*Evolution Letters*, 2024, **8**(3), 361–373
**https://doi.org/10.1093/evlett/qrad064**
Advance access publication 23 January 2024
**Letter**

# Regularized regression can improve estimates of multivariate selection in the face of multicollinearity and limited data

Jacqueline L. Sztepanacz[1], and David Houle[2],

[1]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada
[2]Department of Biology, Florida State University, Tallahassee, FL, United States

Corresponding author: Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada. Email: jsztepanacz@gmail.com

## Abstract

The breeder's equation, $\Delta\bar{\mathbf{z}} = \mathbf{G}\boldsymbol{\beta}$, allows us to understand how genetics (the genetic covariance matrix, $\mathbf{G}$) and the vector of linear selection gradients $\boldsymbol{\beta}$ interact to generate evolutionary trajectories. Estimation of $\boldsymbol{\beta}$ using multiple regression of trait values on relative fitness revolutionized the way we study selection in laboratory and wild populations. However, multicollinearity, or correlation of predictors, can lead to very high variances of and covariances between elements of $\boldsymbol{\beta}$, posing a challenge for the interpretation of the parameter estimates. This is particularly relevant in the era of big data, where the number of predictors may approach or exceed the number of observations. A common approach to multicollinear predictors is to discard some of them, thereby losing any information that might be gained from those traits. Using simulations, we show how, on the one hand, multicollinearity can result in inaccurate estimates of selection, and, on the other, how the removal of correlated phenotypes from the analyses can provide a misguided view of the targets of selection. We show that regularized regression, which places data-validated constraints on the magnitudes of individual elements of $\boldsymbol{\beta}$, can produce more accurate estimates of the total strength and direction of multivariate selection in the presence of multicollinearity and limited data, and often has little cost when multicollinearity is low. We also compare standard and regularized regression estimates of selection in a reanalysis of three published case studies, showing that regularized regression can improve fitness predictions in independent data. Our results suggest that regularized regression is a valuable tool that can be used as an important complement to traditional least-squares estimates of selection. In some cases, its use can lead to improved predictions of individual fitness, and improved estimates of the total strength and direction of multivariate selection.

**Keywords:** natural selection, sexual selection, multicollinearity, regularized regression, quantitative genetics

## Lay Summary

To understand how traits will evolve in populations, it is necessary to determine which traits are under natural or sexual selection, and the strength and direction of that selection. Organisms are comprised of many correlated traits and selection often acts on these traits simultaneously, necessitating multiple traits to be included in evolutionary predictions. Estimating selection on multiple correlated traits, however, is a well-known statistical challenge. In this article, we use simulations and reanalyses of published data to show that modern regularized regression methods can generate more accurate estimates of the total strength and direction of multivariate selection in the presence of correlated traits and have minimal cost when correlations between traits are low. Our results suggest that regularized regression can improve our understanding of some aspects of selection and should be employed alongside traditional least-squares approaches in selection analyses.

## Introduction

Understanding how selection acts on phenotypes is a fundamental challenge in evolutionary biology. Organisms are comprised of many correlated traits and selection acts on these traits simultaneously, necessitating a multivariate approach to the study of selection. In their seminal paper, Lande and Arnold (1983) outlined a method to estimate selection on correlated traits by performing a multiple regression of trait values on relative fitness. The vector of partial regression coefficients, $\boldsymbol{\beta}$, which are obtained from the model, estimates the magnitude and direction of selection acting directly on each trait when all traits under selection are included in the model. The Lande–Arnold approach has since been widely adopted to estimate selection in both wild and lab populations, producing hundreds of estimates of multivariate directional selection (Hereford et al., 2004). It has significantly advanced our understanding of how selection acts in nature.

One challenge to any multiple regression approach, which may be particularly prevalent in multivariate selection analyses, is that predictor variables can be highly correlated, making it difficult to disentangle their independent effects on a response variable (Figure 1). This situation is exacerbated by finite sample sizes and environmental variation that also contribute to uncertainty

Received March 25, 2023; revisions received November 19, 2023; accepted December 6, 2023

in multiple regression; however, our focus in this paper is on multicollinearity. Selection often acts on functional suites of traits that are correlated, such as the cuticular hydrocarbons of insects (Blows & Allan, 1998), volatile floral compounds that produce a scent (Chapurlat et al., 2019), or components of an insect's song that are combined to attract mates (Hoikkala et al., 1998; Hoy et al., 1988; Talyn & Dowse, 2004). We would like to be able to disentangle the independent effects of each correlated trait on fitness to identify the targets of selection.

The issue of multicollinearity in selection analyses was documented by Lande and Arnold (1983) who emphasized that when traits under selection are perfectly correlated, $\boldsymbol{\beta}$ cannot be estimated using the full data. While perfectly correlated traits may be unlikely to occur, highly correlated sets of traits are common, and when there are more than two traits under selection, visually inspecting all the bivariate correlations between them is uninformative of the overall correlation structure, or multicollinearity, in the data. When predictor traits are highly correlated, the estimates of selection gradients have large standard errors. Consider the algebra of ordinary least-squares (OLS) regression:

$$w = \mathbf{X}\boldsymbol{\beta} + e$$

where $w$ is the vector of relative fitnesses of individuals (individual fitnesses divided by population mean fitness), $\mathbf{X}$ is an $n \times p$ matrix of $p$ traits measured on $n$ individuals, $\boldsymbol{\beta}$ is the vector of partial regression coefficients that we aim to solve for, and $\boldsymbol{e}$ is the residual variation not accounted for by the model. Assume that each trait is centered on its mean and divided by its standard deviation, so all the traits have an equal variance $(\sigma^2)$ of one. Solving for $\boldsymbol{\beta}$, the estimated partial regression coefficients (selection gradients) are:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'w,$$

where′ denotes the matrix transpose, and $^{-1}$ denotes the matrix inverse. $\mathbf{X}'\mathbf{X}$ is the sums of squares and cross product matrix of predictors, which indicates how much information about the estimated coefficients is contained in the data. The amount of independent information in the data is indicated by the eigenvalues of $\mathbf{X}'\mathbf{X}$, with larger eigenvalues signifying more information. The inverse of $\mathbf{X}'\mathbf{X}$ determines the variance of the estimated partial regression coefficients and consequently their standard errors:

$$V\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

To uniquely estimate the regression coefficients, every linear combination of the columns of $\mathbf{X}'\mathbf{X}$ must describe independent information in the data. There are at least two reasons why one or more linear combinations of $\mathbf{X}'\mathbf{X}$ may carry no independent information, signified by a corresponding eigenvalue that is very low or zero: (a) the number of individuals measured is less than the number of traits ($n < p$), or (b) one or more of the traits is a linear combination of the others. For many selection analyses, (a) is not likely to be an issue, as many more individuals than traits are typically measured in a study. However, with the increasing use of phenomic technologies (Houle, 2010), it is possible to measure hundreds, or even thousands of traits, which is likely to become an increasing problem. Multicollinearity between traits may often occur, however, particularly if the selection of many related traits is the focus of the study.

Multicollinearity is often diagnosed post hoc by the variance inflation factor (VIF) or related metrics that are produced by regression software. VIFs indicate the increase in variance of regression coefficients due to multicollinearity. A VIF of 8, for example, means that the variance of a regression coefficient is eight times larger than it would be in the absence of multicollinearity. Traits that have high VIFs, based on arbitrary cutoffs such as 10 (e.g., Belsley et al., 2005), are often discarded from the data, and multiple regression analyses rerun to estimate selection on the remaining traits (Mitchell-Olds & Shaw, 1987), or other more complicated methods to estimate selection such as principal component (PC) regression (Lande & Arnold, 1983) or partial least squares are used. However, the interpretability of the partial regression coefficients as estimates of selection from these models is not always straightforward (but see Chong et al., 2018). Removing traits with high VIFs leads to biased estimates of selection on the remaining traits (Morrissey & Ruxton, 2018). We illustrate this in Figure 1 where the sign of the parameter estimate changes when a correlated predictor is omitted from the regression. Estimates of selection on the principal components

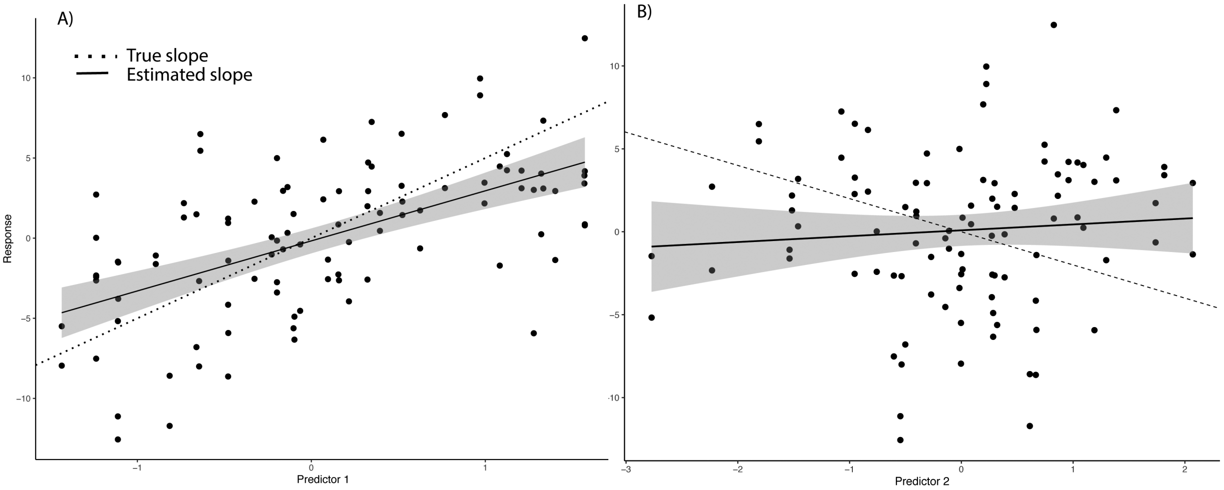

**Figure 1.** Regression of predictor variables (A) X1 and (B) X2 on a response variable from simple linear regression. The true partial regression coefficients of predictors 1 and 2 are 5 and −2, respectively, and the correlation between X1 and X2 is 0.7. When correlated predictors are left out of the model, estimated regression coefficients are biased in magnitude (panels A and B) and sign (panel B). The estimated partial regression coefficient for X1 when X2 is omitted from the model is 3.12 ± 0.45, and the estimated partial regression coefficient for X2 when X1 is omitted from the model is 0.35 ± 0.47.

