Estimating (non)linear selection on reaction norms: A general framework for labile traits

Jordan S. Martin[*1], Yimen Araya-Ajoy[2], Niels J. Dingemanse[3], Alastair J. Wilson[4], & David Westneat[5]

*corresponding author: jordan.martin@uzh.ch

[1]*Human Ecology Group, Institute of Evolutionary Medicine, University of Zurich Switzerland*

[2]*Center for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, Norway*

[3]*Behavioral Ecology Unit, Department of Biology, Ludwig Maximilian University of Munich, Germany*

[4]*Evolution Group, Centre for Biosciences, University of Exeter, United Kingdom*

[5]*Department of Biology, University of Kentucky, United States of America*

Abstract

Individual reaction norms describe how labile phenotypes vary as a function of organisms' expected trait values (intercepts) and plasticity across environments (slopes), as well as their degree of stochastic phenotypic variability or predictability (residuals). These reaction norms can be estimated empirically using multilevel, mixed-effects models and play a key role in ecological research on a variety of behavioral, physiological, and morphological traits. Many evolutionary models have also emphasized the importance of understanding reaction norms as a target of selection in heterogeneous and dynamic environments. However, it remains difficult to empirically estimate nonlinear selection on reaction norms, inhibiting robust tests of adaptive theory and accurate predictions of phenotypic evolution. To address this challenge, we propose generalized multilevel models for estimating stabilizing, disruptive, and correlational selection on the reaction norms of labile traits, which can be applied to any repeatedly measured phenotype using a flexible Bayesian framework. These models avoid inferential bias by accounting for uncertainty in reaction norm parameters and their potentially nonlinear fitness effects. We validate these models in a Bayesian framework using multiple simulation techniques, demonstrating unbiased inference across a broad range of effect sizes and desirable power for large sample sizes. Coding tutorials are further provided to aid empiricists in applying these models to any phenotype of interest using the Stan statistical programming language in R.

Keywords

phenotypic evolution, complex trait, multivariate, adaptation, personality, flexibility

Introduction

A population will evolve by natural selection whenever heritable variation occurs in fitness-relevant phenotypes (Darwin 1859). Measuring the fitness consequences of individual differences in highly labile behavioral, physiological, and morphological traits is, therefore, fundamental for explaining their adaptive evolution. Across a variety of phenotypes and taxa, repeatable individual differences have been observed in organisms' average trait values (Bell, Hankison, & Laskowski 2009; Fanson & Biro 2015; Cauchoix et al. 2018) and in their plasticity across environments (Dingemanse et al. 2010; Stamps 2016; Arnold, Nicotra, & Kruuk 2019), with some individuals consistently being more or less responsive to environmental change than others. In addition, it is increasingly appreciated that individuals may repeatably differ in their degree of stochastic phenotypic variability within a given environment (see **Box 1** below for a conceptual overview; Biro & Adriaenssens 2013; Westneat, Schofield, & Wright 2013; Mitchell, Beckmann, & Biro 2021), a phenomenon which has often been ignored in ecological research (Hansen, Carter & Pélabon 2006). These individual-specific patterns reflect distinct but potentially integrated parameters (intercepts, slopes, and within-individual residuals) of the reaction norms (RNs, i.e. state-dependent functions relating phenotype to environment, **Table 1**) evolving in a population (**Figure 1**). RN models provide a highly generalizable, quantitative framework for investigating the evolution and development of labile traits, with broad applications ranging from social behaviors (Dingemanse & Araya-Ajoy 2015; McNamara & Leimar 2020; Martin, Jaeggi, & Koski 2023) and learning processes (Wright, Haaland, Dingemanse, & Westneat 2022) to thermal performance curves (Svensson, Gomez-Llano, & Waller 2020) and extended phenotypes (Munar-Delgado, Araya-Ajoy, & Edelaar, 2023), such as gall size in insect-host plant interactions (Weis & Gorman 1990). Interest in the evolutionary ecology of RNs has grown steadily across a diverse range of fields in recent decades (e.g. Brommer, Kontiainen, & Pietiäinen 2012; Strickland et al. 2021; Newediuk, Prokopenko, & Wal 2022), generating methodological innovations for estimating RNs subject to measurement error (e.g. Nussey, Wilson, & Brommer 2007; Dingemanse & Dochtermann 2013; Gomulkiewicz et al. 2018; O'Dea, Noble, & Nakagawa 2021; Martin & Jaeggi 2022), as well as theoretical models for explaining the

selection pressures shaping and maintaining individual variation in RNs within populations (e.g. Wolf & Weissing 2010; Dall & Griffith 2014; Sih et al. 2015; Wright et al. 2019). Attention to RNs has also increased in related fields of inquiry such as personality psychology (Denissen & Penke 2008; Nettle & Penke 2010) and evolutionary anthropology (Jaeggi et al. 2016).

RN models are, of course, not only useful for describing phenotypic variation. While classical models largely focused on the consequences of phenotypic selection for RN evolution (e.g. Gavrilets & Sheiner, 1993), many evolutionary frameworks also emphasize that the parameters of RNs (intercepts, slopes, and residuals) may be direct targets of selection, leading to differential patterns of adaptation and extinction in changing environments (Via et al. 1995; Schlichting & Piglucci 1998; Ghalambor, McKay, Carroll, & Reznick 2007; Fox et al. 2019). For instance, evolutionary ecologists have long investigated the role of both cue-induced and stochastic phenotypic plasticity in the colonization of novel habitats (Caño et al. 2008; Volis, Ormanbekova, & Yermekbayev 2015; Hendry 2016; Wang & Althoff 2019). In addition, evolutionary geneticists have shown how plasticity in social environments can magnify heritable variation in mean trait values, accelerating or inhibiting phenotypic evolution in comparison to unresponsive phenotypes (Moore et al. 1997; Bijma & Wade 2008; McGlothlin et al. 2010; Kazancıoğlu, Klug, & Alonzo 2012). Game theorists and behavioral ecologists have further emphasized the importance of understanding selection on RNs due to the prevalence of fluctuating density- and frequency-dependent selection in social environments (Araya-Ajoy, Westneat, & Wright 2020; McNamara & Leimar 2020; Martin, Jaeggi, & Koski 2023), as well as the role of dynamic environments more generally in selecting for learning mechanisms and emotional states rather than specific behaviors per se (Skinner, 1966; Henrich & McElreath 2003; McNamara & Houston 2009; Fawcett, Hamblin, & Giraldeau 2013; Nakahashi & Ohtsuki 2015; Wright et al. 2022). Distinct genetic control of phenotypic stability and change has also been experimentally demonstrated for diverse phenomena from cold tolerance (Ørsted, Rohde, Hoffmann, Sørensen, & Kristensen 2018) to body size (Scheiner & Lyman, 1991) and various forms of developmental polyphenism (Suzuki & Nijhout 2006; Projecto-Garcia, Biddle, Ragsdale 2017), suggesting that differential selection on heritable variation in RN intercepts, slopes, and

residuals, as well as differential patterns of genetic integration between RN parameters (Wagner, Booth, & Bagheri-Chaichian, 1997; Tonsor, Elnaccash, & Scheiner, 2013), can in turn have distinct consequences for phenotypic evolution. Accordingly, divergence has been observed in the RNs of many naturally occurring populations, such as differential plasticity in the growth rates of phytoplankton (*Thalassiosira pseudonana*; Schaum, Buckling, Smirnoff, & Yvon-Durocher 2022), ponderosa pine (*Pinus ponderosa*; de la Mata et al. 2022) and single-leaf pinyon (*Pinus monophylla*; Vasey, Weisberg, & Urza 2022) populations in response to temperature fluctuations and microhabitat heterogeneity. Despite this strong theoretical emphasis and empirical basis, robust statistical methods have not yet been developed for detecting complex patterns of selection on the RNs of labile traits.

Many of the phenotypes commonly studied by evolutionary ecologists are highly labile (i.e. exhibit high degrees of reversible plasticity; Scheiner, 1993) in response to the local environment. This means that repeatable individual differences in the RN underlying these traits tend to account for only a modest proportion of the total variation observed in measurements across space and time (Bell, Hankison, & Laskowski 2009; Fanson & Biro 2015; Cauchoix et al. 2018). This is expected, given that labile traits are often adapted to facilitate flexible responses toward fitness-relevant variation in the environment (Scheiner 1993), such as by up-regulating circulating testosterone in response to social challenges (Wingfield et al. 1990; Eisenegger, Haushofer, & Fehr 2011), temporarily inducing a fear state in response to odor cues of predation (Mathuru et al. 2012), or regulating alloparental care in response to the quality of the local environment (Guindre-Parker & Rubenstein, 2018; Martin et al. 2020). Conversely, labile traits may also be prone to maladaptive plasticity in response to novel or extreme environmental stressors (e.g. Ghalambor et al. 2015). As such, any particular measurement of a labile phenotype will tend to reflect within- rather than among-individual variation, potentially biasing empirical estimates of trait (co)variances and selection gradients estimated across heterogeneous environments (Brommer 2013; Dingemanse & Dochtermann 2013; Niemelä & Dingemanse 2018; Royauté et al. 2018), leading to inaccurate inferences about adaptive evolution (Dingemanse, Araya-Ajoy, & Westneat 2021; Martin & Jaeggi 2022). Classical approaches such as the Lande and Arnold (1983) regression framework do not partition

repeatable and non-repeatable differences across phenotypic measurements and, as a consequence, may lead to downwardly biased estimates of selection gradients for labile traits in field research (Dingemanse et al. 2021). Classical methods can also be biased by unmeasured, within-individual environmental effects on fitness and phenotype that generate spurious signals of selection (Scheiner et al. 2002; Stinchcombe et al. 2002). Using these methods to estimate selection on labile traits with single measures, averages of raw data, or point estimates in multi-stage analyses can, therefore, increase the risk of biased evolutionary inference (Hadfield et al. 2010), particularly when attempting to understand the adaptation of RNs underlying observed phenotypes across environments.

Fortunately, generalized linear mixed-effects models (GLMMs) provide a flexible toolkit for estimating RNs from empirical data, as well as for modelling the effects of RNs on fitness and other biological outcomes of interest. Current variance-partitioning methods rely on the use of multi-response/multivariate GLMMs with covarying random effects to model selection, which effectively account for uncertainty in individuals' RNs and their estimated effects (Hadfield et al. 2010). This approach has been repeatedly introduced to selection studies of RNs in variety of contexts, demonstrating its broad applicability (e.g. Brommer, Kontiainen, & Pietiäinen 2012; Houslay & Wilson 2017; Arnold, Nicotra, & Kruuk 2019), and can be further extended to provide a veritable treasure chest of biological insights (Blows 2007). For example, such models can be used to identify trajectories of phenotypic conservation and divergence among closely related populations (Royauté, Hedrick, & Dochtermann 2020), discover latent behavioral characters among multiple traits (Araya-Ajoy & Dingemanse 2014; Martin et al. 2019), or calculate genetic responses to directional selection (Stinchcombe, Simonsen, & Blows 2014). Therefore, multi-response GLMMs with covarying random effects can be used to accomplish many empirical goals with relative ease, while also avoiding statistical bias due to uncertainty in RNs.

Despite their benefits, these commonly used GLMMs cannot detect nonlinear selection on RNs (i.e disruptive, stabilizing, and correlational selection) because the random effect covariance is defined as an average measure of linear dependency among fitness and phenotype. By failing to describe the curvature of the adaptive landscape, and

thus the ecological phenomena generating fitness saddles, ridges, domes, and cliffs (Lande & Arnold, 1983; Blows & Brooks, 2003; Blows 2007; Vercken et al., 2012), random effect models can provide an incomplete and potentially misleading perspective on the biological processes driving and constraining multivariate evolution. In non-randomized experiments or field settings, ignoring nonlinear selection can further generate biased estimates of directional selection gradients, in addition to biased predictions of the evolutionary response to selection on the expectations and (co)variances of RN parameters (Arnold et al., 2001; Morrissey et al., 2012; Pick et al., 2022). Therefore, despite their clear utility, current covarying random effects models can also limit robust tests of adaptive theory, which often predicts that stabilizing, disruptive, and/or correlational selection will shape RN evolution (e.g. Wagner et al., 1997; Gavrilets & Hastings, 1994). This inhibits accurate predictions of phenotypic evolution more generally (Bulmer 1971; Lande & Arnold 1983; Arnold, Pfrender, & Jones, 2001; Villemereuil et al., 2020).

