

Size correction in biology: how reliable are approaches based on (common) principal component analysis?

Daniel Berner

Received: 12 August 2010 / Accepted: 1 February 2011 / Published online: 22 February 2011
© Springer-Verlag 2011

Abstract Morphological traits typically scale with the overall body size of an organism. A meaningful comparison of trait values among individuals or populations that differ in size therefore requires size correction. A frequently applied size correction method involves subjecting the set of n morphological traits of interest to (common) principal component analysis [(C)PCA], and treating the first principal component [(C)PC1] as a latent size variable. The remaining variation (PC2–PC n) is considered size-independent and interpreted biologically. I here analyze simulated data and natural datasets to demonstrate that this (C)PCA-based size correction generates systematic statistical artifacts. Artifacts arise even when all traits are tightly correlated with overall size, and they are particularly strong when the magnitude of variance is heterogeneous among the traits, and when the traits under study are few. (C)PCA-based approaches are therefore inappropriate for size correction and should be abandoned in favor of methods using univariate general linear models with an adequate independent body size metric as covariate. As I demonstrate, (C)PC1 extracted from a subset of traits, *not* themselves subjected to size correction, can provide such a size metric.

Keywords Bias · Body size · Morphology · Multivariate statistics · Shape

Communicated by Pedro Peres-Neto.

Electronic supplementary material The online version of this article (doi:10.1007/s00442-011-1934-z) contains supplementary material, which is available to authorized users.

D. Berner (✉)
Zoological Institute, University of Basel,
4051 Basel, Switzerland
e-mail: daniel.berner@unibas.ch

Introduction

Morphological traits typically scale with an organism's overall body size. As a consequence, differences in trait values among individuals within populations, and between populations and species, will often arise simply because individuals or populations differ in body size. Although differences in size may be interesting in their own right, biologists are often more interested in trait differences that remain once differences in overall size have been partialled out—i.e., after size correction.

Size correction can be viewed to serve two major purposes: the first is the quantification of variation in and associations among traits *within* a population (note that this paper is generally concerned with multivariate data). In other words, the focus is here the estimation of the phenotypic or genetic variance and covariance (or correlation) structure among morphological variables while controlling for variation in growth among individuals. Information on trait variation and associations is important because it helps us explore the genetic architecture of multivariate phenotypes and characterize a population's raw material for evolution (Lynch and Walsh 1998; Fox and Wolf 2006). The second purpose of size correction is to quantify differences in trait means *among* two or more populations (or species) while controlling for variation in growth among the populations. These trait differences may, for instance, reflect phenotypic plasticity in response to experimental treatments, or adaptive genetic divergence between populations occupying distinct habitats.

Morphological analyses are now often performed using landmark-based geometric morphometrics, where a consensus on the separation of size and shape information has been reached (Zelditch et al. 2004 and references therein). Many organismal features, however, are not amenable to

landmark-based analysis and are instead quantified as linear distance measurements or counts. For these types of morphological datasets, several size correction techniques have been introduced that fall into two families of approaches. In the first family, body size is estimated by a size metric mathematically (but not biologically) independent from the actual morphological traits of interest. An example of a size metric used frequently is an organism's body length. When studying a single population, this approach typically involves linear regression of each focal trait separately on the body size metric. The residuals of the regressions are retained and treated as size-independent morphology (Reist 1985), for instance to estimate the (co)variance or correlation structure (e.g., Losos et al. 1998; Hulsey et al. 2007; Berner et al. 2010). The extension to situations where mean trait differences between two or more populations are the focus involves adding population as a factor in the general linear model (GLM) (i.e., using an ANCOVA design; Reist 1986; Darlington and Smulders 2001; Garcia-Berthou 2001). This is necessary because linear regression with individuals pooled over multiple populations will confound the within- and among-population association between each trait and the size metric. Differences in trait means among populations are then quantified by the population coefficients estimated by the GLMs (e.g., Cotton et al. 2004; Revell et al. 2007; Hoverman and Relyea 2008). Note that ratio-based methods have been introduced as an alternative to the above GLM-based size correction approaches (e.g., Mosimann and James 1979; Jungers et al. 1995). Ratio methodology, however, has been shown to be problematic (Albrecht et al. 1993) and is therefore not considered in this paper.