(PCs) can be hard to interpret because the estimate of selection is on composite traits that are constructed regardless of their importance to fitness (Mitchell-Olds & Shaw, 1987). Although it is straightforward to transform selection estimates on PCs back to the original trait space, the resulting estimates of selection on the original traits are biased (Chong et al., 2018). Projection pursuit regression (Friedman & Stuetzle, 1981) can also help solve interpretability problems by defining orthogonal axes of the multivariate phenotype that maximize the explained variation in fitness (Schluter & Nychka, 1994). However, when this approach is implemented in a reduced data space, the estimates of selection are also biased.

These methods all aim to solve the problem of multicollinearity in selection analyses with the common side effect of producing biased estimates of selection. Multiple regression, however, correctly assesses the effect of multicollinearity in the uncertainty of the parameter estimates it produces, and yields unbiased estimates of selection that are frequently favorable. For example, the goal of many selection analyses is to determine which traits have a direct effect on fitness and what that effect is. For this question, multiple regression produces the appropriate unbiased estimates and their uncertainty. When the estimated coefficients from selection studies are to be used as inputs in further statistical analyses, such as meta-analyses of the strength of selection (e.g. Kingsolver et al., 2001; Morrissey, 2016), it is also important that estimates be unbiased and have a measure of uncertainty, which multiple regression produces. For these reasons, Morrissey and Ruxton (2018) argued that using biased regression methods will be a detriment to biological studies, such as selection gradient analyses, where the values of the individual selection coefficients and their uncertainty are of interest.

For certain questions and studies of selection, however, obtaining unbiased estimates may be of less importance to researchers than obtaining accurate estimates. Methods that decrease the variance of an estimator more than they increase the square of the bias will reduce the mean square error (MSE) of the estimator and increase its accuracy. This is the well-known bias–variance trade-off (Hastie et al., 2009). In many biologically relevant cases, we may be able to obtain more accurate estimates of selection, by introducing bias toward estimates that are *a priori* more reasonable. We already discussed two classes of methods that are biased but may increase the accuracy of estimated selection: subset selection and dimension reduction. Regularization, which is also known as shrinkage, is a third class of methods that may increase the accuracy of estimated coefficients, although it has not been widely applied in selection analyses (but see Morrissey, 2014). Regularized regression fits a model with all of the original traits, but their estimated coefficients are shrunken toward (or to) zero compared to the OLS estimates. This reduces the variance of the estimated coefficients, while also biasing the estimates toward zero. Depending on the relative changes in variance and bias, this can either increase or decrease the accuracy of estimated coefficients.

Ridge regression (Hoerl & Kennard, 1970), the lasso (Tibshirani, 1996), and their combination, the elastic net (Zou & Hastie, 2005), are the common types of regularization that are similar to OLS regression, but add a penalty to the sum of squared errors (ridge) or absolute values (lasso). They find the best-fit regression coefficients by minimizing the following functions:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where $n$ is the number of observations, $p$ is the number of traits, $\beta_0$ is the intercept, and $\lambda$ is the tuning parameter, chosen by the researcher, that determines the relative impact of the usual residual sums of squares vs. the regularization penalty (Hastie et al., 2009).

In the case of ridge regression, where the penalty is applied to squared coefficients (L2 regularization), larger coefficients are disproportionately shrunken compared to small coefficients. In the presence of multicollinearity, these large coefficients also tend to be the ones that are estimated with the most error. In ridge regression, all squared coefficients are shrunk toward zero, but do not become zero unless $\lambda$ is infinity. Therefore, all of the traits with nonzero gradients in an OLS analysis will still have nonzero gradients following ridge regression. Lasso regularization penalizes the absolute value of the coefficients (L1 regularization) rather than their square (Tibshirani, 1996). Consequently, some of the estimated coefficients may be exactly zero when $\lambda$ is sufficiently large, and thereby the lasso performs both shrinkage and variable selection. The elastic net is a combination of ridge and lasso penalties allowing the relative weighting of the two penalties to be adjusted either by choice of the investigator or according to some criteria, such as cross-validation. The elastic net thus can result in both shrinkage and variable selection. Ridge and lasso penalties can also be expressed as Bayesian priors over the model parameters (e.g., ridge = Normal prior; lasso = Laplace prior), and indeed there are several other Bayesian forms of regularization (Gianola, 2013; Melo et al., 2019; Park & Casella, 2008). Bayesian analyses allow for a more flexible set of models to be implemented through a variety of prior distributions that may better match the data, and they also yield uncertainty in their estimates through the posterior distribution. However, this flexibility can come with optimization problems and a high computational cost (e.g., Celeux et al, 2012). In this article, we specifically focus on ridge, lasso, and elastic net penalties in a frequentist framework.

It is important to state at the outset that different aspects of selection may be of primary interest in each study, and consequently, there is no single measure of accuracy and no single regression method that is appropriate to apply across the board to selection analyses. We focus on four different measures of accuracy that capture aspects of the total strength and direction of a multivariate selection vector, and that are likely to be of general interest to evolutionary biologists: (a) The proportional error in the total estimated strength of selection, given by $\frac{\|\hat{\boldsymbol{\beta}}\|}{\|\boldsymbol{\beta}\|}$, (b) the MSE of the estimated multivariate selection gradient, given by the Euclidian distance between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, scaled by the number of traits $\left(\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| / n\right)$, (c) divergence in the direction of estimated and true selection, given by the angle between the vectors $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, and (d) the proportion of individual coefficients (all coefficients and those with $p < .05$ in a standard multiple regression) that estimate selection in the correct direction (i.e., have the correct sign).

Here we investigate the utility of regularized regression for the estimation of the strength and direction of multivariate vectors of linear selection, and give guidance on when these approaches are more accurate compared to OLS regression using these four measures of accuracy. We first assess the frequency and extent of multicollinearity in a sample of studies. Next, we simulate data where the true selection gradient is known, apply the different regression approaches, and quantify the accuracy of each

**Table 1.** Recent studies that estimate multivariate selection using the Lande–Arnold multiple regression framework and that have phenotypic data available either within the article or on the Dryad data repository.

| Study | Species | Number of traits | Trait type(s) | Fitness component(s) | Approximate sample size | Mean VIF | Max. VIF |
|---|---|---|---|---|---|---|---|
| Poissant et al. (2016) | *Parus major* | 4 | Morphological | Lifetime reproductive success | 986 males; 1,095 females | 1.1 | 1.1 |
| Brachi et al. (2012) | *Arabidopsis thaliana* | 7 | Phenological | Total fruit length | 688 | 1.1–10.4 | 27.5 |
| Bartkowska & Johnston, (2015) | *Lobelia cardinalis* | 6 | Phenological and morphological | Total seeds produced per individual | 860 | 2, 2.6 | 4.9 |
| Lindholm et al. (2014) | *Poecilia reticulata* | 7 | Morphological and Reproductive | Number of offspring sired | 30 | 2–4.5 | 9.1 |
| Walker et al. (2014) | *Notiomystis cincta* | 9 | Plumage color and morphological | Number of mates, number of fertilizations, number of fledglings | 79 | 1.2, 3.7 | 6.9 |
| *Sztepanacz and Rundle (2012) | *Drosophila serrata* | 8 | Chemical communication | Binomial mating success | 1,978 | 5.4 | 15.4 |
| Sanjak et al. (2018) | *Homo sapiens* | 25 males 31 females | Biomedical | Lifetime reproductive success | 217,728 females; 158,638 males | 122.9 | 1944 |
| Chapurlat et al. (2019) | *Gymnadenia conopsea* | 19 | Chemical communication, phenology, morphology | Number of fruits × fruit mass | 139; 169 | 3.0, 2.3 | 9.0 |
| *Angell et al. (2020) | *Protopiophila litigata* | 17 | Chemical communication | Binomial mating success | 186; 234 | 21 | 52.3 |
| *Chong et al. (2018) | *Arabidopsis thaliana* | 4 | Phenological | Fruit production | 50 (line means) | 6.2 | 12.5 |
| Morrissey (2014) | *Ovis aries* | 4 | Morphological | Survival | 846 males; 398 females | 2.8 | 4.5 |
| Oh and Shaw (2013) | *Laupala cerasina* | 3 | Courtship song | Binomial acoustic preference | 73 | 1.0 | 1.0 |

*Note.* Studies were haphazardly chosen and are not an exhaustive sample. Mean and max variance inflation factor (VIF) are the average and maximum VIF of predictors in the model. *Studies that were analysed using regularized regression in 'Empirical case studies section'

approach. Finally, we reanalyze data from the three published studies to compare the estimates of selection from standard, principal component, and regularized regression, and the ability of each model to correctly predict fitness in independent data.

