–E47.

Stinchcombe, J. R., A. F. Agrawal, P. A. Hohenlohe, S. J. Arnold, and M. W. Blows. 2008. Estimating nonlinear selection gradients using quadratic regression coefficients: double or nothing? Evolution 62:2435–2440.

Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics, Theory and Methods A7:13–26.

Travis, J. 1989. The role of optimizing selection in natural populations. Ann. Rev. Ecol. Syst. 20:279–296.

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

van Tienderen, P. H. 2000. Elasticities and the link between demographic and evolutionary dynamics. Ecology 81:666–679.

Wade, M. J., and S. Kalisz. 1989. The additive partitioning of selection gradients. Evolution 43:1567–1569.

Weldon, W. F. R. 1895. An attempt to measure the death-rate due to the selective destruction of *Carcinus moenas* with respect to a particular dimension. R. Soc. Lond. B. 57:360–379.

Associate Editor: J. Wolf

Appendix: Monotonicity

We note that aster models, like generalized linear models, are linear in their canonical parameters, which have no simple interpretation. However, estimates of mean value parameters, for example, the expected value of the total number of offspring produced, can readily be obtained from estimates of the canonical parameters. A key result is that the relationship between the unconditional canonical parameter vector $\boldsymbol{\varphi}$ and the unconditional mean value parameter vector $\boldsymbol{\mu}$ of an aster model is strictly multivariate monotone (Rockafellar and Wets 2004, Definition 12.1), which means

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2) > 0 \qquad \text{(A1)}$$

whenever $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ are distinct unconditional canonical parameter vectors and $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are corresponding mean value parameter vectors (Barndorff-Nielsen 1978, Equation 28, p. 121).

Suppose expected fitness is deemed the sum $\sum_{j \in G} \mu_j$ of means of a subset of the response vector. For example, if some components are counts of offspring, then fitness is the sum of those counts. And suppose we model

$$\varphi_j(\mathbf{x}, \mathbf{z}) = \begin{cases} a_j(\mathbf{x}) + q(\mathbf{z}), & j \in G \\ a_j(\mathbf{x}), & j \notin G \end{cases} \qquad \text{(A2)}$$

where $\mathbf{z}$ denotes the vector of measured phenotypic variables for an individual and $\mathbf{x}$ denotes a vector of other covariates. If $\mathbf{z}_1$ and $\mathbf{z}_2$ are phenotype vectors for two individuals having the same covariate vector $\mathbf{x}$, then

$$q(\mathbf{z}_1) > q(\mathbf{z}_2) \quad \text{if and only if} \quad \sum_{j \in G} \mu_j(\mathbf{x}, \mathbf{z}_1) > \sum_{j \in G} \mu_j(\mathbf{x}, \mathbf{z}_2), \qquad \text{(A3)}$$

that is, individuals having higher values of $q(\mathbf{z})$ have higher expected fitness and vice versa. This argument is presented in detail in Section 3.2 and generalized in Appendix A of Geyer and Shaw (2008a).

This argument is the key to sound estimation of fitness landscapes. It allows statistical models that are both simple and valid, guaranteeing that expected fitness is a monotone transformation of the function used to model fitness on the unconditional canonical parameter scale.