The second family of size correction methods does not use a size metric measured separately (and hence mathematically independent) from the traits of interest. Instead, a latent body size variable is estimated by the first principal component (PC1) extracted from the (co)variance (or correlation) matrix of the (typically log-transformed) morphological traits themselves. The rationale is that if there is substantial variation among individuals in overall body size and the focal morphological traits are well correlated with size, PC1 should essentially reflect size variation (Jolicoeur 1963). PC1 will then capture a large proportion of the total trait variation, and all traits will display relatively similar loadings (coefficients) of the same sign on PC1. All the remaining variation (i.e., orthogonal to PC1) is considered size-independent shape information.

When dealing with a single population, this size-independent (co)variation is often analyzed directly using as new variables PC2–PC n (where n is the total number of traits) (e.g., Lande and Arnold 1983; Jockusch 1997; Merilä and Björklund 1999; Badyaev and Hill 2000; Holtmeier 2001; McGuigan et al. 2003; Bolnick and Lau

2008). Alternatively, the original traits are regressed on their PC1 scores and the residuals treated as size-corrected morphological variables (e.g., Schluter and McPhail 1992; Schluter 1996; Berner et al. 2008). This PC1-residuals approach seems particularly appealing because subsequent analyses (e.g., estimation of correlation structure) can still be conducted in terms of the n originally measured morphological variables, rather than $n - 1$ composite variables. A more efficient calculation yielding PC1-corrected trait values identical to PC1-residuals involves projecting individuals into the trait space orthogonal to PC1, followed by back-projection to the original trait axes (Burnaby 1966). This procedure is commonly referred to as Burnaby back-projection (hereafter BBP). It is crucial to recognize that these methods of PC1-based size correction are mathematically equivalent, although the size-corrected data appear in different coordinate systems. For instance, individual scores and trait loadings on PC2 extracted from the original traits are *identical* to scores and loadings on PC1 extracted from PC1-regression residuals or from the variables obtained by BBP.

The extension of PC-based size correction to situations with multiple populations involves the estimation of a latent size variable *shared* among the populations. This is most conveniently done by common principal component analysis (CPCA; Flury 1988), which allows testing whether the assumption of a common first principal component (CPC1) is met. Note that CPC1 will often be approximated well by the pooled within-group PC1. However, extracting PC1 simply from the individuals pooled over the populations will confound within- and among-population trait (co)variation. Next, individuals are projected into the trait space orthogonal to CPC1, and back-projected to the original traits (Burnaby 1966; Rohlf and Bookstein 1987; Klingenberg 1996; McCoy et al. 2006; see also Humphries et al. 1981 for a related procedure called shearing). The new trait values are then considered size-independent and used to quantify differences in population means by univariate or multivariate GLMs (e.g., McCoy and Bolker 2008; Touchon and Warkentin 2010).

To summarize, a key difference between the two families of size correction approaches is the body size metric used; an independent size variable versus first (common) principal component scores extracted directly from the focal traits. In what follows, I will refer to the former as ISMSC (for 'independent size metric size correction'), and to the latter as PCSC (for 'principal component size correction'). The purpose of this paper is to demonstrate that all types of PCSC produce statistical artifacts. Although poor performance of PC-based size correction has sometimes been noticed in specific empirical contexts (e.g., Reist 1985; Jungers et al. 1995), no study has explicitly examined the underlying mechanism, or scrutinized under

which conditions bias is particularly serious. For this reason, most biologists seem to be unaware of the problems inherent in PCSC, as this family of approaches is used frequently (for a few examples, see references above). Moreover, the combination of CPCA and back-projection has only recently been argued to represent the protocol of choice for the quantification of differences in multivariate trait means among multiple populations (McCoy et al. 2006; see also Rohlf and Bookstein 1987; Klingenberg 1996). The approach taken in this paper is to examine the performance of PCSC using simulated data with known variational properties, and real data from natural stickleback fish populations. I consider both situations where the interest is in estimating (co)variances and correlations within a single population, and differences in trait means among multiple populations.

Materials and methods

Simulations with a single population

The first situation simulated was a sample from a single population used to estimate the size-independent (co)variance and correlation structure among multiple morphological traits. Three simulation series were performed to explore different aspects of PCSC.