# Methods

## Quantifying multicollinearity

The VIF is a metric of multicollinearity that is commonly produced by most regression packages and indicates how much the estimated variance of a regression coefficient is increased above what it would be if that trait was uncorrelated with all others in the model. The VIF for trait *i* is

$$\mathrm{VIF}_i = (n-1)\,\sigma_i{}^2 * (\mathbf{X}'\mathbf{X})^{-1}{}_{ii} = \mathbf{P}^{-1}{}_{ii}$$

where **P** is the phenotypic covariance matrix of the traits measured. VIFs produce a sensible measure of the effect of multicollinearity on the precision of the estimate for each regression coefficient. However, arbitrary rules of thumb, such as discarding variables with a VIF greater than some constant are not very sensible (O'brien, 2007).

Unfortunately, many published selection studies do not include the phenotypic covariance matrix necessary for the calculation of the VIF, making it difficult to determine the pervasiveness of multicollinearity. We acquired a handful of relevant studies from different systems to quantify the extent of multicollinearity observed in empirical data. We haphazardly searched for multivariate selection studies that presented the phenotypic covariance matrix of the traits measured, or which had the raw data available as a supplement to the article or published in the Dryad data repository. We identified 10 papers (Angell et al., 2020; Bartkowska & Johnston, 2015; Brachi et al., 2012; Chapurlat et al., 2019; Lindholm et al., 2014; Morrissey, 2014; Oh & Shaw, 2013; Poissant et al., 2016; Sanjak et al., 2018; Sztepanacz & Rundle, 2012; Walker et al., 2014) that estimated multivariate selection on 3–31 traits and calculated the average and range of VIFs for traits included in the published analyses (Table 1). One caveat of this approach is that many studies may have already excluded correlated traits during exploratory analyses (e.g., Chapurlat et al., 2019), leading to an underestimation of the VIFs likely to be found for suites of traits.

## Data simulations

### *Proof of concept simulation*

We start with a proof-of-concept example to demonstrate the behavior of multiple and regularized regression in the absence and presence of multicollinearity. For each simulation, we simulated *n* records of the three traits from a multivariate normal distribution with a covariance structure of an identity matrix **I**, or mid (Supplementary Table S1) and high multicollinearity (Supplementary Table S2), respectively. In the presence of mid multicollinearity, the VIFs for the three traits were 37.3, 5.7, and 27.0, and the average VIF was 23.3. In the presence of high multicollinearity, the VIFs were 79.8, 11.8, and 56.7, and the average VIF was 50.1. This is a high average multicollinearity compared to the published studies shown in Table 1. The simulated selection gradient **β** was [–0.18, 0.30, –0.60], and we drew unique gradients of fitness with respect to phenotype from a Poisson or binomial distribution. Because of the nonlinear mapping between fitness and phenotype in these simulations, the simulated selection gradient is not equal to the true gradient in each simulation. To determine the true selection gradient in each simulation we followed the approach in Morrissey (2014), calculating the selection gradient as the partial derivative of trait value with respect to relative fitness for each trait by finite differences. We generated $10^6$ records of phenotype according to the true value of **P** for each simulation, and then for every individual calculated expected and absolute fitness and averaged them to obtain population mean fitness. For each of the three traits, we recalculated population mean fitness after adding and subtracting 0.03 from each phenotypic record, and then took the difference to determine the partial derivative of trait value with respect to relative fitness.

### *Empirically informed simulations*

The estimates of selection and their standard errors produced by the standard Lande–Arnold approach of multiple regression are influenced by four key parameters: the size of the population sampled, *n*; the number of traits measured, *p*; the phenotypic covariance structure among the traits, **P**; and the distribution of the fitness component measured. To assess the utility of regularized regression in selection gradient analyses more generally, we simulated data spanning an empirically informed parameter space, which represents realistic scenarios that may be encountered in selection studies that are undertaken in the field or the lab. We simulated all combinations of (a) sample size, $n$ = 100, 400, 1,000, (b) number of traits, $p$ = 4, 7, 12, 17, (c) low, and mid/high eigenvalue dispersions of **P** which were informed by the empirical **P** collected in Table 1, and (d) fitness measures which were either binomial or Poisson distributed to represent common scenarios such as mating success and seed count/offspring number, respectively (Supplementary Table S3). For each scenario, we carried out two sets of simulations. In the first set of simulations, all the traits were simulated to be under selection. In the second set of simulations, at least one trait was always under selection, but the other included traits were not all necessarily under selection. The number of traits not under selection was randomly determined in each simulation.

For each simulation, we simulated unique **P** from an inverse Wishart distribution with scale matrix **S** and degrees of freedom $\upsilon$, to span the range of multicollinearity we observed in Table 1. Each **P** was standardized to unit variance for each trait. For each simulation, we simulated *n* records of *p* traits, **z**, with a mean of 0 and covariance **P** from a multivariate normal distribution. We drew unique gradients of fitness with respect to phenotype, **b**, from a normal distribution ($\mu = 0, \quad \sigma = 3$) and normalized **b** to unit length. We then simulated individual fitness from a Poisson or binomial distribution with expected values of $\exp(\mathbf{zb}')$ or $\text{logit}^{-1}(\mathbf{zb}')$, respectively. To calculate the true selection gradients $\boldsymbol{\beta}$ for each simulation, we used the method of finite differences described for the proof-of-concept example above.

## Analyses

### *Proof of concept and empirically informed simulations*

We estimated the vector of linear selection gradients, $\hat{\boldsymbol{\beta}}$, for each of the simulated data sets using least-squares multiple regression, ridge regression, lasso, and the elastic net. Our analyses were implemented in R (R Core Team, 2023) using the lm function for least-squares multiple regression, and the elastic net package (Friedman et al., 2010) for regularized regression. The shrinkage parameter $\lambda$ for regularized regressions was chosen using 10-fold cross validation (CV) to minimize the MSE of the models. Equal weight ($\alpha = 0.5$) was given to ridge and lasso penalties for elastic net regression. The response variables were either Poisson or binomially distributed, however, we fit linear models with a Gaussian function for all methods to make them comparable to the standard Lande–Arnold regression approach.

### *Empirical case studies*

For 3 of the 10 studies in Table 1 that are identified with an asterisk, we reanalyzed the data using standard linear regression and regularized regression approaches. Each case study data set was randomly split into an estimation (85% of the data) and a test set (15% of the data), with the estimation set used to estimate standard and regularized selection gradients, as described in the previous section. For regularized models, the shrinkage parameter $\lambda$ was chosen using 10-fold CV to minimize the MSE of the models, and for elastic net regression equal weight was given to ridge and lasso penalties ($\alpha = 0.5$). The best-fitting model from the estimation set was used to predict fitness in the test set. To obtain confidence intervals on the predictions, we bootstrapped the analysis 500 times, redrawing unique estimation and test sets. We quantified the accuracy for models with binomial fitness measures by classification accuracy, and positive and negative predictive values. For Poisson distributed fitness, we classified the accuracy by the percent variance of true fitness that was explained by predicted fitness in the test data ($R^2$).

# Results

## Published selection gradients

Most studies of multivariate selection that we surveyed focus on relatively few traits (three to eight traits) and have modest data sets of tens to hundreds of observations (Table 1). A few studies have thousands of observations, while one study of humans includes hundreds of thousands of observations. The median of mean VIFs over studies is 5.4, while the median of the maximum VIF over studies is 8.5. Two studies have average VIFs over 10, and these are both studies that include large numbers of traits (Angell et al., 2020; Sanjak et al., 2018). Five studies have maximum VIF values over 10. One possible factor shaping VIF in these studies is the exclusion of highly correlated predictors a priori, a type of exploratory data filtering not necessarily described in the methods of papers.

## Proof-of-concept

The results of the proof-of-concept example are shown in Figure 2. We examined the accuracy of each method of analysis in response to the average VIF of the model. VIFs reflect the combination of sample size, trait number, and data structure that leads to multicollinearity, and therefore, provide a more complete assessment of multicollinearity than any one of these individual parameters we varied. In cases where there was no multicollinearity in the simulated data, all regression methods performed well. For Poisson distributed fitness, there is a small upward bias in the total strength of estimated selection at low multicollinearity for all regression methods and for multiple regression with binomial fitness (Figure 2A). Some upward bias is expected, as a consequence of taking the L2 norm of a vector estimated with error in a finite sample (Morrissey, 2014). As the sample size increases, the magnitude of this bias should decrease. For high multicollinearity, multiple regression performed poorly for both fitness distributions, overestimating the total strength of selection by as much as three times its true value (Figure 2A and B).