Here we address this challenge by introducing multi-response/multivariate GLMMs for unbiased estimation of nonlinear selection on RNs, building on well-established approaches to estimating linear selection (e.g. Brommer, Kontiainen, & Pietiäinen 2012; Houslay & Wilson 2017; Arnold, Nicotra, & Kruuk 2019; Araya-Ajoy, Dingemanse, Westneat, & Wright 2023). The proposed GLMMs are applicable to any labile and repeatedly measured phenotype. We begin by reviewing so-called double hierarchical GLMMs for estimating RNs from longitudinal, repeated measures data (Westneat, Schofield, & Wright, 2013; O'Dea et al. 2021) and formally introduce multi-response/multivariate models estimating linear and nonlinear selection on RNs, applicable to both Gaussian and non-Gaussian measurements. We then consider their implementation in a Bayesian framework, using a simulation-based calibration procedure to validate that the proposed models are unbiased for statistical inference. We also explore statistical power for Bayesian hypothesis tests across a range of sampling designs and selection effect sizes. Guided tutorials are further provided (see **data**

194 **availability**) to aid researchers in implementing and interpreting these models for their
195 own data using the Stan statistical programming language (Carpenter et al. 2017).

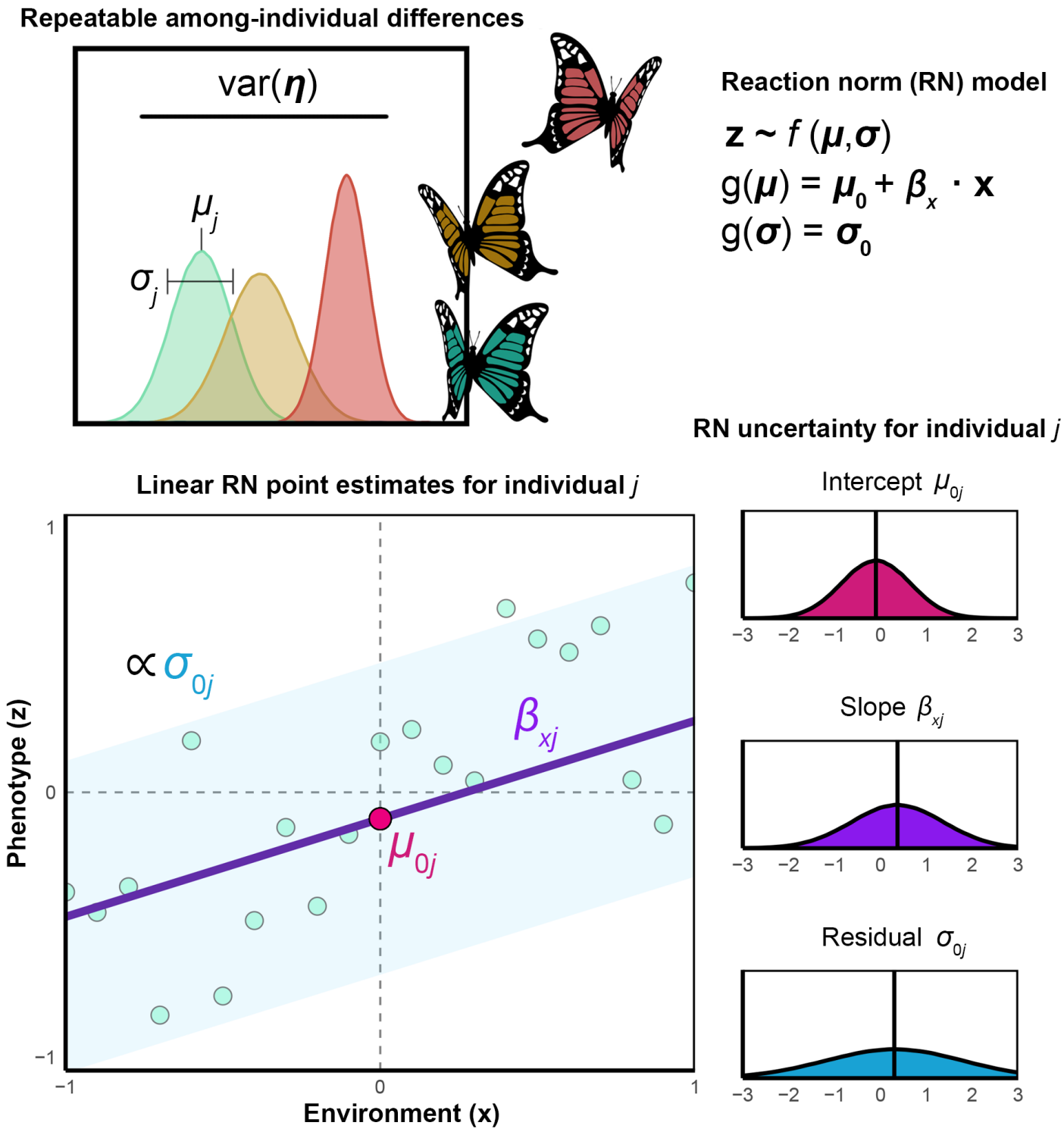

Figure 1. *Empirical estimation of reaction norms*. Repeatable among-individual differences $\mathrm{var}(\boldsymbol{\eta})$ (*top left*) in the expected value $\boldsymbol{\mu}$ and dispersion $\boldsymbol{\sigma}$ of observed phenotype **z** can be predicted with a RN model (*top right*) using link functions ***g*** and three (or more) distinct parameters: RN intercept parameters $\boldsymbol{\mu_0}$ describing each individual's average phenotype across a mean-centered environment or in the absence of the environment (i.e. when the environmental state x = 0); RN slope parameters $\boldsymbol{\beta_x}$ describing each individual's systematic change in phenotype across environmental states **x**; and RN residual parameters $\boldsymbol{\sigma_0}$ reflecting each individual's degree of stochastic variability (or, conversely, their predictability/precision) in phenotype within a given environment. See **Eq. 1** for index rather than matrix notation. These parameters will be unknown in empirical research and must, therefore, be estimated using raw longitudinal measurements (teal circles) across environmental states (*bottom left*). An example is shown for a simple linear RN with a log-link on the dispersion of a normal distribution, so that an individual's residual parameter, expressed as a variance on the squared log scale $\mathrm{sqrt}\left(\exp\left(\sigma_0 + \sigma_{0j}\right)\right)$, is proportional to ($\propto$) the spread of observed residuals on the original data scale, as shown here by a 95% credible interval. Failure to account for uncertainty around point estimates of individual *j*'s RN parameters (*bottom right*) leads to anti-conservative inference and increased risk of false positives (Hadfield et al. 2010).

214 **Table 1.** Notation and terminology.

Term	Symbol	Description
Individual reaction norm (RN)	$f(\mu, \sigma)$	A probabilistic function f with parameters predicting the expectation μ and dispersion σ of an individual's phenotype in response to a measurable aspect of the environment.
RN intercept	μ_0, μ_{0j}	The expected phenotype in the average environment or in the absence of an environmental factor. Individual RN intercept μ_{0j} is expressed as a deviation from population RN intercept μ_0.
RN slope	β_x, β_{x_j}	The expected change in phenotype in response to a measured environment *x*. Individual RN slope β_{x_j} is expressed as a deviation from the population average slope β_x.
RN residual	σ_0, σ_{0j}	The magnitude of stochastic variability in phenotype within a given environment, i.e. the inverse of predictability (O'Dea et al., 2021) and precision (Hansen et al., 2006). Individual RN residual parameter σ_{0j} is expressed as a deviation from population average residual parameter σ_0, which together determine the magnitude of variation in observed residuals.
RN trait value/ character state	η_{jt}	The repeatable trait value predicted by individual *j*'s reaction norm being expressed within the environmental state at time *t.* This context-specific trait value is also referred to as a character state in quantitative genetics.
Repeatable among-individual differences	$\mathrm{var}(\boldsymbol{\eta})$	The total amount of among-individual variation in the phenotype available to natural selection over the sampling period, which reflects consistent individual differences in RN expression across environments (i.e. the variance of character states).
Link functions	$g_\mu, g_\sigma, g_\theta$	Transformations that facilitate modelling of non-Gaussian phenotypes and fitness measures on a linear scale.
Fitness	W, $f(\theta, \delta)$	A measure of an individual's observed survival, reproduction, and/or performance *W,* as predicted by the expectation θ and dispersion δ parameters of distribution *f.* These quantifiable 'fitness components' are used to approximate the repeatable, differential rate of zygote production across individuals.
Directional selection	$\mathbf{b}, \boldsymbol{\beta}$	Selection gradients $\boldsymbol{\beta}$ quantify the magnitude of direct selection on the population means of reaction norm parameters. Regression coefficients **b** approximate these effects on the transformed scale of a GLMM.
Quadratic selection	$\mathbf{q}, \boldsymbol{\gamma}$	Selection gradients $\boldsymbol{\gamma}$ quantify the magnitude of direct selection on the (co)variances of reaction norm parameters. Regression coefficients $\boldsymbol{q}$ approximate these effects on the transformed scale of a GLMM.
Fluctuating selection	$\boldsymbol{\Delta\beta}, \boldsymbol{\Delta\gamma}$	Environmental change that shifts the magnitude of selection on RNs $\boldsymbol{\Delta\beta}, \boldsymbol{\Delta\gamma}$ across space and/or time (see *supplementary appendix* for details)

Modelling nonlinear selection on labile traits

The models we propose in this section are straightforward extensions of the multi-response/multivariate random effects GLMMs discussed above. Our trait-based approach shifts estimation of fitness effects from random effect covariances to flexibly parameterized linear and nonlinear selection coefficients. This approach builds on a long tradition of measurement error models in biostatistics (Loken & Gelman 2017; Ponzi et al. 2018; Martin & Jaeggi 2022), also known as structural equation (Bollen & Noble 2011; Araya-Ajoy & Dingemanse 2014; Martin et al. 2019) or errors-in-variables models (Dingemanse et al. 2021), which allow for latent trait values such as RN intercept, slope, and residual parameters to simultaneously affect multiple response models. The basic structure of these models has been previously introduced in the broader context of phenotypic selection analysis by Ponzi et al. (2018), Dingemanse et al. (2021), and Araya-Ajoy et al. (2023), who considered Gaussian models of selection on repeatable trait values. Here, we generalize and extend these models to allow for estimating (non)linear selection on RN intercepts, slopes, and residuals (and any other distributional parameters of interest), as well as to estimate directional and quadratic selection gradients on RN parameters with non-Gaussian phenotype and fitness measures.

Reaction norm model

The first step in any selection analysis is to define the trait of interest. For repeatedly expressed traits that exhibit plasticity, the 'traits' of interest may be latent properties of a RN, which researchers can estimate as functional parameters. As shown in **Figure 1**, individual variation in a linear RN can be decomposed into underlying repeatable differences in individuals' RN intercept $\boldsymbol{\mu_0}$, slope $\boldsymbol{\beta_x}$, and residual parameters $\boldsymbol{\sigma_0}$. Note that we use $\boldsymbol{\beta_x}$ here to reference any slope defined over a non-social environmental state (see Martin & Jaeggi, 2022 for a treatment of social effects). **Table 1** provides a glossary of formal notation and terminology used throughout the paper. GLMMs effectively describe the RNs of non-Gaussian phenotypes using additive linear functions on a transformed latent scale (Bolker et al., 2009; Villemereuil et al., 2016). Extensive prior work has been done on appropriate study design and GLMM

implementation for RN research in evolutionary ecology (e.g. see Nussey, Wilson, & Brommer 2007; Martin, Nussey, Wilson, & Réale, 2010; Dingemanse & Dochtermann 2013; O'Dea et al. 2021 among others). Therefore, we avoid reviewing this material in detail here, instead focusing on the introduction of a general form and notation for RN models of any labile trait.