In the first series, I examined how changes in the extent to which morphological traits reflected body size-related versus size-independent variation influences the performance of PCSC. For the default simulation, I first created a ‘latent size’ variable by drawing 50 random numbers (sample size was 50 in all simulations) from a normal distribution with zero mean and variance equal to five. Next, I used the formulae given in Leon-Garcia (2008) to generate four random variables with zero mean whose expected (co)variance matrix had five in all diagonals and zero in all off-diagonals. I then added each of these size-independent and mutually uncorrelated variables to the latent size variable. The result was a collection of four ‘morphological traits’, each with an expected total variance of 10. Half of this variance was attributable to the shared latent size variable (causing correlation among the traits), while the other half reflected shape information unique to each trait. (Situations with non-zero size-independent covariance were also simulated but are not presented because they did not produce qualitatively novel insights.) The morphological traits were then size-corrected by computing the first principal component (PC1) of their (co)variance matrix, followed by BBP against PC1 (Burnaby 1966). [Throughout all simulations, extracting (C)PC1 from the correlation instead of the (co)variance matrix led to very similar results.] The resulting variables

were used to calculate the (co)variance and correlation matrix. Data generation and size correction was carried out 1,000 times to estimate the mean and standard deviation for each (co)variance and correlation matrix element.

This default protocol was then expanded such that the contribution from the latent size variable to the total trait variance (kept constant at a value of 10) spanned the full range from 0 to 0.9 in increments of 0.1. The relative proportion of size versus shape information in each trait thus increased gradually over the runs in this simulation series. I emphasize that at least some datasets included in this series would generally be considered well-suited for PCSC, as they consisted of multiple, normally distributed and intercorrelated traits, each containing substantial latent size variation as well as unique shape information.

In the second simulation series, I explored how PCSC was influenced when the magnitude of variance was heterogeneous among the four traits. For this, I modified the above default protocol to obtain datasets where the variance in the first trait increased up to threefold over five steps relative to the variance in the other three traits (kept constant at a value of 10). This was done by inflating proportionally size-related and size-independent variance in the first trait, thereby altering the traits’ covariance but not correlation matrix. In the third series, I explored the effect of variation in the number of traits under investigation. I again used the default protocol except that I increased the number of morphological traits gradually from 2 to 14 in increments of 2. This covered the number of traits analyzed in typical empirical studies. (The median trait number across the 12 studies using PCSC cited as examples in the “Introduction” was 8.5, range 3–21; only 1 study used more than 14 variables.)

Simulations with multiple populations

To examine the performance of PCSC in estimating differences in trait means among multiple groups, I conducted simulations with two populations (adding more populations did not influence the conclusions). Again, three different simulation series were performed to explore different aspects of PCSC.

Analogous to the situation with a single population, the first series explored the effect of variation in the relative proportion of size-related and size-independent trait variance. The default started with the generation of a random latent size variable with zero mean and a variance of five, and eight size-independent, uncorrelated random variables with zero mean and a variance of five. Four of the size-independent variables were added to latent size to produce the morphological data for the first population. The other four size-independent variables were used similarly to create the dataset for the second population, except that an

arbitrary constant of two was added to the first trait (the magnitude of this constant did not influence the conclusions). In other words, the full dataset included two populations with identical expected trait (co)variance structure among the four traits, but with one population shifted by 2 units along the first trait axis.

These data were then subjected to size correction by first estimating the shared within-population PC1. As the (co)variance structure was specified a priori to be equal in both populations, this was done simply by centering all traits to zero mean within each population, pooling each trait over the two populations, and extracting PC1 from the pooled traits' (co)variance matrix. Inspection of a few simulation iterations showed that this pooled within-population PC1 was always nearly identical to the common PC1 (CPC1) estimated by CPCA [Flury 1988; implemented by Phillips and Arnold 1999; (co)variance matrices were generally equal or proportional]. Using CPC1 instead of the pooled PC1 produced quantitatively very similar results. Next, the original (uncentered) trait matrices of the two populations were joined and subjected to BBP against the pooled PC1 (Burnaby 1966; Rohlf and Bookstein 1987; Klingenberg 1996; McCoy et al. 2006). The result was a new set of morphological traits with variation along the shared PC1 eliminated. These data were used to compute size-independent mean differences between the populations for each trait using univariate GLMs with population as factor (McCoy et al. 2006). Again, data generation and analysis was performed in 1,000 replicates to quantify the mean and standard deviation for the estimated trait differences.

As with a single population, I modified this default protocol such that the size-related fraction of the total variance covered the range from 0 to 0.9. Note that the two populations' PC1 became increasingly dissimilar in orientation as the size-related fraction of total variation decreased, to the extent that PCSC would no longer be recommended (McCoy et al. 2006).