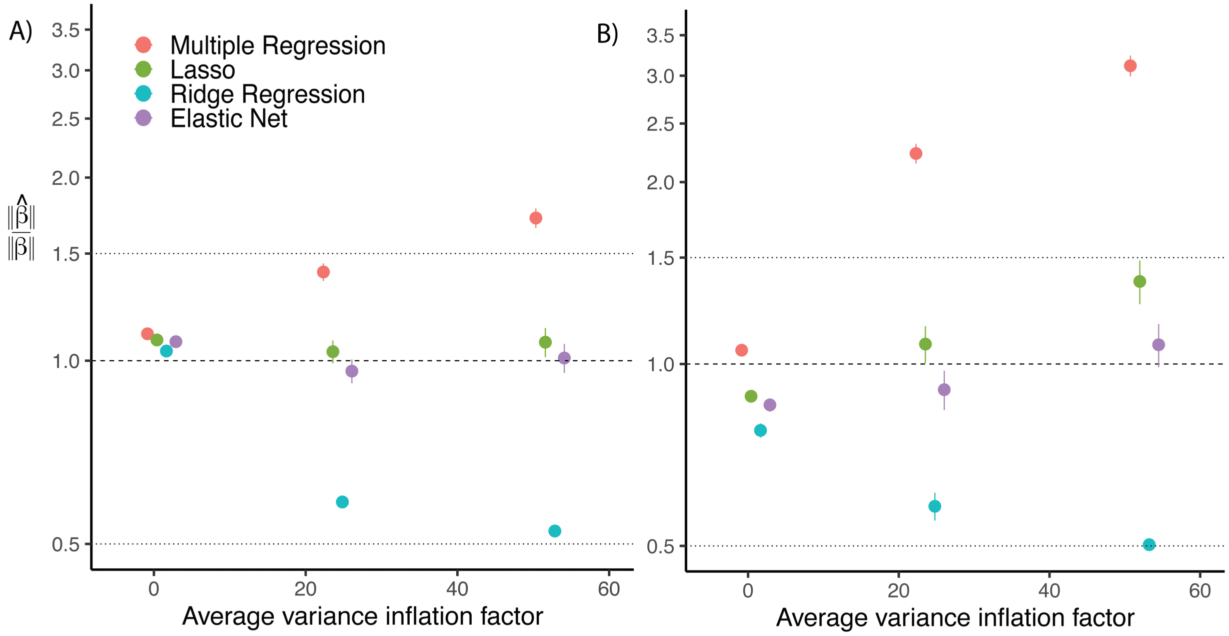

**Figure 2.** Proportional error in the estimated length of β (total strength of selection) when fitness has a (A) Poisson or (B) binomial distribution, as a function of the average variance inflation factor (VIF) of predictors. Points depict the mean of 300 simulations of each parameter combination and bars show the 95% confidence intervals of the estimates.

Ridge regression, on the other hand, consistently underestimated the total strength of selection, but by less than 50%. Estimates were further from the truth for all methods when fitness was binomially distributed for all methods, and this behavior was exacerbated when multicollinearity was high.

## Empirically informed simulations

As above, we examined the accuracy of each method of analysis as a function of the average VIF of the model. The performance of all regression methods in terms of the error in the total estimated strength of selection is shown in Figure 3A–D. In all simulations, regularized models produce better estimates of the total strength of selection than Lande–Arnold multiple regression when VIF is high. Multiple regression tends to overestimate the total strength of selection and overestimate it more than regularized models at moderate to high VIF. The difference between the methods is small when multicollinearity is low, with OLS regression performing better than regularized regression in some cases. Overall, however, regularized models perform as well at high multicollinearity as multiple regression does at low multicollinearity.

The type of regularization that performs best depends on the distribution of the fitness component measured. Ridge regression was best when fitness had a Poisson distribution (Figure 3A and C) and lasso was best when fitness was binomially distributed (Figure 3B and D). In the simulations, the weighting of L2 vs. L1 regularization ($\alpha$) in the elastic net was fixed at 0.5. There was not much difference between lasso and elastic net over the range of multicollinearity or fitness distributions studied. Overall, the total strength of selection was more poorly estimated by all methods when fitness is binomially distributed (Figure 3C and D) and when some traits that are included in the model are not under selection (Figure 3D).

The performance of all regression methods in terms of our measure of MSE is shown in Supplementary Figure S1A–D. At low multicollinearity, all regression methods perform similarly, but regularized regression performs slightly worse at the lowest VIF than OLS regression. With high VIF, OLS always showed the highest MSE, and ridge regression always performed the best. When all traits were subject to selection, lasso and elastic net regressions had slightly higher MSE than ridge regression, but when some traits were not under selection, MSE of lasso and elastic net regressions was markedly higher at high VIF (Supplementary Figure 1C and D). This behavior is unexpected as one of the potential advantages of lasso and elastic net is to perform feature selection or shrink some coefficients to zero.

Regularized and multiple regression both estimated similar directions of multivariate selection (Figure 4). When the average VIF of the traits was less than 10 there was little difference between the methods. At high multicollinearity (VIF ≥ 40) regularized models were closer, on average, to the true direction of selection by about 10 degrees when sample sizes were low and only a few degrees when sample sizes were high. However, the confidence intervals on these estimates are larger than the average difference between the methods.

The final measure of accuracy we considered was whether the estimated gradients had the correct sign. The results for simulations where all traits were under selection are shown in Figure 5. The lasso and elastic net can shrink estimates to 0, and when they did, we considered this to be a direction error when there was simulated selection on that trait. Similarly, if the simulated selection on the trait was 0 and the method estimated a nonzero gradient that was also considered a direction error, although this rarely happened in this set of simulations. We considered the accuracy of all individual gradients together, and the accuracy of the subset of gradients that were far enough from zero to be statistically supported in a multiple regression. On average, all methods did a good job of estimating the direction of the subset of individual gradients that were statistically supported in multiple regression ($p < .05$) (Figure 5A and B solid bars). However, multiple regression and ridge regression had the lowest estimation error, suggesting they are superior to lasso and elastic net. When considering all coefficients together (Figure 5A and B transparent bars), ridge

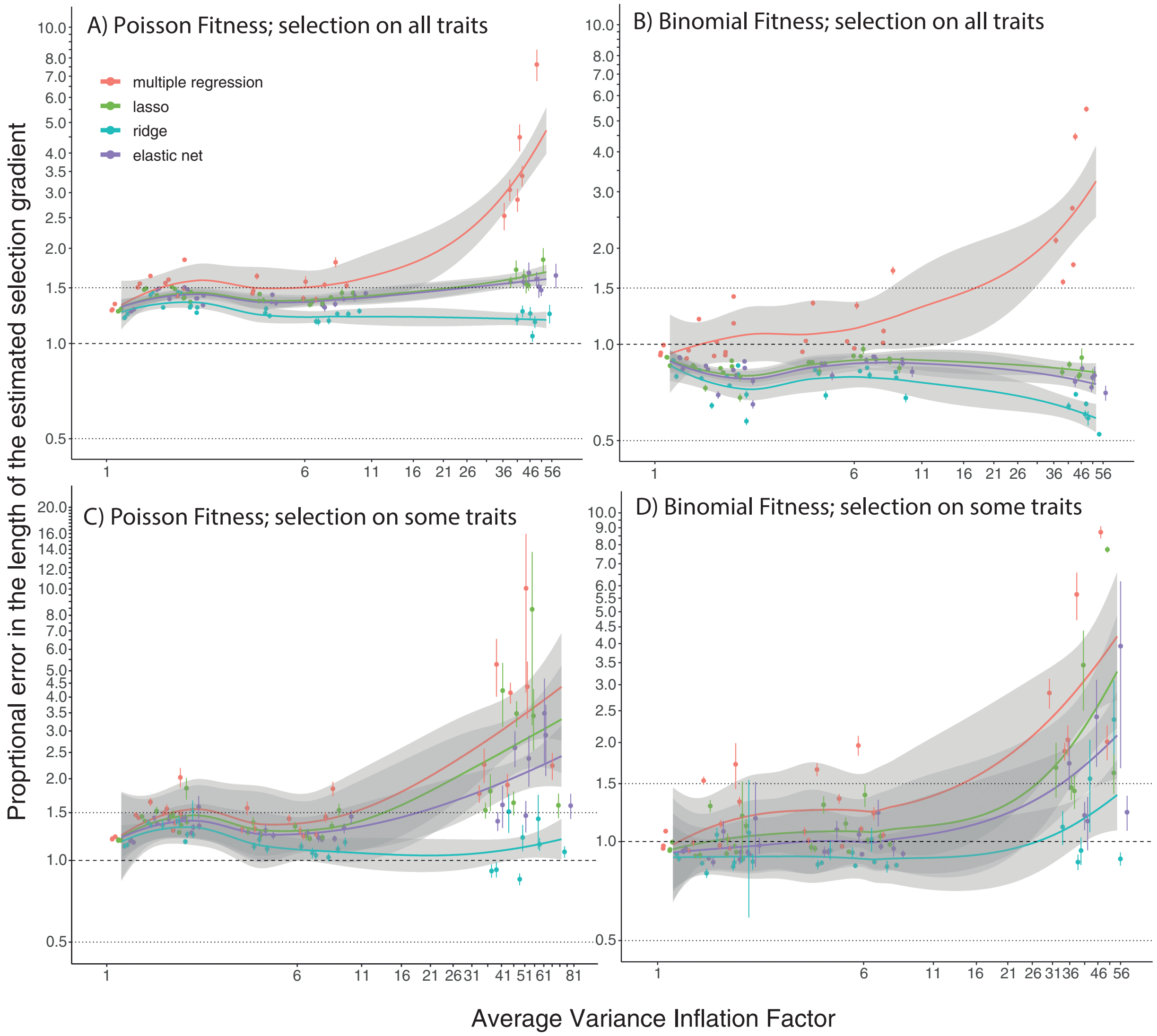

**Figure 3.** Proportional error in the estimated length of β (total strength of selection) when fitness has a binomial or Poisson distribution as a function of the average VIF of predictors. Points depict the mean of 300 simulations of each parameter combination and bars show the 95% confidence intervals of the estimates. Line with shading shows a LOWESS smooth curve ± one standard error. "Top panel" shows results for simulations where all traits are under selection and have a (A) Poisson (B) binomial distribution. "Lower panel" shows results from simulations where only some traits are under selection and fitness has a (C) Poisson (D) binomial distribution. Note the difference in scale of the y-axis for different panels.

regression and elastic net outperformed multiple regression when multicollinearity was high, and performed similarly to each other and to multiple regression when multicollinearity was low. Lasso performed the worst in all cases, and in particular, when all gradients (regardless of statistical support) were considered together. This is because in Figure 5, no traits were truly unselected, so any coefficients shrunk to 0 were treated as an error.