Consider a GLMM for repeated measure *t* of individual *j*, who expressed labile phenotype z_{jt} in environmental state x_{jt}. The distribution of measurements can be predicted using a probability function $f(\mu, \sigma)$ with mean, location, or rate parameter μ and dispersion, shape, or scale parameter σ (e.g. as with normal, gamma, and beta distributions). Link functions g_μ and g_σ are used for modelling the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ across observations so that the RN parameters $\boldsymbol{\mu_0}$, $\boldsymbol{\beta_x}$, and $\boldsymbol{\sigma_0}$ can be expressed as additive linear effects on a transformed scale, irrespective of the assumed distribution of the raw data. For instance, $g_\mu = \mathrm{identity}(\mu)$ and $g_\sigma = \log(\sigma^2)$ are sensible choices for a Gaussian measure. The generalized form of the model is given by

$$z_{jt} \sim f\left(\mu_{jt}, \sigma_j\right) \tag{1}$$

$$g_\mu\left(\mu_{jt}\right) = \mu_0 + \mu_{0j} + \left(\beta_x + \beta_{xj}\right)x_t$$

$$g_\sigma\left(\sigma_j\right) = \sigma_0 + \sigma_{0_j}$$

$$[\boldsymbol{\mu_0}, \boldsymbol{\beta_x}, \boldsymbol{\sigma_0}]^{\mathrm{T}} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{P}): \mathbf{P} = \begin{bmatrix} \mathrm{var}(\boldsymbol{\mu_0}) & \dots & \dots \\ \mathrm{cov}(\boldsymbol{\beta_x}, \boldsymbol{\mu_0}) & \mathrm{var}(\boldsymbol{\beta_x}) & \vdots \\ \mathrm{cov}(\boldsymbol{\sigma_0}, \boldsymbol{\mu_0}) & \mathrm{cov}(\boldsymbol{\sigma_0}, \boldsymbol{\beta_x}) & \mathrm{var}(\boldsymbol{\sigma_0}) \end{bmatrix}$$

Here μ_0, β_x, σ_0 are the average values for the RN intercept, slope, and residual parameters in the population, expressed on the scale of the link functions. Repeatable individual differences in RN parameters are in turn estimated as deviations from these averages using random effects μ_{0j}, β_{xj}, and σ_{0j}. For simplicity, the model assumes environmental exposures $\boldsymbol{x}$ are randomized across individuals, but it may be necessary in non-experimental contexts to center covariates within individuals for appropriate scaling of RN slopes (Schaeffer, 2004; van de Pol & Wright 2009; Araya-Ajoy, Mathot, & Dingemanse, 2015; Westneat et al., 2020; Fay, Martin, & Plard 2022). The magnitude of

among-individual (co)variance in these RN parameters is described by the **P** matrix. See the *supplementary appendix* for further model extensions and **Box 1** for further discussion of the RN residual parameter.

Box 1. ***Interpreting among-individual differences in RN residuals.***

The functional role of the RN residual parameters $\boldsymbol{\sigma_0}$ can be ambiguous because these individual effects are modelled on the dispersion $\boldsymbol{\sigma}$ of the phenotypic distribution, rather than the expectation $\boldsymbol{\mu}$ (**Eq. 1**). Phenotypic variance due to dispersion is generally interpreted as noise or measurement error around individuals' repeatable mean trait values (Brommer 2013), which are determined by the expression of RN intercepts $\boldsymbol{\mu_0}$ and slopes $\boldsymbol{\beta_x}$ across measured environments. However, the residuals of labile traits may also contain repeatable and fitness-relevant variation in how organisms intrinsically regulate their phenotype (Westneat, Wright, & Dingemanse 2015), such as in their assessment and response toward developmental noise within a given environment (Gavrilets & Hastings, 1994; Hansen et al., 2006; Mitchell et al. 2021). Such repeatable *among*-individual differences in *within*-individual variation, described by $\boldsymbol{\sigma_0}$, may arise from a variety of mechanisms regulating patterns of stochastic expression in behavior or other labile traits (Prentice, Houslay, Martin, & Wilson, 2020). For instance, stochasticity can be generated through the repeatable activities of the organism, such as by random sampling of the environment, which can be shaped via reinforcement and punishment to facilitate adaptive learning in novel or uncertain environments (Niv et al. 2002; Barrett 2011; Wright et al., 2022). As a consequence, intrinsic variability may evolve in conjunction with learning mechanisms to track unpredictable shifts in fitness optima during development (Borenstein, Feldman, & Aoki 2008). Predation may also select for greater variability in movement, so as to reduce predators' capacity to predict prey escape trajectories (Hugie, 2003; Moore et al. 2017), while reduced variability may instead be adaptive for reputation formation and trust in repeated social interactions (McNamara & Leimar, 2010). Stochasticity may also result from exogenous factors, such that individual differences in $\boldsymbol{\sigma_0}$ reflect how organisms regulate in response to the environment. For example, when environmental states fluctuate rapidly in an unpredictable and uncontrollable manner, negative selection may act on the RN residual parameter to

promote phenotypic canalization, decreasing susceptibility of the phenotype to developmental perturbation (Flatt 2005; Siegal & Leu 2014; Westneat et al., 2015).

In empirical research, it will often be challenging to distinguish variance in residuals due to intrinsically stochastic variability or unmeasured processes of cue-induced plasticity and individual-by-environment interaction (Westneat et al. 2015; Prentice et al., 2020). Estimates of $\mathrm{var}(\boldsymbol{\sigma_0})$ in the field may, for example, reflect repeatable functional interactions between unmodelled RN slopes and stochastic environmental exposures. Therefore, caution is warranted when inferring the mechanistic underpinnings of $\boldsymbol{\sigma_0}$ outside of well-controlled experiments. Poorly specified statistical models, in which predicted residual processes do not accurately describe observed phenotypic variance, will also inhibit accurate biological inference of RNs (Mitchell, Dujon, Beckmann, & Biro, 2020; Ramakers, Visser, & Gienapp, 2020). Nonetheless, to the degree that individual differences in residuals are repeatable across time and not due to unbalanced sampling or pseudo-repeatability (Dingemanse & Dochtermann 2013), selection can still shape RN residuals, irrespective of whether within-individual deviations arise from mechanisms of intrinsically stochastic or cue-induced trait expression. Therefore, we suggest that researchers in both observational and experimental systems focus their attention on functionally interpreting and operationally defining RN residual parameters with respect to theoretically motivated RN slopes, defined over measured dimensions of environmental change (**Figure 1**).

Box 2. ***Repeatable among-individual differences due to RNs.***

Selection on the RNs of labile traits can only occur if individuals differ in their intercepts, slopes, and residual parameters across time. The covariance matrix **P** in **Eq. 1** describes these repeatable among-individual differences and, therefore, ultimately determines the total amount of trait (co)variation available to natural selection on phenotype $\boldsymbol{z}$ over the sampling period of interest, given that RN parameters $\boldsymbol{\mu_0}$, $\boldsymbol{\beta_x}$, and $\boldsymbol{\sigma_0}$ predict how organisms will repeatedly express their phenotype within and across environments. We denote the total magnitude of repeatable among-individual differences in **z** due to RNs as

$\mathrm{var}(\boldsymbol{\eta})$, which in the general case sets an upper limit on the heritability of a phenotype due to direct genetic effects (see Bijma, 2011 for social traits) and thus provides a useful phenotypic proxy of the evolvability of a trait (Martin et al., 2023). The trait values $\boldsymbol{\eta}$ represent the repeatable character states that organisms are expected to express within and across sampled environments, as predicted by their individual RNs (**Fig. 1** top left). Conversely, any variance in observed trait values $\mathbf{z}$ due to non-repeatable effects $\mathrm{var}(\boldsymbol{\xi}) = \mathrm{var}(\boldsymbol{z}) - \mathrm{var}(\boldsymbol{\eta})$ introduces noise into the estimation of selection gradients defined across sampled environments. Failure to distinguish non-repeatable $\mathrm{var}(\boldsymbol{\xi})$ and repeatable $\mathrm{var}(\boldsymbol{\eta})$ variance in measured phenotypes $\mathrm{var}(\mathbf{z}) = \mathrm{var}(\boldsymbol{\eta}) + \mathrm{var}(\boldsymbol{\xi})$ can thus lead to biased estimates of directional $\boldsymbol{\beta}^*$ and quadratic $\boldsymbol{\gamma}^*$ selection gradients (**Figure 2**). For evolutionary ecologists, correlations between fitness and phenotype that are repeatable over time and potentially heritable across generations will generally be of primary interest, motivating partitioning of $\mathrm{var}(\boldsymbol{\eta})$ from $\mathrm{var}(\mathbf{z})$ with a GLMM (Martin & Jaeggi, 2022).

O'Dea et al. (2022) and de Villemereuil et al. (2016), among others, provide exact analytic solutions and numeric methods for calculating $\mathrm{var}(\boldsymbol{\eta})$ with many commonly used GLMMs. For the general case, $\mathrm{var}(\boldsymbol{\eta})$ can always be approximated on the original data scale, irrespective of model complexity, by using simulation to compare the variance of model predicted phenotypic distributions in the presence $\mathrm{var}\left(\boldsymbol{z}_{\mathrm{pred}}\right)_{\boldsymbol{\eta}}$ and absence $\mathrm{var}\left(\boldsymbol{z}_{\mathrm{pred}}\right)_{-\boldsymbol{\eta}}$ of repeatable individual effects $\boldsymbol{\mu_0}, \boldsymbol{\beta_x}$, and $\boldsymbol{\sigma_0}$, using a large number of random samples.

$$\mathrm{var}(\boldsymbol{\eta}) \approx \mathrm{var}\left(\boldsymbol{z}_{\mathrm{pred}}\right)_{\boldsymbol{\eta}} - \mathrm{var}\left(\boldsymbol{z}_{\mathrm{pred}}\right)_{-\boldsymbol{\eta}} \quad (3)$$

Model predictions can also be used to approximate the total repeatability of among-individual differences in the phenotype on the original data scale for any GLMM

$$\mathrm{R}_{\eta} \approx \frac{\mathrm{var}(\boldsymbol{\eta})}{\mathrm{var}\left(\boldsymbol{z}_{\mathrm{pred}}\right)_{\boldsymbol{\eta}}} \quad (4)$$

The bias of estimated selection gradients will increase as the R_{η} of a phenotype decreases and $\mathrm{var}(\boldsymbol{\xi})$ in turn increases (Spearman, 1904; Searle, 1961). Therefore,

failure to remove non-repeatable causes of variation from observed phenotypic measures is a particularly serious issue when estimating selection on labile traits across heterogeneous environments (**Figure 2**; Dingemanse et al. 2021).

(Non)linear selection model

To model selection on the individual-specific RN parameters μ_{0j}, β_{xj}, and σ_{0j}, the RN GLMM in **Eq. 1** can be expanded to include an additional response model predicting measure *t* of fitness component or proxy ***W***. Linear ***b*** and quadratic ***q*** selection coefficients, as well as other more complex forms of nonlinear selection, can then be estimated directly for the RN parameters.

$$z_{jt} \sim f(\mu_{jt}, \sigma_j) \quad (5)$$

$$g_\mu(\mu_{jt}) = \mu_0 + \mu_{0j} + (\beta_x + \beta_{xj})x_t$$

$$g_\sigma(\sigma_j) = \sigma_0 + \sigma_{0j}$$

$$W_{jt} \sim f(\theta_{jt}, \delta_j)$$

$$g_\theta(\theta_{jt}) = W_0 + W_{0j} + b_1\mu_{0j} + b_2\beta_{xj} + b_3\sigma_{0j}$$

$$+q_1\mu_{0j}^2 + q_2\beta_{xj}^2 + q_3\sigma_{0j}^2 + q_4\mu_{0j}\beta_{xj} + q_5\mu_{0j}\sigma_{0j} + q_6\beta_{xj}\sigma_{0j}$$

Fitness *W* for individual *j* at measurement *t* is described by a GLMM with expectation parameter θ and dispersion parameter δ. The full model thus estimates the RN parameters and their accompanying selection coefficients in the fitness model simultaneously using a multivariate analysis. **Figure 2** visualizes this model structure and explains how it avoids bias by partitioning repeatable sources of (non)linear association between phenotype and fitness. Parameter W_0 is the average fitness on the transformed scale given by link function g_θ. When repeated fitness measures *t* are available, an individual random effect W_{0j} should be estimated to capture repeatable among-individual differences in fitness that are not due to the modelled phenotypes (i.e. unexplained selection). If only a single fitness measure is available, $\mathrm{var}(\boldsymbol{W_0})$ cannot be identified

separately from fitness residual dispersion δ, so these effects should instead be excluded from the analysis.

The polynomial regression in **Eq. 5** can be used to infer short-term population trajectories on the adaptive landscape, under the assumption that a quadratic function effectively approximates the local shape of the individual selection surface on the latent transformed scale (Lande & Arnold 1983; Phillips & Arnold 1989; see *supplementary appendix* for further discussion). However, the values of the ***b*** and ***q*** regression coefficients should only be interpreted as measures of directional and quadratic selection gradients when fitness is a mean-scaled Gaussian response, after appropriately scaling the coefficients (see Stinchcombe et al. 2008; Dingemanse et al. 2021 for details). Analytic expressions can also be used for direct interpretation of coefficients in a log-normal fitness model (Bollen, Morrissey, & Kruuk 2019). However, in the general case, it will be necessary to further process regression coefficients from the fitness model before making quantitative inferences about directional and quadratic selection on the scale of the original data, which is generally of greater biological interest.