The second simulation series again examined the effect of heterogeneity in total variance among traits. I here used the default with the modification that, in both populations, variance was raised gradually up to threefold in the one trait that also differed in mean value between the populations. The third series was similar to the default except that the number of total traits in each population was increased gradually from 2 to 14 (the population means always differed in only the first trait).

Data from natural populations

I first examined whether patterns identified by simulations with a single population could also be recovered when size-correcting real data with PCSC versus an approach using

an independent size metric (ISMSC). I here used measurements of five morphological traits from the three-spine stickleback (*Gasterosteus aculeatus*) fish population inhabiting Farewell Lake on Vancouver Island, British Columbia, Canada. The traits included body length, body depth, gape width, gill raker number, and gill raker length. In addition, body mass was available as an independent size metric. Sample size was 300 individuals (details on the population, traits, and measurements are provided in Bolnick 2004; Bolnick and Lau 2008; Berner et al. 2010). In a first step, each trait was divided by its mean to decouple variances from measurement scales and means (Houle 1992; Hansen and Houle 2008; log-transformation instead of mean-scaling produced very similar results in all analyses). For PCSC, I then extracted PC1 from the (co)variance matrix of the mean-scaled traits, performed BBP against PC1, and used the resulting variables to estimate (co)variances and correlations. For ISMSC, I regressed each trait against linearized (cube-root transformed; Sneath and Sokal 1973) body mass. The residuals were treated as size-free morphology (Reist 1985) and used to compute the (co)variance and correlation matrices.

To compare results from PCSC and ISMSC applied to multiple populations, measurements ($n = 300$ individuals) from the stickleback population from Gray Lake (Bolnick and Lau 2008; Berner et al. 2010) were added to the above dataset. Mean-scaling of each trait followed, but this time based on the grand mean in order to preserve differences in population means. PCSC started with common principal component analysis (CPCA; Flury 1988; Phillips and Arnold 1999) on grand mean-scaled morphology to test whether the two stickleback populations shared their first principal component. Indeed, CPCA supported (co)variance matrix proportionality just slightly better than equality (details not presented). The populations thus satisfied the criterion of a shared PC1 (CPC1), as required for size correction following McCoy et al. (2006). Variation along CPC1 was then removed by back-projection against that vector (Burnaby 1966; Rohlf and Bookstein 1987; Klingenberg 1996; McCoy et al. 2006). [Using as size vector PC1 extracted from the pooled within-group (co)variance matrix produced very similar results.] The new back-projected traits were subjected to univariate GLMs with population as factor to test and estimate population differences in means.

ISMSC was carried out by fitting a univariate GLM to each grand mean-scaled trait. Size (linearized body mass) was entered as covariate in the model, and population as a factor (Reist 1986; Darlington and Smulders 2001; Garcia-Berthou 2001). For standard length, body depth, and gape width, the size by population interaction was non-significant ($P > 0.21$). For gill raker number and length, an interaction was indicated ($P = 0.047$ and 0.02). These

interactions, however, were weak in magnitude (note the large sample size and hence high detection power), and the body size distribution was nearly identical among the two populations. I therefore considered it meaningful to analyze population differences in all five traits by the population coefficient determined by GLMs without interaction (Reist 1986). All simulations, analyses, and plotting were carried out using the R statistical language (R Development Core Team 2009; codes are provided on request).

Results

Size-independent (co)variance and correlation structure within a single population

The first series of simulations with a single population explored the effect on PCSC of variation in the proportion of size-related versus size-independent variance in the traits. These simulations revealed that the (co)variance and correlation structure of the size-corrected traits was systematically biased: true trait variance was underestimated and artificial covariance and correlation among the traits was introduced (Fig. 1a). Bias arising from PCSC was also evident in the trait loadings on PC2–PC4 extracted from the original variables (or, equivalent, loadings on PC1 extracted from PC1-residuals or back-projected traits). For instance, trait loadings on PC2 during the last iteration with the default settings were 0.86, −0.21, −0.25, and −0.4,

indicating negative covariance between the first and the other traits.

Intriguingly, bias was, over most of the parameter range explored, independent from the extent to which the traits reflected size variation as opposed to trait-specific shape variation. Only when the fraction of size-related variation approached zero, the artificial trait covariances and correlations tended to disappear, but the underestimation of trait variance became even stronger. In other words, changes in the proportion of size-related variance in the traits affected mainly the fraction of the total variance captured by PC1; the nature and relative magnitude of bias in the (co)variance and correlation structure in dimensions orthogonal to PC1 remained constant.