The results for simulations where some of the traits included in the model are not under selection are shown in Figure 6. First, we considered classification accuracy for lasso and elastic net methods, i.e., did the model correctly include (exclude) a trait that was (not) under selection (Figure 6A). Ridge and multiple regression estimate selection on all traits in the model, precluding a meaningful metric of classification accuracy for these methods. Both lasso and elastic net performed similarly to each other and for both distributions of fitness. On average, classification accuracy was between 50% and 60%, and this did not depend on the multicollinearity of the traits. We next considered how well each regression method performed at estimating the correct direction of selection (Figure 6B). The estimated selection gradient was considered incorrect if it was estimated to be nonzero for traits that were not under selection, and if it was zero or of different sign for traits under selection. On average the proportion of gradients estimated in the correct direction was less than 50% across the range of multicollinearity studied. All methods performed similarly poorly to each other and for both distributions of fitness, and much worse than in simulations where all traits were under selection.

## Case studies

We reanalyzed data from two case studies where the fitness measure was binomially distributed and one where fitness was Poisson distributed, using multiple and regularized regression. We do not know the true selection gradient in real data, so we quantified the performance of each model by how well it predicted fitness in the test data. For the Sztepanacz and Rundle (2012) data, there was no difference in predictive ability between any of the regression methods (Table 2). The estimated selection gradients were very

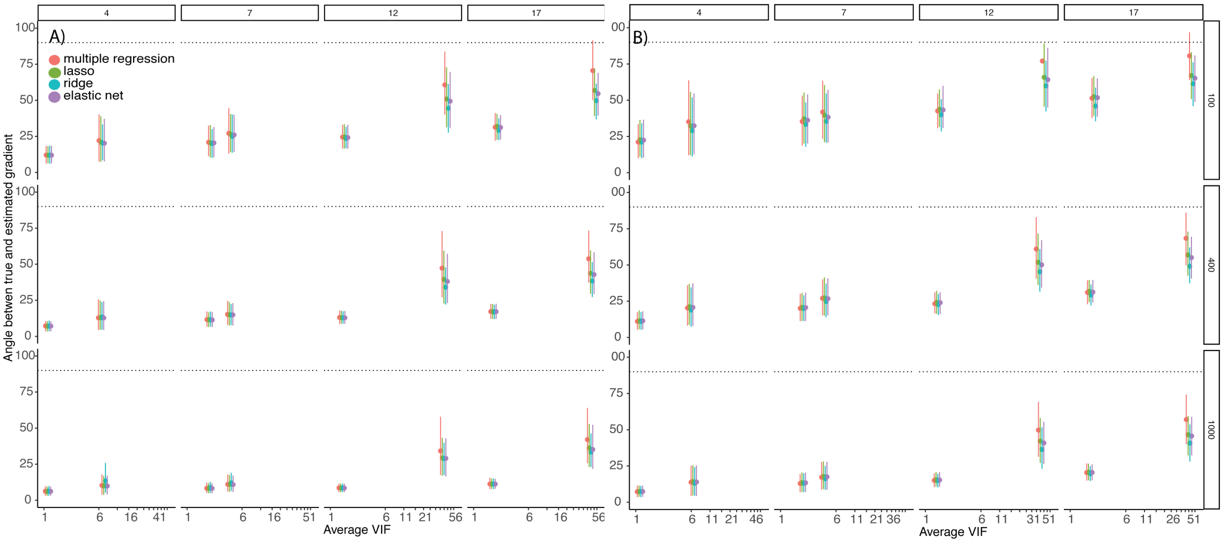

**Figure 4.** Angle between the vector of estimated multivariate selection and the true selection gradient (accuracy of the estimated multivariate direction of selection) when fitness has a (A) Poisson and (B) binomial distribution. Points depict the mean of 300 simulations of each parameter combination and bars show the 95% confidence intervals of the estimates.

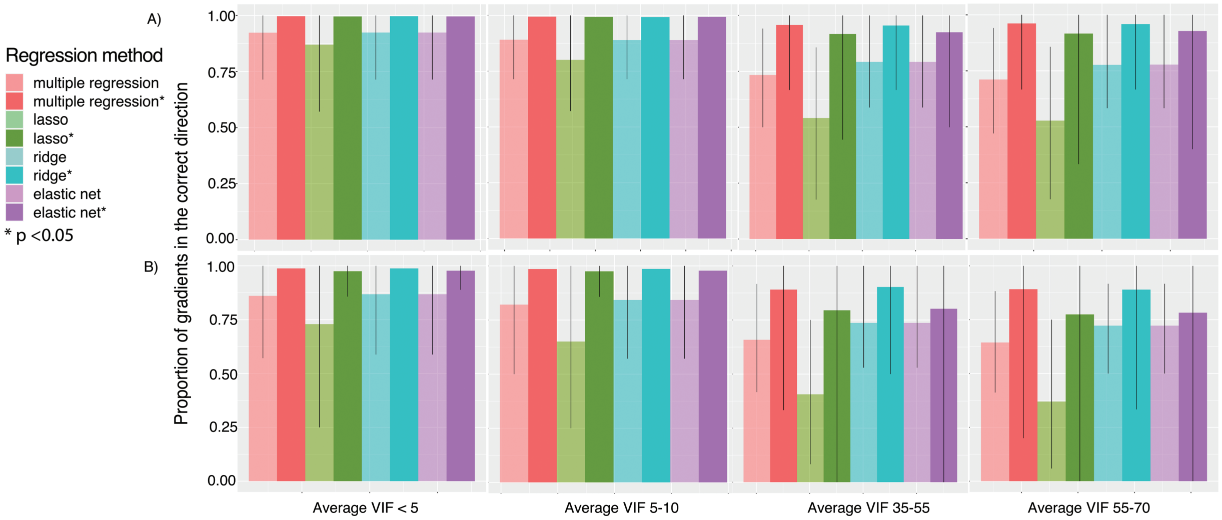

**Figure 5.** The proportion of individual selection gradients estimated to be in the correct direction when there is selection on all traits for (A) Poisson and (B) binomial distributed fitness. Transparent bars with 95% confidence intervals are determined with all estimated gradients. Solid bars with 95% confidence intervals are determined for estimated gradients that have a p-value <.05 in a standard multiple regression.

similar, and consistent in both magnitude and sign between the methods. For these data, regularization does not appear to be an improvement or cost compared to standard Lande–Arnold regression. This is not surprising, considering the average VIF of these data was relatively low (5.4), eight traits were studied, and the sample size was very large compared to most studies ($n$ = 1,978).

In the original analysis of Angell et al. (2020), selection was estimated on the first 9 of 17 principal components and transformed to estimates of selection on individual traits (Fig. 3A of Angell et al., 2020) using the approach outlined in Chong et al., (2018). We show these gradients in Table 3 alongside our regularized estimates. The VIF was 21 in these data and the sample size was 186 individuals. In the original paper, the authors found statistical support for selection acting on 10 of 18 traits studied using the Chong et al., (2018) approach. For three of these traits, OLS and the three regularized regression methods estimated selection in the opposite direction. Despite the differences in the estimated selection gradients, all of the regression methods, including Lande–Arnold regression, predicted fitness with similar accuracy as indicated by the positive and negative predicted values (Table 3).

Chong et al., (2018) also estimated selection gradients on principal component scores (first 4 of 5 PCs) and transformed them into the original traits. Their results suggested that selection on flowering time and flowering duration was of equal strength and in opposite directions. Ridge regression yielded a similar result, with selection for earlier flowering and for longer flowering time of approximately equal magnitude (Table 4). Lasso regression, on the other hand, estimated selection for earlier flowering that was about twice as strong, but estimated no selection on flowering duration. Using the ridge regression model to predict fitness in an independent set of data explained more of the variation in fitness than the lasso model, indicating that it is a better model than the lasso in this case. All of the regularized models had