Following Lande and Arnold (1983) and Morrissey and Sakrejda (2013), directional $\boldsymbol{\beta}$ and quadratic $\boldsymbol{\gamma}$ selection gradients can be numerically calculated for any GLMM by taking the first ∂ and second ∂^2 partial derivatives of the estimated fitness function with respect to the expected population-level RN parameters $\bar{\mu}_0$, $\bar{\beta}_x$, and $\bar{\sigma}_0$.

$$\beta_{\mu_0} = \frac{\partial\, \mathrm{E}(\bar{W}, \bar{\mu}_0)}{\partial\, \bar{\mu}_0}\, \bar{W}^{-1} \ldots\ \gamma_{\mu_0} = \frac{\partial^2\, \mathrm{E}(\bar{W}, \bar{\mu}_0)}{\partial\, \bar{\mu}_0\, \delta\, \bar{\mu}_0} \bar{W}^{-1} \ldots\ \gamma_{\beta_x \sigma_0} = \frac{\partial^2\, \mathrm{E}(\bar{W}, \bar{\beta}_x)}{\partial\, \bar{\beta}_x\, \delta\, \bar{\sigma}_0} \bar{W}^{-1} \qquad (\mathbf{6.1})$$

where $\bar{W}$ is the expected population fitness on the original data scale, as predicted by the fitness function defined with ***b*** and ***q*** coefficients on the link scale in **Eq. 5**. The directional gradients β_{μ_0}, β_{β_x}, and β_{σ_0} indicate the direction and magnitude of selection on the expected values of population RN parameters, with respect to the original untransformed scale of the data. Quadratic selection gradients γ_{μ_0}, γ_{β_x} and γ_{σ_0} in turn indicate convex or concave curvature in the selection surface shaping the variance of RN parameters (Stinchcombe et al. 2008); and $\gamma_{\mu_0 \beta_x}$, $\gamma_{\mu_0 \sigma_0}$, and $\gamma_{\beta_x \sigma_0}$ indicate further curvature due to the presence of correlational selection between RN parameters (Blows & Brooks, 2003).

These gradients can be expressed in standardized units for effect size comparison between traits and parameters using the appropriate variances and standard deviations (Lande & Arnold 1983)

$$\beta_{\mu_0}^{\mathrm{sd}} = \beta_{\mu_0}\,\mathrm{sd}(\boldsymbol{\mu_0}) \;\ldots\; \gamma_{\mu_0}^{\mathrm{sd}} = \gamma_{\mu_0}\mathrm{var}(\boldsymbol{\mu_0}) \ldots\; \gamma_{\beta_x\sigma_0}^{\mathrm{sd}} = \gamma_{\beta_x\sigma_0}\mathrm{sd}(\boldsymbol{\beta_x})\mathrm{sd}(\boldsymbol{\sigma_0}) \qquad (6.2)$$

Standardized gradients are particularly useful for GLMMs because the magnitude of variances may differ appreciably between the distinct transformed link scales used for estimating RNs and selection, which makes it challenging to meaningfully distinguish between small and large effect sizes across models.

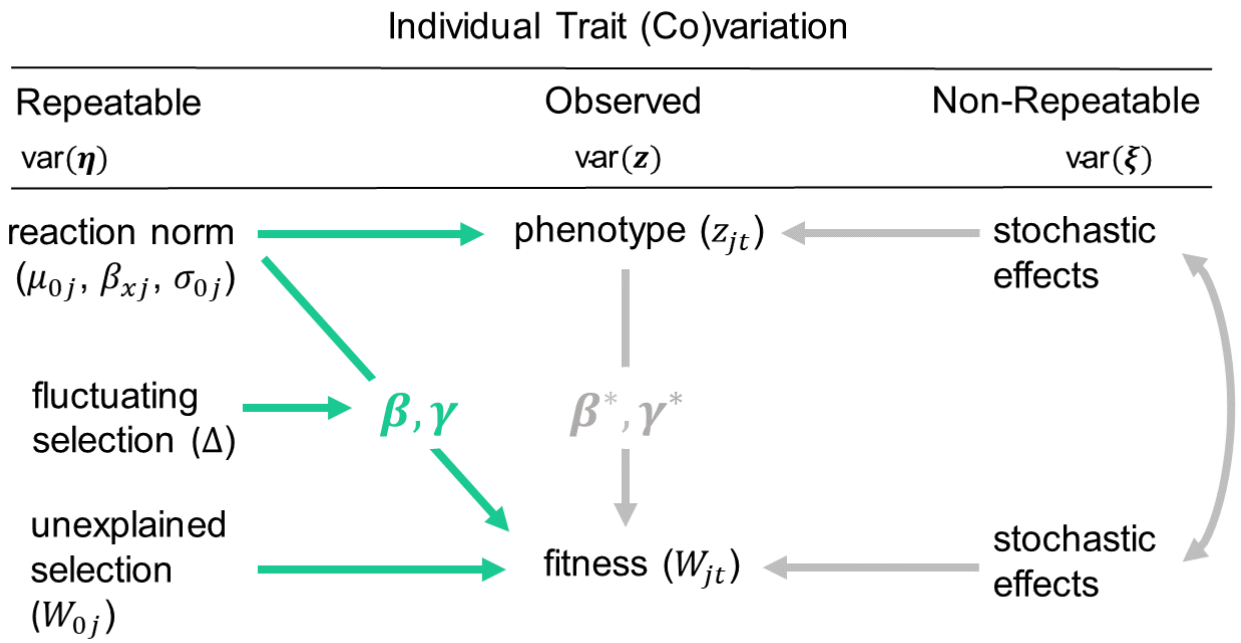

Figure 2. ***Removing non-repeatable effects from selection gradients*****.** The diagram shows causal pathways (directional arrows) by which repeatable (green) and non-repeatable (grey) effects can influence selection gradients of fitness (W) on phenotype (z). Non-repeatable, stochastic effects influence both fitness and phenotype (directional arrows) and may be correlated (double-headed arrow), introducing statistical noise into the selection analysis. This leads to biased directional $\boldsymbol{\beta}^*$ and quadratic gradients $\boldsymbol{\gamma}^*$ when observed variance in the phenotype var(***z***) is used to estimate selection across environments. However, if the (non)linear relationships between phenotype and fitness are modelled independently of

stochastic effects on the phenotype var($\boldsymbol{\xi}$), using RN parameters $\boldsymbol{\mu_0}, \boldsymbol{\beta_x}$, and $\boldsymbol{\sigma_0}$ (**Eq. 1-5**), unbiased selection gradients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be estimated (**Eq. 6**) directly for repeatable among-individual differences in the phenotype var($\boldsymbol{\eta}$) (see **Box 2**). Spatiotemporal fluctuations Δ in these selection gradients can also be described by additional coefficients (see *supplementary appendix* **Eq. 9**), and any repeatable among-individual differences in fitness unexplained by RN parameters can be estimated with random effects $\boldsymbol{W_0}$ when repeated fitness measures are available (**Eq. 5**).

Statistical inference

Bayesian estimation

The proposed models cannot currently be estimated using popular GLMM software packages, due to the need for latent RN parameters to be simultaneously estimated with random and fixed effects across different response models. Fortunately, the Stan statistical programming language (Carpenter et al. 2017), which relies on cutting-edge and computationally efficient Markov Chain Monte Carlo (MCMC) sampling algorithms, provides the flexibility needed for estimating these novel GLMMs within a Bayesian framework. Researchers unfamiliar with the general motivations of Bayesian inference are encouraged to see McElreath (2020) and Gelman et al. (2020) for helpful tips on developing an effective workflow for data analysis. The brms package (Bürkner, 2018) is also a very helpful bridge for writing complex (non)linear Bayesian GLMMs in Stan using familiar R formula syntax. We provide guided tutorials (see **data availability**) for various implementations of the models presented here in Stan.

Prior distributions need to be specified for all the population-level parameters in a Bayesian model. While flat or highly diffuse priors are often recommended in the literature (e.g. Ellison 2004; Villemereuil et al. 2016; Houslay and Wilson 2017), weakly informative or regularizing priors, which place relatively low probability on extreme effect sizes, facilitate more robust inferences with limited sample sizes and should generally be preferred over flat priors (Gelman & Tuerlinckx 2000; Lemoine 2019; McElreath 2020).

This does not necessarily require strong a priori assumptions; general-purpose priors can be used to increase the generalizability and robustness of parameter estimates, even in a state of relative ignorance about the true effect size. See Lemoine (2019) for more detailed discussion and recommendations.

Model validation

Previous work has validated the performance of our general approach in Stan for modest effect sizes, showing robust estimates of directional selection on RN intercepts and slopes with many repeated measures and sample sizes of N = 100 - 300 (Martin & Jaeggi, 2022). To provide more general validation, we further conducted a simulation-based calibration (SBC; Talts et al. 2018; Säilynoja, Bürkner, & Vehtari, 2022) procedure to assess whether the proposed models are unbiased estimators of nonlinear selection under a broader range of scenarios. SBC is a procedure for validating the performance of any Bayesian algorithm across many possible parameter values, as defined by the prior distributions of a generative model. This approach removes the arbitrariness of setting a limited range of fixed parameter values for assessing performance, which can lead to unexpected sources of bias being overlooked in uninvestigated regions of parameter space (e.g. rare but possible combinations of phenotypic variances and selection coefficients). Instead, random parameter values are repeatedly sampled across many simulated datasets. Visual inspection of the correspondence between the generative distributions used to simulate datasets and the subsequent posterior distributions inferred from these datasets allows for detecting sources of bias such as overdispersion, overestimation, or inconsistent model performance for extreme values. A GLMM validated through SBC is thus an unbiased Bayesian estimator with respect to the range of effect sizes described by the prior generative model.

Particular attention was given to the estimation of directional and quadratic selection coefficients during SBC, using 300 simulated datasets assuming conditions of very minimal sampling effort (N = 100 subjects with 3 repeated phenotypic measurements and 2 repeated fitness measures). Parameters were simulated such that $\mu_0, \beta_x, \sigma_0, \boldsymbol{b}, \boldsymbol{q} \sim N(0,1), \mathrm{sd}([\boldsymbol{\mu_0}, \boldsymbol{\beta_x}, \boldsymbol{\sigma_0}, \boldsymbol{W_0}]), \delta \sim \mathrm{exponential}(2), \text{and } \mathrm{cor}([\boldsymbol{\mu_0}, \boldsymbol{\beta_x}, \boldsymbol{\sigma_0}]) \sim \mathrm{LKJ}(2)$.

Note that LKJ refers to the Lewandowski-Kurowicka-Joe distribution, which is useful for generating positive-definite correlation matrices (Gelman et al., 2013). These priors led to a broad range of very small to large selection effect sizes, as well as very small to large effects for the standard deviations and correlations of RNs and the residual fitness standard deviation (δ). Phenotype and fitness were assumed to be Gaussian for computational efficiency, with mean fitness fixed to 1. Following the recommendations of Säilynoja et al. (2022), we computed and visualized the difference in expected cumulative distribution functions between the generative and inferred parameters to perform a quantitative graphical test of the model's performance. As shown in **Figure 3**, our results demonstrated with probability ≥ 0.95 that the posterior distributions of inferred selection coefficients were not systematically higher or lower than the prior distributions used to generate expected selection coefficients. The proposed model thus provides unbiased inference of nonlinear selection on RNs across a broad range of effect sizes, even under conditions of minimal sampling effort.

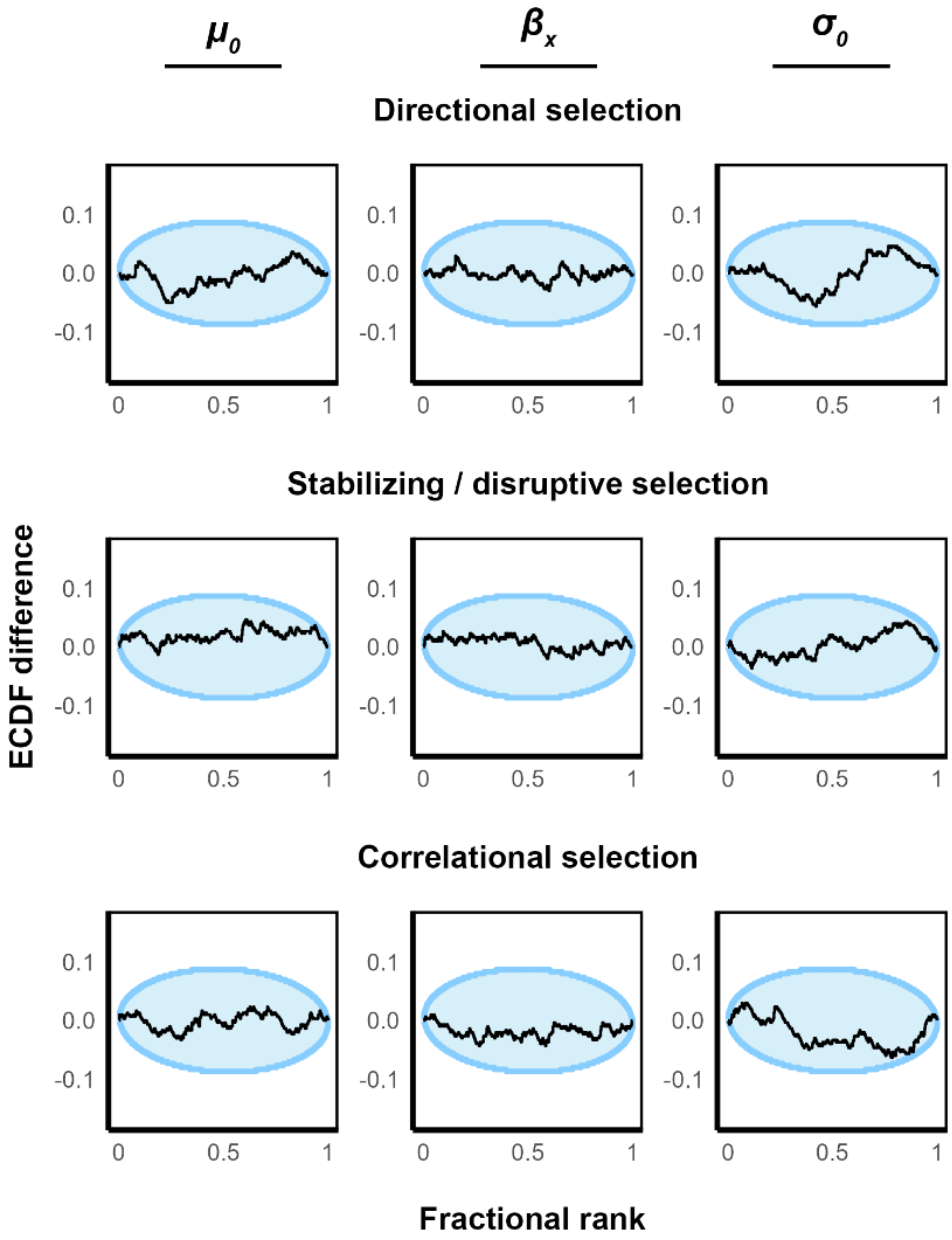

Figure 3. ***Simulation-based calibration of the nonlinear selection model*.** Results are shown for analyses of 300 simulated datasets (N = 100 subjects, 3 repeated phenotype measures and 2 repeated fitness measures) generated from prior distributions defined over the parameters of a Gaussian nonlinear selection model for RNs (**Eq. 5**). Plots show the difference between the expected cumulative density functions (y-axis) for directional and quadratic selection gradients, based on their generative prior distributions $N(0,1)$, and the estimated cumulative density functions based on inferred posterior distributions.

The x-axis indicates the ordered fractional ranks across posterior samples used for computing these comparisons. Blue circles show 90% Bayesian credible intervals for regions of concordance between the estimated and expected parameter distributions, and the black line reflects the observed difference between the expected and inferred distribution (a perfectly horizontal line would thus indicate perfect concordance with the simulated parameters in every dataset). Consistent deviations of the black line beyond the blue region would provide evidence of systematic inferential bias during model estimation. Note that due to stochasticity, fluctuations of the black line within the blue circle are expected at computationally efficient sample sizes.

Power analysis

The SBC procedure demonstrated that our model facilitates unbiased Bayesian estimation across a broad range of parameter values (**Figure 3**). We also conducted an additional simulation study to provide concrete guidelines for empiricists designing studies to assess nonlinear selection on RNs, investigating how the power to detect the direction of selection gradients is influenced by the number of subjects and repeated measures per subject for phenotypes and fitness proxies. For simplicity and ease of effect size comparison, we modelled Gaussian phenotype and fitness measures. Fitness effects for the nonlinear selection model (modified from **Eq. 5**) were simulated such that $\boldsymbol{b}, \boldsymbol{q} \sim U(0.1, 0.5)$, resulting in selection effects ranging from statistically weak to strong in strength, with a mean effect size of $|0.3|$ across datasets. For simplicity, we assumed $W_0 = \delta = 1$ and $\mu_0 = \beta_0 = 0$. Continuous environmental variation ($\boldsymbol{x}$) for quantifying reaction norm slopes was treated as a standardized variable drawn from $\boldsymbol{x} \sim N(0,1)$. Repeatable among-individual differences in RNs were fixed to $\mathrm{sd}([\boldsymbol{\mu_0}, \boldsymbol{\beta}, \boldsymbol{\sigma_0}]) = 0.55$ with correlations drawn from $\mathrm{cor}([\boldsymbol{\mu_0}, \boldsymbol{\beta}, \boldsymbol{\sigma_0}]) \sim LKJ(5)$, and the residual standard deviation of the phenotype was fixed to $\mathrm{sqrt}(\exp(\sigma_0)) = 0.77$, so that repeatable and residual random effect variances were 0.3 and 0.6 respectively. This resulted in each RN parameter exhibiting modest repeatability, $R = 0.2 = \frac{0.3}{3(0.3)+0.6}$ in the absence of phenotypic correlations. Unexplained selection was also fixed to $\mathrm{sd}(\mathbf{W_0}) = 0.55$ for the fitness model.

Power to detect the appropriate direction of selection coefficients was explored with 1000 datasets of varying size drawn from $N \sim U(200, 1000)$ subjects with $t_z \sim U(3,7)$ repeated phenotype and $t_w \sim U(1,5)$ repeated fitness measures per subject. Classical frequentist methods define power with respect to a binary decision rule based on the desired significance level of a null hypothesis test. In Bayesian analysis, 'power' is not precisely defined but may instead refer to the continuous level of support provided for a direct (rather than null) hypothesis test, such as the posterior probability of positive selection occurring on a trait. The power of a Bayesian analysis thus reflects how confident a model is likely to be in the existence and direction of a true selection effect, with p = 0.5 indicating no confidence (+ and – values are equally likely) and p = 1.0 indicating complete confidence in the effect. We herein use 'power' in this sense to refer to the expected posterior probability supporting positive directional and quadratic selection effects on RN parameters.

Power for detecting selection across simulated scenarios is visualized in **Figure 4**, with second-order polynomial lines plotted across datasets to infer general patterns expected in empirical research. As expected, we find that Bayesian power for inferring directional and quadratic selection increases with a greater number of subjects (N) and repeated phenotype (t_z) and fitness measures (t_w), as well as with greater selection effect sizes ($\boldsymbol{b}, \boldsymbol{q}$), while larger absolute phenotypic correlations among RN parameters ($\overline{\mathrm{cor}}$) reduce power, particularly for detecting quadratic selection. Power to detect quadratic selection is lower than for directional selection across small to moderate sample and effect sizes, with power for correlational selection also being relatively lower than stabilizing/disruptive selection except under ideal conditions. This implies that research particularly focused on detecting correlational selection of RNs will require larger samples to attain confident inferences. Power is also consistently lower for detecting all types of selection on RN residual parameters in comparison to RN intercepts and slopes, indicating a need for greater sampling effort in selection studies on phenotypic variability. As with any multivariate selection model, these results show that large sample sizes and sufficient repeated measurements are crucial for robust hypothesis testing, particularly in the presence of weak selection. As a rule of thumb, sample sizes of at least N = 500-1000 will be desirable to appropriately reduce the risk of false negatives, particularly in the

absence of many repeated phenotype and/or fitness measures. The negative effect of RN parameter correlations on power also shows that (non)linear selection will be much easier to detect when RN parameters vary quasi-independently among individuals within a population.

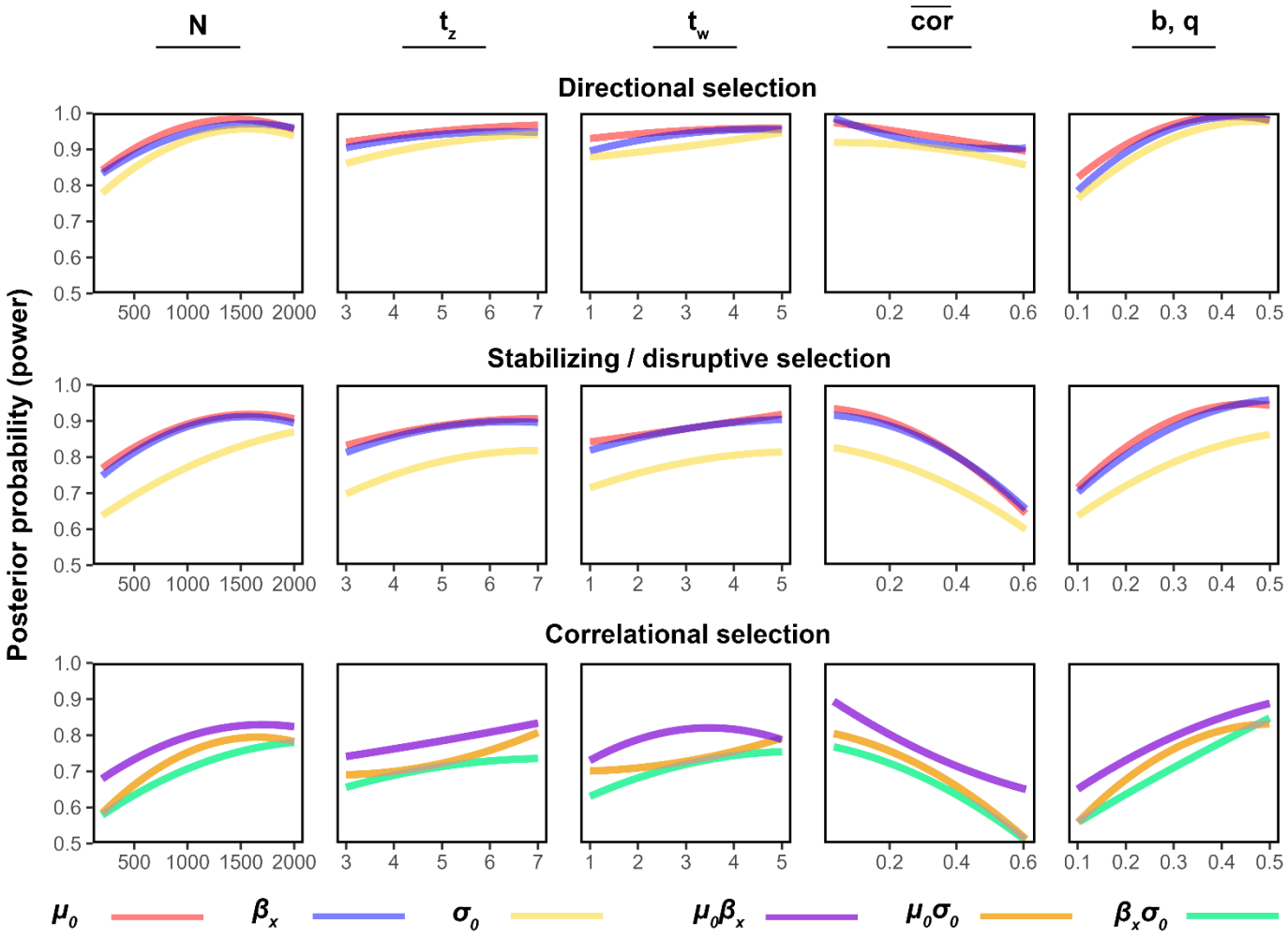

Figure 4. *Bayesian power analysis of the nonlinear selection model*. Results are shown for directional hypothesis tests of selection effects across 1000 simulated datasets used to estimate the nonlinear selection model for RNs (**Eq. 5**) with Gaussian phenotype and fitness measures. Plots show the expected posterior probability ('power', y-axis) supporting selection effects as a function of variation in sampling conditions across simulated datasets (x-axis): the number of subjects/sample size (N), the number of

phenotypic measures per subject (t_z), the number of fitness measures per subject (t_w), the mean absolute correlation among RN parameters ($\overline{cor}$), and the size of linear (b) and nonlinear (q) selection effects. General patterns were inferred using second-order polynomials across conditions, which are color-coded by RN parameter (red = intercepts, blue = slopes, yellow = residuals, purple = intercepts x slopes, orange = intercepts x residuals, and green = slopes x residuals).