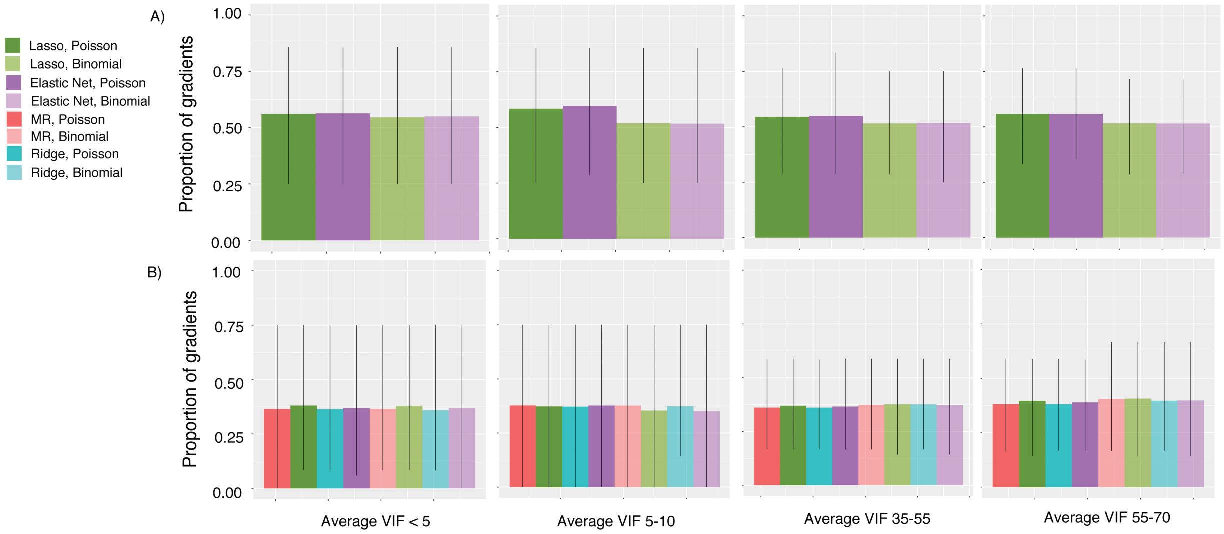

**Figure 6.** (A) Proportion of individual selection gradients that were correctly selected (or not) as being under selection all traits included in the analysis are under selection. (B) Proportion of individual selection gradients that were estimated in the correct direction when only some traits included in the analysis are under selection. Height of the bar is the average proportion with 95% confidence intervals from 300 simulations. Classification of feature accuracy was from a contingency table.

**Table 2.** Estimated sexual selection gradients on male cuticular hydrocarbons of *Drosophila serrata* from Sztepanacz and Rundle (2012).

| Regression method | | OLS | Lasso | Ridge | Elastic net |
|---|---|---|---|---|---|
| | **Coefficients (SE)** | | | | |
| | l2 | **0.066 (0.031)** | 0.062 | 0.060 | 0.066 |
| | l3 | –**0.074 (0.290)** | –0.068 | –0.069 | –0.071 |
| | l4 | 0.005 (0.044) | . | –0.002 | 0.001 |
| | l5 | –0.043 (0.050) | –0.010 | –0.034 | –0.012 |
| | l6 | –**0.232 (0.058)** | –0.209 | –0.212 | –0.214 |
| | l7 | 0.060 (0.085) | – | 0.041 | – |
| | l8 | **0.557 (0.041)** | 0.545 | 0.538 | 0.551 |
| | l9 | –**0.234 (0.052)** | –0.205 | –0.220 | –0.206 |
| **Accuracy** | – | 0.65 [0.572, 0.680] | 0.625 [0.566, 0.677] | 0.625 [0.569, 0.677] | 0.625 [0.569, 0.677] |
| **PPV** | – | 0.635 [0.522, 0.732] | 0.634 [0.518, 0.737] | 0.635 [0.520, 0.742] | 0.635 [0.519, 0.735] |
| **NPV** | – | 0.631 [0.521, 0.738] | 0.630 [0.517, 0.732] | 0.631 [0.521, 0.740] | 0.631 [0.520, 0.737] |

*Note*. OLS estimates from multiple regression were presented in the original paper and regularized estimates are presented alongside. Coefficients shown in bold were statistically supported in multiple regression at $\alpha = 0.05$. OLS = ordinary least squares; PPV = positive predictive value; NPV = negative predictive value.

higher predictive ability than standard Lande–Arnold regression for these data, but none of the differences were statistically significant.

## Discussion

Dealing with multicollinearity in selection analyses has been a challenge since Lande and Arnold (1983) proposed the estimation of selection on multivariate trait combinations using multiple regression. Regularized regression is an approach that has become particularly popular in recent years with the rise in machine learning and "big data" (e.g., Acharjee et al., 2013; Okser et al., 2014). Our simulations show that, for some measures of accuracy, using regularized regression in multivariate studies of selection can provide a substantial benefit compared to standard multiple regression. For other measures of accuracy, regularized regression may perform more poorly than least-squares analyses. Below, we evaluate general trends that come from our analyses and the usefulness of regularized regression for estimating the strength and direction of selection on multivariate trait combinations.

Overall, our simulations of known selection gradients show that when multicollinearity is low there is no detectable cost of using regularized regression when accuracy is assessed as the total strength of selection, MSE of the multivariate selection gradient, or in the overall direction of selection. When multicollinearity is high, all the regularized models perform better than OLS multiple regression for these measures of accuracy. However, which regularized model performs best depends on the measure of accuracy, whether fitness has a binomial (viability-like) or Poisson (fecundity-like) distribution, and whether some traits included in the analysis are not under selection. Improvements in the accuracy of the total strength of selection and MSE can be rather large, but improvements in the estimation of the direction of selection are always modest.

We observe the largest benefit of regularization when estimating the total strength of selection acting on a suite of traits (Figure 3). In the presence of any estimation error, the total strength of selection (calculated by the length of the estimated vector) is biased upward, which we observe at low levels of multicollinearity (Figure 3). This is the bias highlighted by Hereford et al. (2004) for univariate estimates of the strength of selection, determined by absolute values

**Table 3.** Estimated sexual selection gradients on male cuticular hydrocarbons of *Protopiophila litigata* from Angell et al. (2020).

| Regression method | | PC (9PCs) | OLS | Lasso | Ridge | Elastic net | PC (17 PCs) |
|---|---|---|---|---|---|---|---|
| | **Coefficients (+/– SE)** | | | | | | |
| – | FID peak 1 | 0.195 | 0.023 (0.621) | – | 0.086 | – | 0.491 |
| – | FID peak 2 | **0.396** | −0.384 (0.835) | −0.294 | −0.287 | −0.475 | −0.080 |
| – | FID peak 3 | **0.438** | −0.947 (0.904) | −0.851 | −0.817 | −1.228 | 1.646 |
| – | FID peak 4 | −**0.086** | 0.440 (0.581) | 0.256 | 0.291 | – | −0.140 |
| – | FID peak 5 | 0.520 | 0.372 (1.133) | – | 0.142 | – | 0.472 |
| – | FID peak 6 | **1.446** | 0.640 (0.860) | 0.303 | 0.454 | – | 0.783 |
| – | FID peak 7 | −0.291 | −0.241 (0.552) | −0.176 | −0.162 | – | −0.403 |
| – | FID peak 8 | −**0.266** | 0.612 (1.296) | – | 0.023 | – | −1.164 |
| – | FID peak 9 | −**0.332** | −0.537 (0.758) | −0.580 | −0.182 | – | −0.035 |
| – | FID peak 10 | −0.066 | −0.232 (0.976) | – | 0.025 | – | −0.272 |
| – | FID peak 11 | 0.003 | −0.124 (1.071) | – | −0.104 | – | −0.014 |
| – | FID peak 12 | −0.232 | −0.353 (0.906) | – | −0.182 | – | −0.340 |
| – | FID peak 13 | −**0.807** | −1.080 (0.801) | −0.823 | −0.547 | – | −0.558 |
| – | FID peak 14 | −**0.659** | 0.526 (0.764) | – | 0.202 | – | −1.528 |
| – | FID peak 15 | **0.170** | **1.799** (0.650) | 1.312 | 1.263 | 1.181 | 0.311 |
| – | FID peak 16 | −**0.370** | −0.345 (0.749) | – | 0.029 | – | 0.452 |
| – | FID peak 17 | −0.429 | 0.499 (1.091) | – | 0.191 | – | −0.051 |
| – | FID peak 18 | 0.371 | NA | – | 0.067 | – | 0.431 |
| **Accuracy** | – | – | 0.616 [0.474, 0.763] | 0.609 [0.447, 0.763] | 0.616 [0.474, 0.763] | 0.610 [0.447, 0.763] | – |
| **PPV** | – | – | 0.613 [0.400, 0.824] | 0.604 [0.400, 0.813] | 0.610 [0.417, 0.818] | 0.606 [0.400, 0.800] | – |
| **NPV** | – | – | 0.621 [0.421, 0.810] | 0.616 [0.412, 0.813] | 0.623 [0.412, 0.824] | 0.617 [0.412, 0.810] | – |

*Note*. PC estimates were presented in the original paper and we show the regularized and OLS estimates alongside. The selection gradients shown are for the data collected in 2013. Coefficients shown in bold were statistically supported in multiple regression at $\alpha = 0.05$. OLS = ordinary least squares; PPV = positive predictive value; NPV = negative predictive value.

**Table 4.** Estimated selection on *Arabidopsis* phenology from Chong et al., (2018).

| Regression method | | OLS regression | PC regression (4 PCs) | Lasso | Ridge | Elastic net |
|---|---|---|---|---|---|---|
| | **Coefficients (+/– SE)** | | | | | |
| – | Flowering time | −0.299 (0.191) | −0.181 | −0.371 | −0.143 | −0.249 |
| – | Flowering duration | 0.050 (0.182) | 0.186 | – | 0.109 | 0.048 |
| – | Branch number | 0.080 (0.085) | 0.034 | – | 0.083 | 0.067 |
| – | Rosette diameter | 0.059 (0.068) | 0.072 | – | 0.072 | 0.070 |
| – | Rosette leaf number | 0.061 (0.097) | 0.033 | – | −0.006 | – |
| **R$^2$** | – | 0.55 [0.007, 0.939] | – | 0.680 [0.081, 0.997] | 0.964 [0.838, 0.997] | 0.696 [0.059, 0.996] |

*Note*. PC estimates were presented in the original paper and we show the regularized estimates alongside. OLS = ordinary least squares.

of selection gradients. They showed that the bias is large when the standard errors are as large as the estimated coefficients, a common situation in selection gradient analysis, but that the bias rapidly decreases as relative error decreases. This upward bias is on the order of $\sqrt{1+\left(\frac{s}{b}\right)^2}$ where $s$ is the standard error and $b$ is the absolute value of the regression coefficient (Morrissey, 2014). When multicollinearity is high and the sample size is low, standard errors of the estimates are often much larger than the estimates themselves, and therefore regularization helps reduce this bias. We found that the total strength of selection is estimated with more error, and consequently, a larger upward bias, when fitness is binomially distributed than Poisson distributed, for a given sample size and multicollinearity. This is a consequence of the higher residual variance in binomial than Poisson models. The stronger shrinkage of coefficients by ridge regression compared to lasso or elastic net better attenuates the larger upward bias. However, the weaker shrinkage of coefficients imposed by lasso and elastic net is more helpful at low multicollinearity when fitness has a Poisson distribution.

In some scenarios, it may be enough for researchers to simply determine whether a trait is or is not under selection, and the direction of selection acting on it. This is the rationale for our final measure of accuracy, the proportion of estimates that are in the correct direction. In this context, least-squares regression always performs as well as regularized regression, showing the benefits of a lack of bias. As expected, ridge regression performs just as well as least squares, since it never alters the sign of estimates relative to least squares. Lasso was notably inferior to the other estimates, and the elastic net shares some of this disadvantage.

An unexpected result is that both the lasso and elastic net regularization perform poorly when some of the traits are not under selection. One of the advertised advantages of these methods is that they allow shrinkage of small estimates to 0, which would seem to give them an advantage in detecting the absence of selection. For suites of traits where only some are under selection, lasso and elastic net correctly identify the traits under selection about 55% of the time, on average (Figure 6A). This result underscores that the traits identified as/as not being under selection will be those that in combination with other traits in the model yield better predictions of fitness, not those that are more likely to be/not be under selection. Selection is rarely estimated in the correct direction for any of the regression methods when some of the traits included in the model are not under selection, for the same reason (Figure 6B). Our results suggest that unless there is an a priori reason to predict that correlated traits are also under selection, including them in a selection analysis will lead to less accurate estimates of selection on focal traits, than not including them.

Regularization introduces bias into estimated regression coefficients by placing constraints on the magnitudes of estimated coefficients, which also reduces their variance. Ultimately, the question of interest and structure of data will determine whether and which type of regression is likely to be the most useful. We show that the decrease in the variance of regularized estimates can outweigh the increase in bias that regularization introduces, leading to more accurate estimates of summaries of multivariate selection by some of our accuracy measures. However, for other measures of accuracy, the benefit of decreased variance may not outweigh the increase in bias, presenting an argument for OLS multiple regression in those cases. When the estimate of selection on a particular trait is the parameter of interest, applying regularization to the entire estimated selection gradient may not produce a more accurate estimate of selection for any individual trait. It depends on whether the decrease in variance for that trait is outweighed by the increase in bias. OLS, however, will always provide unbiased estimates of selection on individual traits.

Other traditional solutions to the problems of multicollinearity and limited data are to perform subset selection or use principal component or other data reduction techniques (Lande & Arnold, 1983; Mitchell-Olds & Shaw, 1987). However, these approaches can have undesirable properties, such as producing estimates of selection in the wrong direction as demonstrated in Figure 1. Choosing how to reduce a data set can invite arbitrary decisions regarding the number of important principal components to include. A frequent, but underdocumented analog of this data reduction problem is in the initial choice of variables to measure, where researchers may choose to exclude highly correlated traits a priori, in the interests of reducing the variance inflation that the study of correlated traits brings. Our results demonstrate that, in some cases, we can study aspects of selection on highly correlated traits in a meaningful way using regularized regression. A previous study that explored regularized regression for estimating selection in Soay sheep suggested that it did not improve our understanding of selection in this system, but highlighted the possible utility in other situations (Morrissey, 2014). Our more comprehensive simulations show that regularization can improve the accuracy of the total strength and direction of estimated multivariate selection gradients under realistic scenarios.

Our reanalysis of the three published data sets provided mixed results for the benefits of regularization. We obtained different estimates of selection than Angell et al. (2020) who estimated sexual selection via binomial mating success on pheromone traits in *Protopiophila litigata*. Using principal component regression, they found statistically supported selection on 10 of 18 traits. Our reanalysis using all the regularization techniques, as well as least squares, estimated selection on three of those 10 traits in the opposite direction. We also estimated the trait under the strongest selection ($\beta = 1.446$) in the original analysis to be under almost five times weaker selection ($\beta = 0.3$). The reconstituted selection gradients in the original analysis represent selection on only 50% of the multivariate trait space (selection was estimated on 9 of 18 PCs), while the regularized estimates encompass the entire space. In principle this could be one explanation for the difference in the estimated selection gradients, however, that does not seem to be the case here. The reconstituted selection gradients from an analysis of 17 of 18 PCs are not more similar to the regularized estimates (Table 3). Despite producing different estimates of selection, all the models did a similar job at predicting fitness in an independent data set (Table 3), demonstrating that all estimated multivariate selection gradients, including those in the original paper, provide important biological information about selection. However, they do not provide the same answers with regard to the strength and direction of selection on individual traits, highlighting that the question at hand will determine which method will perform best.

For the other two case studies we reanalysed, regularized regression provided similar results to the original analyses. The estimated sexual selection gradient in Sztepanacz and Rundle (2012) was very close to that produced by regularized regression, showing no benefit (or cost) to using the regularized approach when data are abundant, and multicollinearity is low. The regularized reanalysis of Chong and Stinchcombe (2018) recapitulated the results they found from performing principal component regression and back-transforming the coefficients to the original traits.

There are similarities between regularized regression and principal component regression, however, regularized regression allows more flexibility. Although PC regression allows researchers to set some multivariate trait combinations to zero, unless the loading of a trait on all principal components is zero, it will be estimated to have a nonzero coefficient. Therefore, PC regression shares more similarity to ridge regression than the lasso. One particular benefit of regularization compared to PC regression is that it is possible to omit some traits from the regularization and not others by applying separate penalty factors to each trait. If the penalty factor for a particular trait is set to 0, no shrinkage is applied to the estimated coefficient. This option is straightforward to implement in the glmnet R package and is very useful if a researcher wants unbiased estimates of selection for some traits, or is particularly interested in an estimate of the strength of selection acting on one focal trait and fitting others as covariates. Regularized regression also determines optimal shrinkage based on the data, by using CV, obviating the need for arbitrary decisions by the analyst. In traditional PC regression, the researcher chooses how many PCs to regress on relative fitness, omitting those that explain the least variation which also contribute the most to multicollinearity (Jolliffe, 2002), often based on some arbitrary cutoff of variance explained or visual break in a scree plot. This has been criticized on the grounds that these PCs can be important in predicting relative fitness (Chong et al, 2018; Jolliffe, 2002; Mitchell-Olds & Shaw, 1987). There are at least two alternative ways of choosing which PCs to include that could mitigate this problem. The first is the employment of cross-validation in the choice of variables. The second is to determine whether the variance explained by a given principal component is larger than the sampling error using the Tracy–Widom distribution (Johnstone, 2008; Saccenti et al., 2011; Sztepanacz & Blows, 2017). Neither approach has, to our knowledge, been widely applied in selection studies.

Despite the advantages we highlight, regularized regression has drawbacks. The estimates it produces are biased, unlike least-squares estimators. Although we demonstrate the bias–variance trade-off can work in favor of regularized regression in the range of parameters and for the summaries of multivariate selection that we studied, the introduced bias reduces the utility of regularized selection gradients in meta-analyses (Morrissey & Hadfield, 2012; Siepielski et al., 2013). We, therefore, suggest that regularized gradients be presented alongside OLS multiple regression estimates and their standard errors. In the frequentist framework we employed, regularized regression does not directly yield standard errors of estimated coefficients, standard error, or $p$-values. Implementing such error estimation by bootstrapping or cross-validation, or by using Bayesian approaches is possible but more time-consuming. Some researchers may view the lack of $p$-values and standard errors as a drawback of using regularized regression in selection analyses. We do not share this opinion, given the rampant misunderstanding of the meaning of statistical significance, and that biological and statistical significance are often unrelated (Motulsky, 2015; Wasserstein et al., 2019). The magnitudes of the coefficients will indicate their relative importance. However, when estimates of individual selection coefficients and their uncertainty are the parameters of most interest, OLS regression or Bayesian approaches to regularization that allow for uncertainty of parameter estimates to be retained through the posterior distribution, will likely be a more favorable approach. Regularized regression may also have more limited utility when the goal is to obtain estimates of mean-standardized selection gradients (Hereford et al., 2004) on traits that are measured on different scales. Traits measured on different scales will have different magnitudes of the regularization penalty applied to them, simply due to differences in the scale of the data (Hastie et al, 2009).

Overall, our analyses show that regularized regression is a straightforward and easily implementable approach that can provide more accurate estimates of the total strength and direction of multivariate selection than traditional approaches. Our reanalysis of the three case studies showed that it can help the interpretation of selection when multicollinearity is high and data are limited, and that it produces similar selection gradients to multiple regression in the remaining cases. Ultimately, the question and data structure at hand will determine which regression approach will be best for estimating selection in each case. Regularized regression is a promising method for future studies of multivariate selection and will become particularly important as phenomic technologies increase and the number of traits that we can estimate selection on in a single study.