Conclusion

Studying selection on highly labile traits is essential for explaining how and why organisms adapt to environmental change. RN models are a crucial tool for characterizing such phenotypes, but their application to selection analysis remains hampered by the limitations of current methods. A major challenge is to avoid inferential bias caused by non-repeatable, stochastic effects and other sources of measurement error in RNs and their fitness effects (Hadfield et al. 2010; **Figure 1-2**). A common solution is to use multi-response/multivariate random effect GLMMs to account for uncertainty in selection on RNs. However, this approach restricts analyses to focus on linear effects and directional selection. Ignoring quadratic selection caused by nonlinear effects fundamentally inhibits researchers' capacity to study the adaptive landscape of labile traits (Bulmer 1971; Arnold et al., 2001; Blows & Brooks, 2003).

To overcome this limitation, we proposed a novel Bayesian GLMM framework for studying complex patterns of nonlinear selection on RNs, which we validated over a broad range of possible parameter values using a simulation-based calibration approach (**Figure 3**). We also found that these models exhibited desirable statistical power under reasonable sampling conditions for many long-term field research projects (**Figure 4**). This modeling framework synthesizes the well-established Lande and Arnold (1983) approach to error-free selection analysis with measurement error or error-in-variables models (Ponzi et al. 2018; Dingemanse et al. 2021; Martin & Jaeggi 2022) and double hierarchical (Westneat et al. 2013; O'Dea et al. 2021), multi-response GLMMs (Brommer et al. 2012; Houslay & Wilson 2017; Arnold et al. 2019). These models can be applied to estimate directional and quadratic selection irrespective of the distribution of the data and

the potential nonlinearity of the RN or fitness function, allowing researchers to construct more realistic models of the processes underlying their measurements. This focuses attention on accurate description of observed data rather than the restrictive assumptions of linear regression. With the analytic toolkit of quantitative genetics (Lande & Arnold 1983; Morrissey & Sakrejda 2013), estimates from these models can also be transformed to quantify selection gradients, visualize multivariate selection, and predict ongoing adaptation. The proposed modeling framework should, therefore, readily enhance tests of adaptive theory for labile traits in the wild.

Data availability statement

R and Stan code with detailed tutorials for implementing the models presented in this paper are available online through a GitHub public repository https://github.com/Jordan-Scott-Martin/Selection-on-RNs.

Acknowledgements

JSM would like to thank Adrian Jaeggi, Adam Hunt, Camila Scaff, and Gabriel Šaffa for their helpful feedback on previous versions of this manuscript, as well as the University of Zurich Candoc/Forschungskredit PhD grant FK-20-034 and Statistical Quantification of Individual Differences (SQuID) educational group postdoctoral fellowship for financial support.

621

References

Abdul-Rahman, F., Tranchina, D., & Gresham, D. (2021). Fluctuating environments maintain genetic diversity through neutral fitness effects and balancing selection. *Molecular Biology and Evolution*, *38*(10), 4362–4375.

Araya-Ajoy, Y. G., & Dingemanse, N. J. (2014). Characterizing behavioural 'characters': An evolutionary framework. *Proceedings of the Royal Society B*, *281*, 20132645.

Araya-Ajoy, Y. G., Dingemanse, N. J., Westneat, D. F., & Wright, J. (2023). The evolutionary ecology of variation in labile traits: selection on its among- and within-individual components. *Evolution*, *77*, 2246–2256.

Araya-Ajoy, Y. G., Mathot, K. J., & Dingemanse, N. J. (2015). An approach to estimate short-term, long-term and reaction norm repeatability. *Methods in Ecology and Evolution*, *6*(12), 1462-1473.

Arnold, S. J., Pfrender, M. E., & Jones, A. G. (2001). The adaptive landscape as a conceptual bridge between micro- and macroevolution. *Genetica*, *112-113*, 9–32.

Arnold, P. A., Nicotra, A. B., & Kruuk, L. E. (2019). Sparse evidence for selection on phenotypic plasticity in response to temperature. *Philosophical Transactions of the Royal Society B*, *374*, 20180185.

Barrett, L. (2011). *Beyond the brain: How body and environment shape animal and human minds*. Princeton University Press.

Bell, A. M., Hankison, S. J., & Laskowski, K. L. (2009). The repeatability of behaviour: A meta-analysis. *Animal Behaviour*, *77*, 771–783.

Bijma, P. (2011). A general definition of the heritable variation that determines the potential of a population to respond to selection. *Genetics*, *189*(4), 1347–1359.

Bijma, P., & Wade, M. J. (2008). The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. *Journal of Evolutionary Biology*, *21*(5), 1175–1188.

Biro, P. A., & Adriaenssens, B. (2013). Predictability as a personality trait: Consistent differences in intraindividual behavioral variation. *The American Naturalist*, *182*, 621–629.

Blows, M. W. (2007). A tale of two matrices: Multivariate approaches in evolutionary biology. *Journal of Evolutionary Biology*, *20*, 1–8.

Blows, M. W., & Brooks, R. (2003). Measuring nonlinear selection. *The American Naturalist*, *162*, 815–820.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135.

Bollen, K. A., & Noble, M. D. (2011). Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences*, *108*, 15639–15646.

Bollen, T., Morrissey, M. B., & Kruuk, L. E. (2019). Estimation of genetic variance in fitness, and inference of adaptation, when fitness follows a log-normal distribution. *Journal of Heredity*, *110*, 383–395.

Bolnick, D. I., Svanbäck, R., Fordyce, J. A., Yang, L. H., Davis, J. M., Hulsey, C. D., & Forister, M. L. (2003). The ecology of individuals: incidence and implications of individual specialization. *The American Naturalist*, *161*(1), 1–28.

Oomen, R. A., & Hutchings, J. A. (2022). Genomic reaction norms inform predictions of plastic and adaptive responses to climate change. *The Journal of Animal Ecology*, *91*, 1073–1087.

Borenstein, E., Feldman, M. W., & Aoki, K. (2008). Evolution of learning in fluctuating environments: When selection favors both social and exploratory individual learning. *Evolution*, *62*, 586–602.

Brommer, J. E. (2013). On between-individual and residual (co) variances in the study of animal personality: Are you willing to take the ’individual gambit’? *Behavioral Ecology and Sociobiology*, *67*, 1027–1032.

Brommer, J. E., Kontiainen, P., & Pietiäinen, H. (2012). Selection on plasticity of seasonal life-history traits using random regression mixed model analysis. *Ecology and Evolution*, *24*, 695–704.

Bulmer, M. G. (1971). The effect of selection on genetic variability. *The American Naturalist*, *201-211*.

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411.

Caño, L., Escarré, J., Fleck, I., Blanco-Moreno, J. M., & Sans, F. X. (2008). Increased fitness and plasticity of an invasive species in its introduced range: a study using Senecio pterophorus. *The Journal of Ecology*, *96*(3), 468–476.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., & A. Riddell... (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *74*.

Catalina, A., Bürkner, P. C., & Vehtari, A. (2020). Projection predictive inference for generalized linear and additive multilevel models. *arXiv*. http://arxiv.org/abs/2010.06994

Cauchoix, M., Chow, P. K. Y., Horik, J. O. V., Atance, C. M., Barbeau, E. J.,...G. B.-J., & Cauchard, L. (2018). The repeatability of cognitive performance: A meta-analysis. *Philosophical Transactions of the Royal Society B*, *373*, 20170281.

Dall, S. R. X., & Griffith, S. C. (2014). An empiricist guide to animal personality variation in ecology and evolution. *Frontiers in Ecology and Evolution*, *14*, 3.

Darwin, C. (1859). *On the origin of species by means of natural selection*. London: John Murray.

Denissen, J. J. A., & Penke, L. (2008). Motivational individual reaction norms underlying the five-factor model of personality: First steps toward a theory-based conceptual framework. *Journal of Research in Personality*, *69*, 1285–1302.

Dingemanse, N. J., & Araya-Ajoy, Y. G. (2015). Interacting personalities: behavioural ecology meets quantitative genetics. *Trends in Ecology & Evolution*, *30*(2), 88–97.

Dingemanse, N. J., Araya-Ajoy, Y. G., & Westneat, D. F. (2021). Most published selection gradients are underestimated: Why this is and how to fix it. *Evolution*, *Early View*.

Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *Journal of Animal Ecology*, *82*, 39–54.

Dingemanse, N. J., Kazem, A. J., Réale, D., & Wright, J. (2010). Behavioural reaction norms: Animal personality meets individual plasticity. *Trends in Ecology and Evolution*, *25*, 81–89.

Dingemanse, N. J., & Wolf, M. (2010). Recent models for adaptive personality differences: a review. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*, 3947-3958.

Eisenegger, C., Haushofer, J., & Fehr, E. (2011). The role of testosterone in social interaction. *Trends in Cognitive Sciences*, *15*, 263–271.

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, *7*, 509–520.

Estes, S., & Arnold, S. J. (2007). Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. *The American Naturalist*, *169*, 227–244.

Fanson, K. V., & Biro, P. A. (2015). Meta-analytic insights into factors influencing the repeatability of hormone levels in agricultural, ecological, and medical fields. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *316*, R101–R109.

Fawcett, T. W., Hamblin, S., & Giraldeau, L.-A. (2013). Exposing the behavioral gambit: the evolution of learning and decision rules. *Behavioral Ecology*, *24*, 2–11.

Fay, R., Martin, J., & Plard, F. (2022). Distinguishing within from between individual effects: How to use the within-individual centering method for quadratic pattern. *Journal of Animal Ecology*, *91*, 8–19.

Flatt, T. (2005). The evolutionary genetics of canalization. *The Quarterly Review of Biology*, *80*, 287–316.

Fox, R. J., Donelson, J. M., Schunter, C., Ravasi, T., & Gaitán-Espitia, J. D. (2019). Beyond buying time: the role of plasticity in phenotypic adaptation to rapid environmental change. *Philosophical Transactions of the Royal Society B*, *374*, 20180174.

Gavrilets, S., & Hastings, A. (1994). A quantitative-genetic model for selection on developmental noise. *Evolution*, *48*, 1478–1486.

Gavrilets, S., & Scheiner, S. M. (1993). The genetics of phenotypic plasticity. VI. Theoretical predictions for directional selection. *Journal of Evolutionary Biology*, *6*, 49–68.

Gelman, A., & Tuerlinckx, F. (2000). Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics*, *15*, 373–390.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., & M. Modrák... (2020). Bayesian workflow. *arXiv Preprint*, *arXiv:2011.01808*.

Ghalambor, C. K., McKay, J. K., Carroll, S. P., & Reznick, D. N. (2007). Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Functional Ecology*, *21*(3), 394–407.

Ghalambor, C. K., Hoke, K. L., Ruell, E. W., Fischer, E. K., Reznick, D. N., & Hughes, K. A. (2015). Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature. *Nature*, *525*(7569), 372–375.

Gomulkiewicz, R., Kingsolver, J. G., Carter, P. A., & Heckman, N. (2018). Variation and evolution of function-valued traits. *Annual Review of Ecology, Evolution, and Systematics*, *49*, 139–164.

Guindre-Parker, S., & Rubenstein, D. R. (2018). Multiple benefits of alloparental care in a fluctuating environment. *Royal Society Open Science*, *5*, 172406.

Hadfield, J. D., Wilson, A. J., Garant, D., & Sheldon, B. C. (2010). The misuse of BLUP in ecology and evolution. *The American Naturalist*, *175*, 116–125.

Hansen, T. F., Carter, A. J. R., & Pélabon, C. (2006). On adaptive accuracy and precision in natural populations. *The American Naturalist*, *168*(2), 168–181.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, *36*, 531–547.

Hendry, A. P. (2016). Key questions on the role of phenotypic plasticity in eco-evolutionary dynamics. *The Journal of Heredity*, *107*, 25–41.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology*, *12*(3), 123–135.

Houslay, T. M., & Wilson, A. J. (2017). Avoiding the misuse of BLUP in behavioural ecology. *Behavioral Ecology*, *28*, 948–952.

Houston, A. I., & McNamara, J. M. (1999) *Models of adaptive behaviour*. Cambridge, MA: Cambridge University Press.

Hugie, D. M. (2003). The waiting game: a “battle of waits” between predator and prey. *Behavioral Ecology, 14*(6), 807–817.

Jaeggi, A. V., Boose, K. J., White, F. J., & Gurven, M. (2016). Obstacles and catalysts of cooperation in humans, bonobos, and chimpanzees: Behavioural reaction norms

can help explain variation in sex roles, inequality, war and peace. *Behaviour*, *153*, 1015–1052.

Kazancıoğlu, E., Klug, H., & Alonzo, S. H. (2012). The evolution of social interactions changes predictions about interacting phenotypes. *Evolution*, *66*, 2056–2064.

Kinsler, G., Schmidlin, K., Newell, D., Eder, R., Apodaca, S., Lam, G., Petrov, D., & Geiler-Samerotte, K. (2023). Extreme sensitivity of fitness to environmental conditions: Lessons from #1BigBatch. *Journal of Molecular Evolution*, *91*, 293–310.

Lande, R., & Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution*, *37*, 1210–1226.

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in bayesian analyses. *Oikos*, *128*.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585.

Martin, J. G., Nussey, D. H., Wilson, A. J., & Réale, D. (2011). Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. *Methods in Ecology and Evolution*, *2*(4), 362-374.

Martin, J. S., & Jaeggi, A. V. (2022). Social animal models for quantifying plasticity, assortment, and selection on interacting phenotypes. *Journal of Evolutionary Biology*, *35*, 520-538.

Martin, J. S., Jaeggi, A. V., & Koski, S. E. (2023). Social evolution of individual differences: Future directions for a comparative science of personality in social behavior. *Neuroscience & BioBehavioral Reviews*, *144*, 104980.

Martin, J. S., Massen, J. J., Šlipogor, V., Bugnyar, T., Jaeggi, A. V., & Koski, S. E. (2019). The EGA+ GNM framework: An integrative approach to modelling behavioural syndromes. *Methods in Ecology and Evolution*, *10*, 245–257.

Martin, J. S., Ringen, E. J., Duda, P., & Jaeggi, A. V. (2020). Harsh environments promote alloparental care across human societies. *Proceedings of the Royal Society B*, *287*, 20200758.

de la Mata, R., Zas, R., Bustingorri, G., Sampedro, L., Rust, M., Hernandez-Serrano, A., & Sala, A. (2022). Drivers of population differentiation in phenotypic plasticity in a temperate conifer: A 27-year study. *Evolutionary Applications*, *15*, 1945–1962.

Mathuru, A. S., Kibat, C., Cheong, W. F., Shui, G., Wenk, M. R., Friedrich, R. W., & Jesuthasan, S. (2012). Chondroitin fragments are odorants that trigger fear behavior in ffsh. *Current Biology*, *22*, 538–554.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in r and stan* (2nd ed.). CRC Press.

McGlothlin, J. W., Moore, A. J., Wolf, J. B., & Brodie, E. D., 3rd. (2010). Interacting phenotypes and the evolutionary process. III. Social evolution. *Evolution*, *64*(9), 2558–2574.

McNamara, J. M., & Houston, A. I. (2009). Integrating function and mechanism. *Trends in Ecology & Evolution*, *24*(12), 670–675.

McNamara, J. M., & Leimar, O. (2010). Variation and the response to variation as a basis for successful cooperation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*, 2627–2633.

McNamara, J. M., & Leimar, O. (2020). *Game theory in biology*. Oxford University Press.

Mitchell, D. J., Beckmann, C., & Biro, P. A. (2021). Understanding the unexplained: The magnitude and correlates of individual differences in residual variance. *Ecology and Evolution*, *11*, 7201–7210.

Mitchell, D. J., Dujon, A. M., Beckmann, C., & Biro, P. A. (2020). Temporal autocorrelation: a neglected factor in the study of behavioral repeatability and plasticity. *Behavioral Ecology*, *31*, 222–231.

Moore, A. J., Brodie, E. D., 3rd, & Wolf, J. B. (1997). Interacting phenotypes and the evolutionary process: I. Direct and indirect genetic effects of social interactions. *Evolution*, *51*(5), 1352–1362.

Moore, T. Y., Cooper, K. L., Biewener, A. A., & Vasudevan, R. (2017). Unpredictability of escape trajectory explains predator evasion ability and microhabitat preference of desert rodents. *Nature Communications*, *8*, 1–9.

Morrissey, M. B., Parker, D. J., Korsten, P., Pemberton, J. M., Kruuk, L. E. B., & Wilson, A. J. (2012). The prediction of adaptive evolution: Empirical application of the secondary theorem of selection and comparison to the breeder's equation. *Evolution*, *66*, 2399–2410.

Morrissey, M. B., & Sakrejda, K. (2013). Unification of regression-based methods for the analysis of natural selection. *Evolution*, *67*, 2094–2100.

Mouchet, A., Cole, E. F., Matthysen, E., Nicolaus, M., Quinn, J. L., Roth, A. M., Tinbergen, J. M., van Oers, K., van Overveld, T., & Dingemanse, N. J. (2021). Heterogeneous selection on exploration behavior within and among West European populations of a passerine bird. *Proceedings of the National Academy of Sciences*, *118*.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, *33*, 341–365.

Munar-Delgado, G., Araya-Ajoy, Y. G., & Edelaar, P. (2023). Estimation of additive genetic variance when there are gene–environment correlations: Pitfalls, solutions and unexplored questions. *Methods in Ecology and Evolution*, *14*, 1245-1258.

Nakahashi, W., & Ohtsuki, H. (2015). When is emotional contagion adaptive? *Journal of Theoretical Biology*, *380*, 480–488.

Nettle, D., & Penke, L. (2010). Personality: Bridging the literatures from human psychology and behavioural ecology. *Philosophical Transactions of the Royal Society B*, *365*, 4043–4050.

Newediuk, L., Prokopenko, C. M., & Wal, E. V. (2022). Individual differences in habitat selection mediate landscape level predictions of a functional response. *Oecologia*, 1–12.

Niemelä, P. T., & Dingemanse, N. J. (2018). Meta-analysis reveals weak associations between intrinsic state and personality. *Proceedings of the Royal Society B*, *285*, 20172823.

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior*, *10*, 5–24.

Nussey, D. H., Wilson, A. J., & Brommer, J. E. (2007). The evolutionary ecology of individual phenotypic plasticity in wild populations. *Journal of Evolutionary Biology*, *20*, 831–844.

O'Dea, R. E., Noble, D. W., & Nakagawa, S. (2021). Unifying individual differences in personality, predictability and plasticity: A practical guide. *Methods in Ecology and Evolution*, *13*, 278-293.

Ørsted, M., Rohde, P. D., Hoffmann, A. A., Sørensen, P., & Kristensen, T. N. (2018). Environmental variation partitioned into separate heritable components. *Evolution*, *72*, 136–152.

Phillips, P. C., & Arnold, S. J. (1989). Visualizing multivariate selection. *Evolution*, *43*, 1209–1222.

Pick, J. L., Lemon, H. E., Thomson, C. E., & Hadfield, J. D. (2022). Decomposing phenotypic skew and its effects on the predicted response to strong selection. *Nature Ecology & Evolution*, *6*, 774–785.

Pol, M. van de, & Wright, J. (2009). A simple method for distinguishing within- versus between-subject effects using mixed models. *Animal Behaviour*, *77*, 753–758.

Ponzi, E., Keller, L. F., Bonnett, T., & Muff, S. (2018). Heritability, selection, and the response to selection in the presence of phenotypic measurement error: Effects, cures, and the role of repeated measurements. *Evolution*, *72*, 1992–2004.

Prentice, P. M., Houslay, T. M., Martin, J. G. A., & Wilson, A. J. (2020). Genetic variance for behavioural “predictability” of stress response. *Journal of Evolutionary Biology*, *33*(5), 642–652.

Projecto-Garcia, J., Biddle, J. F., & Ragsdale, E. J. (2017). Decoding the architecture and origins of mechanisms for developmental polyphenism. *Current Opinion in Genetics & Development*, *47*, 1-8.

Ramakers, J. J. C., Visser, M. E., & Gienapp, P. (2020). Quantifying individual variation in reaction norms: Mind the residual. *Journal of Evolutionary Biology*, *33*, 352–366.

Reynolds, R. J., de Los Campos, G., Egan, S. P., & Ott, J. R. (2016). Modelling heterogeneity among fitness functions using random regression. *Methods in Ecology and Evolution*, *7*, 70–79.

Royauté, R., Berdal, M. A., Garrison, C. R., & Dochtermann, N. A. (2018). A meta-analysis of the pace-of-life syndrome hypothesis. *Behavioral Ecology and Sociobiology*, *72*, 1–10.

Royauté, R., Hedrick, A., & Dochtermann, N. A. (2020). Behavioural syndromes shape evolutionary trajectories via conserved genetic architecture. *Proceedings of the Royal Society B*, *287*, 20200183.

Säilynoja, T., Bürkner, P.-C., & Vehtari, A. (2022). Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison. *Statistics and Computing*, *32*, 32.

Sasaki, A., & Ellner, S. (1997). Quantitative genetic variance maintained by fluctuating selection with overlapping generations: Variance components and covariances. *Evolution*, *51*, 682–696.

Schaum, C.-E., Buckling, A., Smirnoff, N., & Yvon-Durocher, G. (2022). Evolution of thermal tolerance and phenotypic plasticity under rapid and slow temperature fluctuations. *Proceedings. Biological Sciences / The Royal Society*, *289*(1980), 20220834.

Scheiner, S. M., & Lyman, R. F. (1991). The genetics of phenotypic plasticity. II. Response to selection. *Journal of Evolutionary Biology*, *4*, 23-50.

Scheiner, S. M. (1993). Genetics and Evolution of Phenotypic Plasticity. *Annual Review of Ecology and Systematics*, *24*(1), 35–68.

Scheiner, S. M., Donohue, K., Mazer, L. A. D. S. J., & Wolfe, L. M. (2002). Reducing environmental bias when measuring natural selection. *Evolution*, *56*, 2156–2167.

Schlichting, C. D., & Pigliucci, M. (1998). *Phenotypic evolution: a reaction norm perspective*. Sinauer Associates: Sunderland, MA.

Schluter, D., & Nychka, D. (1994). Exploring fitness surfaces. *The American Naturalist*, *143*, 597–616.

Searle, S. R. (1961). Phenotypic, genetic and environmental correlations. *Biometrics*, *17*, 474–480.

Sigourney, D. B., Munch, S. B., & Letcher, B. H. (2012). Combining a Bayesian nonparametric method with a hierarchical framework to estimate individual and temporal variation in growth. *Ecological Modelling*, *247*, 125–134.

Siegal, M. L., & Leu, J. Y. (2014). On the nature and evolutionary impact of phenotypic robustness mechanisms. *Annual Review of Ecology, Evolution and Systematics*, *45*, 495–517.

Sih, A., Mathot, K. J., Moirón, M., Montiglio, P. O., Wolf, M., & Dingemanse, N. J. (2015). Animal personality and state–behaviour feedbacks: A review and guide for empiricists. *Trends in Ecology and Evolution*, *30*, 50–60.

Silk, J. B., Roberts, E. R., Barrett, B. J., Patterson, S. K., & Strum, S. C. (2017). Female–male relationships influence the form of female–female relationships in olive baboons, Papio anubis. *Animal Behaviour*, *131*, 89–98.

Skinner, B. F. (1966). The phylogeny and ontogeny of behavior. *Science*, *153*, 1205–1213.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.

Stamps, J. A. (2016). Individual differences in behavioural plasticities. *Biological Reviews*, *91*, 534–567.

Stinchcombe, J. R., Agrawal, A. F., Hohenlohe, P. A., Arnold, S. J., & Blows, M. W. (2008). Estimating nonlinear selection gradients using quadratic regression coefficients: Double or nothing? *Evolution*, *68*.

Stinchcombe, J. R., Rutter, M. T., Burdick, D. S., Tiffin, P., Rausher, M. D., & Mauricio, R. (2002). Testing for environmentally induced bias in phenotypic estimates of natural selection: theory and practice. *The American Naturalist*, *160*, 511–523.

Stinchcombe, J. R., Simonsen, A. K., & Blows, M. W. (2014). Estimating uncertainty in multivariate responses to selection. *Evolution*, *68*.

Strickland, K., Mitchell, D. J., Delmé, C., & Frère, C. H. (2021). Repeatability and heritability of social reaction norms in a wild agamid lizard. *Evolution*, *75*.

Suzuki, Y., & Nijhout, H. F. (2006). Evolution of a polyphenism by genetic accommodation. *Science*, *311*, 650–652.