## Supplementary material

Supplementary material is available online at *Evolution Letters*.

## Data and code availability

All simulated data used in the present article's analyses are generated using scripts presented in Supporting Information. Empirical data that were used for reanalysis is publicly available from the original publications.

## Author contributions

J.L.S. and D.H. conceived the study. J.L.S. did the simulations, analyses, and wrote the manuscript. J.L.S. and D.H. edited the manuscript.

## Funding

This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant RGPIN-2020-05838 to J.L.S and the National Science Foundation (NSF) grant DEB-1556774 to D.H.

*Conflict of interest:* The authors declare no conflict of interest.

## Acknowledgments

We thank Thomas F. Hansen, John Stinchcombe, Aneil Agrawal, Stephen Wright, and Asher Cutter for the helpful discussion. We also thank Timothée Bonnet and Michael Morrissey for their helpful comments on the manuscript.

## References

Acharjee, A., Finkers, R., Visser, R. G., & Maliepaard, C. (2013). Comparison of regularized regression methods for ~omics data. *Metabolomics*, 3, 1.

Angell, C. S., Curtis, S., Ryckenbusch, A., & Rundle, H. D. (2020). Epicuticular compounds of *Protopiophila litigata* (Diptera: Piophilidae): Identification and sexual selection across two years in the wild. *Annals of the Entomological Society of America*, *113*(1), 40–49. https://doi.org/10.1093/aesa/saz056.

Bartkowska, M. P., & Johnston, M. O. (2015). Pollen limitation and its influence on natural selection through seed set. *Journal of Evolutionary Biology*, 28(11), 2097–2105. https://doi.org/10.1111/jeb.12741

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Blows, M. W., & Allan, R. A. (1998). Levels of mate recognition within and between two Drosophila species and their hybrids. *The American Naturalist*, 152(6), 826–837. https://doi.org/10.1086/286211

Brachi, B., Aimé, C., Glorieux, C., Cuguen, J., & Roux, F. (2012). Adaptive value of phenological traits in stressful environments: Predictions based on seed production and laboratory natural selection. *PLoS One*, *7*(3), e32069–e32015. https://doi.org/10.1371/journal.pone.0032069

Celeux, G., El Anbari, M., Marin, J. M., & Robert, C. P. (2012). Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, *7*(2), 477–502.

Chapurlat, E., Ågren, J., Anderson, J., Friberg, M., & Sletvold, N. (2019). Conflicting selection on floral scent emission in the orchid *Gymnadenia conopsea*. *New Phytologist*, 222(4), 2009–2022. https://doi.org/10.1111/nph.15747. Wiley Online Library

Chong, V. K., Fung, H. F., & Stinchcombe, J. R. (2018). A note on measuring natural selection on principal component scores. *Evolution Letters*, 2(4), 272–280. https://doi.org/10.1002/evl3.63

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*(376), 817–823. https://doi.org/10.1080/01621459.1981.10477729.

Friedman J, Tibshirani R, Hastie T (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22. https://doi.org/10.18637/jss.v033.i01

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, *194*(3), 573–596. https://doi.org/10.1534/genetics.113.151753

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. (Vol. 2). Springer.

Hereford, J., Hansen, T. F., & Houle, D. (2004). Comparing strengths of directional selection: How strong is strong? *Evolution*, 58(10), 2133–2143. https://doi.org/10.1111/j.0014-3820.2004.tb01592.x

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*(1), 80–86. https://doi.org/10.1080/00401706.2000.10485983.

Hoikkala, A., Aspi, J., & Suvanto, L. (1998). Male courtship song frequency as an indicator of male genetic quality in an insect species, *Drosophila montana*. *Proceedings Biological Sciences*, 265(1395), 503–508. https://doi.org/10.1098/rspb.1998.0323

Houle, D. (2010). Numbering the hairs on our heads: The shared challenge and promise of phenomics. *Proceedings of the National Academy of Sciences*, *107*(Suppl 1), 1793–1799. https://doi.org/10.1073/pnas.0906195106

Hoy, R. R., Hoikkala, A., & Kaneshiro, K. (1988). Hawaiian courtship songs: Evolutionary innovation in communication signals of Drosophila. *Science*, 240(4849), 217–219. https://doi.org/10.1126/science.3127882.

Johnstone, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Annals of statistics*, 36(6), 2638–2716. https://doi.org/10.1214/08-AOS605

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Kingsolver, J. G., Hoekstra, H. E., Hoekstra, J. M., Berrigan, D., Vignieri, S. N., Hill, C. E., Hoang, A., Gibert, P., & Beerli, P. (2001). The strength of phenotypic selection in natural populations. *The American Naturalist*, 157(3), 245–261. https://doi.org/10.1086/319193

Lande, R., & Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution*, *37*(6), 1210–1226. https://doi.org/10.1111/j.1558-5646.1983.tb00236.x

Lindholm, A. K., Head, M. L., Brooks, R. C., Rollins, L. A., Ingleby, F. C., & Zajitschek, S. R. K. (2014). Causes of male sexual trait divergence in introduced populations of guppies. *Journal of Evolutionary Biology*, 27(2), 437–448. https://doi.org/10.1111/jeb.12313

Melo, D., Marroig, G., & Wolf, J. B. (2019). Genomic perspective on multivariate variation, pleiotropy, and evolution. *The Journal of Heredity*, 110(4), 479–493. https://doi.org/10.1093/jhered/esz011.

Mitchell-Olds, T., & Shaw, R. G. (1987). Regression analysis of natural selection: Statistical inference and biological interpretation. *Evolution*, *41*(6), 1149–1161. https://doi.org/10.1111/j.1558-5646.1987.tb02457.x.

Morrissey, M. B. (2014). In search of the best methods for multivariate selection analysis. *Methods in Ecology and Evolution*, 5(10), 1095–1109. https://doi.org/10.1111/2041-210x.12259

Morrissey, M. B. (2016). Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology*, 29(10), 1882–1904. https://doi.org/10.1111/jeb.12950

Morrissey, M. B., & Hadfield, J. D. (2012). Directional selection in temporally replicated studies is remarkably consistent. *Evolution*, 66(2), 435–442. https://doi.org/10.1111/j.1558-5646.2011.01444.x.

Morrissey, M. B., & Ruxton, G. D. (2018). Multiple regression is not multiple regressions: The meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology*, 10(20220112), 2–24. https://doi.org/10.3998/ptpbio.16039257.0010.003

Motulsky, H. J. (2015). Common misconceptions about data analysis and statistics. *Pharmacology Research & Perspectives*, 3(1), e00093. https://doi.org/10.1002/prp2.93.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6.

Oh, K. P., & Shaw, K. L. (2013). Multivariate sexual selection in a rapidly evolving speciation phenotype. *Proceedings Biological Sciences*, *280*(1761), 20130482. https://doi.org/10.1098/rspb.2013.0482.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., & Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. *PLoS Genetics*, *10*(11), e1004754. https://doi.org/10.1371/journal.pgen.1004754.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Poissant, J., Morrissey, M. B., Gosler, A. G., Slate, J., & Sheldon, B. C. (2016). Multivariate selection and intersexual genetic constraints in a wild bird population. *Journal of Evolutionary Biology*, 29(10), 2022–2035. https://doi.org/10.1111/jeb.12925

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Saccenti, E., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. W. B. (2011). Tracy-Widom statistic for the largest eigenvalue of autoscaled real matrices. *Journal of Chemometrics*, 25, 644–652.

Sanjak, J. S., Sidorenko, J., Robinson, M. R., Thornton, K. R., & Visscher, P. M. (2018). Evidence of directional and stabilizing selection in contemporary humans. *Proceedings of the National Academy of Sciences*, 115(1), E4732.

Schluter, D., & Nychka, D. (1994). Exploring fitness surfaces. *The American Naturalist*, *143*(4), 597–616. https://doi.org/10.1086/285622. University of Chicago Press

Siepielski, A. M., Gotanda, K. M., Morrissey, M. B., Diamond, S. E., DiBattista, J. D., & Carlson, S. M. (2013). The spatial patterns of directional phenotypic selection. *Ecology Letters*, *16*(11), 1382–1392. https://doi.org/10.1111/ele.12174.

Sztepanacz, J. L., & Blows, M. W. (2017). Accounting for sampling error in genetic eigenvalues using random matrix theory. *Genetics*, *206*(3), 1271–1284. https://doi.org/10.1534/genetics.116.198606

Sztepanacz, J. L., & Rundle, H. D. (2012). Reduced genetic variance among high fitness individuals: Inferring stabilizing selection on male sexual displays in Drosophila serrata. *Evolution*, 66(10), 3101–3110. https://doi.org/10.1111/j.1558-5646.2012.01658.x.

Talyn, B. C., & Dowse, H. B. (2004). The role of courtship song in sexual selection and species recognition by female *Drosophila melanogaster*. *Animal Behaviour*, 68(5), 1165–1180. https://doi.org/10.1016/j.anbehav.2003.11.023

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Walker, L. K., Ewen, J. G., Brekke, P., & Kilner, R. M. (2014). Sexually selected dichromatism in the hihi *Notiomystis cincta*: Multiple colours for multiple receivers. *Journal of Evolutionary Biology*, 27(8), 1522–1535. https://doi.org/10.1111/jeb.12417

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "*p* < 005". *American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.