Svensson, E. I., Gomez-Llano, M., & Waller, J. T. (2020). Selection on phenotypic plasticity favors thermal canalization. *Proceedings of the National Academy of Sciences USA*, *117*.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv*. http://arxiv.org/abs/1804.06788

Tonsor, S. J., Elnaccash, T. W., & Scheiner, S. M. (2013). Developmental instability is genetically correlated with phenotypic plasticity, constraining heritability, and fitness. *Evolution*, *67*, 2923–2935.

Vasey, G. L., Weisberg, P. J., & Urza, A. K. (2022). Intraspecific trait variation in a dryland tree species corresponds to regional climate gradients. *Journal of Biogeography*, *49*, 2309–2320.

Vercken, E., Wellenreuther, M., Svensson, E. I., & Mauroy, B. (2012). Don't fall off the adaptation cliff: when asymmetrical fitness selects for suboptimal traits. *PloS One*, *7*, e34889.

Via, S., Gomulkiewicz, R., Jong, G. D., Scheiner, S. M., Schlichting, C. D., & Tienderen, P. H. V. (1995). Adaptive phenotypic plasticity: Consensus and controversy. *Trends in Ecology and Evolution*, *10*, 212–217.

de Villemereuil, P., Charmantier, A., Arlt, D., Bize, P., Brekke, P., Brouwer, L.,..., & Chevin, L. M. (2020). Fluctuating optimum and temporally variable selection on breeding date in birds and mammals. *Proceedings of the National Academy of Sciences*, *117*, 31969–31978.

Villemereuil, P. de, Schielzeth, H., Nakagawa, S., & Morrissey, M. (2016). General methods for evolutionary quantitative genetic inference from generalized mixed models. *Genetics*, *204*, 1281–1294.

Volis, S., Ormanbekova, D., & Yermekbayev, K. (2015). Role of phenotypic plasticity and population differentiation in adaptation to novel environmental conditions. *Ecology and Evolution*, *5*(17), 3818–3829.

Wagner, G. P., Booth, G., & Bagheri-Chaichian, H. (1997). A population genetic theory of canalization. *Evolution*, *51*, 329–347.

Wang, S. P., & Althoff, D. M. (2019). Phenotypic plasticity facilitates initial colonization of a novel environment. *Evolution*, *73*(2), 303–316.

Weis, A. E., & Gorman, W. L. (1990). Measuring selection on reaction norms: An exploration of the eurosta-solidago system. *Evolution*, *44*, 820–831.

Westneat, D. F., Araya-Ajoy, Y. G., Allegue, H., Class, B., Dingemanse, N., Dochtermann, N. A., ... & Schielzeth, H. (2020). Collision between biological process and statistical analysis revealed by mean centring. *Journal of Animal Ecology*, *89*(12), 2813-2824.

Westneat, D. F., Schofield, M., & Wright, J. (2013). Parental behavior exhibits among-individual variance, plasticity, and heterogeneous residual variance. *Behavioral Ecology*, 24, 598-604.

Westneat, D. F., Wright, J., & Dingemanse, N. J. (2015). The biology hidden inside residual within-individual phenotypic variation. *Biological Reviews*, *90*, 729–743.

Wingfield, J. C., Hegner, R. E., Jr., A. M. D., & Ball, G. F. (1990). The 'challenge hypothesis': Theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. *The American Naturalist*, *136*, 829–846.

Wolf, M., & Weissing, F. J. (2010). An explanatory framework for adaptive personality differences. *Philosophical Transactions of the Royal Society B*, *365*, 3959–3968.

Wright, J., Bolstad, G. H., Araya-Ajoy, Y. G., & Dingemanse, N. J. (2019). Life-history evolution under fluctuating density-dependent selection and the adaptive alignment of pace-of-life syndromes. *Biological Reviews*, *94*, 230–247.

Wright, J., Haaland, T. R., Dingemanse, N. J., & Westneat, D. F. (2022). A reaction norm framework for the evolution of learning: how cumulative experience shapes phenotypic plasticity. *Biological Reviews*. https://doi.org/10.1111/brv.12879

Yamahira, K., Kawajiri, M., Takeshi, K., & Irie, T. (2007). Inter- and intrapopulation variation in thermal reaction norms for growth rate: evolution of latitudinal compensation in ectotherms with a genetic constraint. *Evolution*, *61*, 1577–1589.

Estimating (non)linear selection on reaction norms:

A general framework for labile traits

Supplementary appendix:

Model extensions

Jordan S. Martin[*1], Yimen Araya-Ajoy[2], Niels J. Dingemanse[3],
Alastair J. Wilson[4], & David Westneat[5]

*corresponding author: jordan.martin@uzh.ch

[1]*Human Ecology Group, Institute of Evolutionary Medicine,*
University of Zurich Switzerland

[2]*Center for Biodiversity Dynamics, Department of Biology,*
Norwegian University of Science and Technology, Norway

[3]*Behavioral Ecology Unit, Department of Biology,*
Ludwig Maximilian University of Munich, Germany

[4]*Evolution Group, Centre for Biosciences,*
University of Exeter, United Kingdom

[5]*Department of Biology,*
University of Kentucky, United States of America

Model extensions

Simplified models are presented in the main text (**Eq. 1**, **Eq. 5**) to aid interpretation, but it will often be necessary to specify more complex models for explaining empirically observed variation in fitness and phenotype. Various model extensions can be straightforwardly accomplished using the basic toolkit of GLMMs and related regression frameworks, along with appropriate study design and sufficient repeated sampling for reliable estimation. Below we briefly consider three key areas for model extension and provide references for further consideration. Implementation for social traits and interactions is discussed by Martin and Jaeggi (2022).

Adjusted and nonlinear effects

As with any regression analysis, additional fixed and random effects may need to be adjusted for to facilitate appropriate biological inference. Predation may, for instance, cause differential mortality as a function of repeatable differences in behavior across sex and age classes, but this selection will not generate an evolutionary response on behavioral variation within sexes or age classes. This motivates estimating repeatable individual variation adjusted for the effects of sex and age, among other commonly studied factors such as size and morphology (Bolnick et al., 2003). Unadjusted environmental effects on fitness and phenotype can also bias estimates of selection and among-individual variation in both field and laboratory settings (Scheiner et al. 2002; Stinchcombe et al., 2022; Kinsler et al., 2023; Munar-Delgado et al., 2023). It is, therefore, often useful to include additional environmental covariates (e.g. average temperature and rainfall, date within season, resource availability), including potential interaction effects, and random factors (e.g. nesting site, spatial position, batch, observer identity) to adjust fitness variation during the selection analysis. As discussed in **Box 2**, model predictions can always be used to quantify and better understand how adjusting for these effects changes the repeatable variation available to selection in any multivariate GLMM.

Relationships between fitness, phenotype, and the local environment may also be best described by additional terms beyond quadratic regression coefficients. For example, RN slopes of thermoregulatory and life history traits such as growth rate are often highly nonlinear in response to temperature (Oomen & Hutchings, 2022),

violating the assumption of **Eq. 5** that individuals' phenotypic deviations from the linear RN slope β_x are multivariate normally distributed. Polynomials (Henderson, 1982; Yamahira, Kawajiri, Takeshi, & Irie, 2007) or generalized additive effects such as splines or Gaussian processes (Schluter & Nychka, 1994; Sigourney, Munch, & Letcher, 2012; Pederson, Miller, Simpson, & Ross, 2019; Catalina, Bürkner, & Vehtari, 2020) can be used to account for nonlinearity in the population RN and ensure the statistical model more accurately predicts observable phenotypic and fitness variation. In the general case, the basic model (**Eq. 5**) can be expanded to include any generalized additive function *s*() describing how expected phenotypic μ_{jt} or fitness values θ_{jt} change in response to the environment

$$g_\mu(\mu_{jt}) = \mu_0 + \mu_{0j} + s(x_t) + \beta_{xj}x_t \quad (7)$$

$$g_\theta(\theta_{jt}) = W_0 + W_{0j} + s(x_t) + b_1\mu_{0j} + b_2\beta_{x_j} + b_3\sigma_{0j} \ldots$$

Extensive tutorials for incorporating such nonlinear effects into Bayesian regression models in Stan are freely available online (see https://mc-stan.org/users/documentation/case-studies for worked examples of fitting splines and Gaussian processes). Code from Stan models constructed using familiar R syntax in the brms package (Bürkner, 2019) also provides a helpful reference point for getting started. By allowing for arbitrarily complex average RN shapes across subjects, individual *deviations* $\boldsymbol{\beta_x}$ from the average slope for phenotype as well as for fitness are much more likely to exhibit multivariate normality. This general approach allows researchers to accurately describe trait change across complex and dynamic environments, while still using standard theory from quantitative genetics to quantify selection gradients and predict short-term evolutionary responses.

Additional individual effects

The RN model presented in the main text (**Eq. 1**) does not account the fact that phenotypic dispersion σ may also be plastic across environments, a phenomenon broadly referred to as 'malleability' (see O'Dea, Noble, & Nakagawa 2021 for discussion). Malleability in residuals can be estimated by including population- and individual-level slopes in the linear predictor of the dispersion parameter (Westneat et al., 2013). For example,

$$g_\sigma(\sigma_{jt}) = \sigma_0 + \sigma_{0j} + (\rho + \rho_j)x_t \quad (8)$$

if observation-level variation in environmental measure **x** is expected to have effect ρ on average differences in phenotypic residuals. Malleability can then be treated as a further RN parameter that is also potentially under selection. Some statistical distributions such as the Poisson lack an explicit dispersion parameter, due to deterministic mean-variance relationships, and thus at first glance only provide scope for selection on the RN intercepts and slopes of expected values. However, in many empirical datasets, there is more variance observed in the phenotype than predicted by these distributions (overdispersion), which can be accounted for through the inclusion of further random effects capturing stochastic, observation-level deviations from model expectations (i.e. residuals; Harrison, 2014). Taking the same approach described in **Eq. 5** and **Eq. 8**, the dispersion of these observation-level random effects can then be modelled as a function of individual-level intercepts and slopes, similar to a standard Gaussian model, providing scope for estimating selection on phenotypic variability using a broad range of RN GLMMs.

More generally, any theoretically relevant component of a statistical distribution may be modelled as a function of further individual-level effects and conceptualized as a RN parameter regulating the expression of phenotypes within and across environments. Hurdle models, for example, combine multiple distributions together to distinguish effects on the presence/absence of trait expression from effects on the subsequent magnitude or intensity of trait expression (Mullahy 1986; Heilbron 1994). This is particularly useful for phenotypes such as allogrooming behavior in primates, which can vary repeatably among individuals both in its probability of occurring as well as its intensity and duration once expressed (Silk et al., 2017). These processes are interdependent but may nonetheless be subject to distinct selection pressures (e.g. whom should be groomed and how much), which can be investigated by estimating separate RN intercepts and/or slopes on both model components.

Fluctuating selection

Fluctuating selection on RNs may occur due to variation in the density of mates and competitors, resource availability and seasonality, bodily condition and age, the availability of local niches, or any other state that modulate the fitness costs and

benefits of labile traits (Houston & McNamara, 1999; Sih et al., 2015). Fluctuating selection is also expected to be a key mechanism for explaining patterns of macroevolutionary stasis (Estes & Arnold, 2007), as well as the adaptive maintenance of individual and genetic variation within populations (e.g. Sasaki & Ellner, 1997; Dingemanse & Wolf, 2010; Wolf & Weissing, 2010; Wright et al., 2019; Abdul-Rahman, Tranchina, & Gresham, 2021; Martin et al., 2023). In many cases, it will be informative to estimate spatiotemporal heterogeneity in selection even if the underlying causes of fluctuations are not directly measured (Reynolds, de Los Campos, Egan, & Ott 2016). For example, long-term field studies can be used to investigate the adaptive maintenance of RN variation by yearly fluctuations in selection, even if the mechanisms underpinning these fluctuations remain unclear (e.g. de Villemereuil et al., 2020; Mouchet et al., 2021). To incorporate these effects, the basic fitness model (**Eq. 5**) can also be extended by including fixed or random interaction effects on the selection coefficients, which will estimate continuous or discrete fluctuations $\boldsymbol{\Delta\beta}$ and $\boldsymbol{\Delta\gamma}$ (**Figure 1**) across space and time. For example,

$$g_\theta(\theta_{jt}) = W_0 + W_{0j} + (b_1 + b_{1x}x_t + u_{tb_1})\mu_{0j} + \cdots (q_1 + q_{1x}x_t + u_{tq_1})\mu_{0j}^2 + \cdots \quad (9)$$

where b_{1x} and q_{1x} describe how the (non)linear selection coefficients change as a function of x_t, and u_{tb_1} and u_{tq_1} describe changes due to a random factor at time *t*.