The second simulation series showed that bias in the (co)variance and correlation structure increased dramatically when variance was heterogeneous among traits (Fig. 1b). For instance, over the range explored, the underestimation of variance in the first trait increased from 25 to 80%, and the negative correlation between the first two traits increased twofold from −0.32 to −0.61.

The third simulation series demonstrated that the number of traits under study strongly influenced the magnitude of bias in (co)variance and correlation structure (Fig. 1c). With only two traits, true size-independent trait variance was underestimated by 50%, and a perfect ($r = -1$) negative artificial correlation emerged. Bias tapered off, but did not disappear, as the number of traits was raised up to 14. All these simulation series produced quantitatively

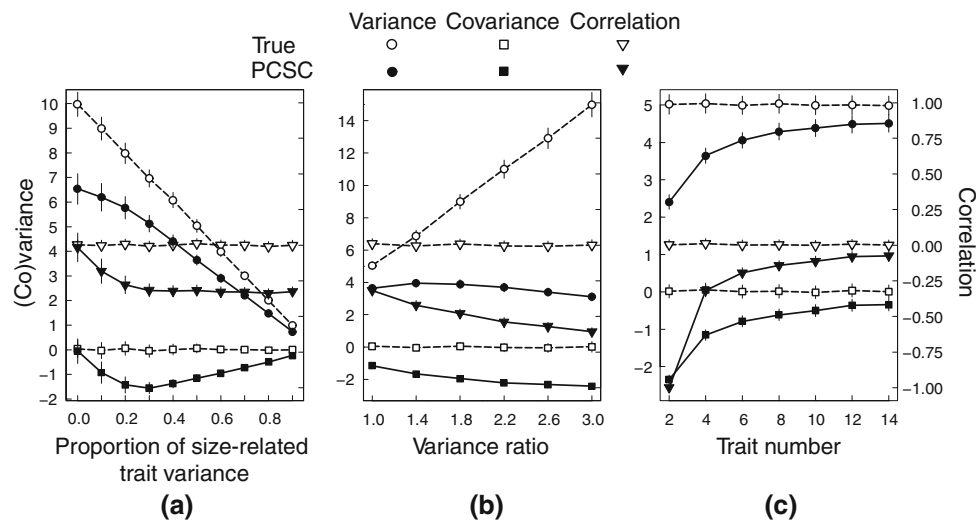


Fig. 1 Size-independent (co)variance and correlation structure among simulated traits. Shown is the variance in the first trait, and the covariance between the first and second trait, both referring to the left axis. The correlation between the first and second trait refers to the right axis, same scale for (a), (b), and (c). Open symbols indicate the true size-independent values realized in the simulations, filled symbols give their estimates after PCSC (BBP against PC1), both

averaged over 1,000 replicates (error bars are 0.25 standard deviation). Illustrated are the effects of **a** a gradual increase in the relative proportion of size-related variance in all traits, **b** an increase in the total variance in the first trait relative to the others (expressed as the expected ratio of the variance in the first and second trait), and **c** variation in the number of traits under investigation

similar results when repeated with PC1 extracted from the correlation instead of the (co)variance matrix (not presented).

In line with the findings from simulated data, qualitatively different results were obtained when variation in and associations among morphological traits in the Farewell lake stickleback population were estimated with PCSC versus ISMSC (Table 1). For instance, compared to ISMSC, PCSC estimated substantially lower variance in gill raker length but returned strong negative covariance and correlation between gill raker length and three other traits (Fig. 2 provides a bivariate illustration). The discrepancy between the methods was also evident when interpreting trait loadings (Table 1).

Size-independent differences in trait means among multiple populations

PCSC also led to biased estimates of the difference in trait means between the two simulated populations. The true difference in the first trait was greatly underestimated, and an artificial difference of opposed sign in the mean of the other traits was generated (Fig. 3a). As with a single population, this artifact was independent from the proportion to which the traits reflected size-related versus size-independent variance, except when this proportion was near zero. Quantitatively similar results were obtained when PC1 was extracted from the pooled correlation instead of the (co)variance matrix (details not presented).

Bias was greatly increased by heterogeneity among traits in the magnitude of total variance. For instance,

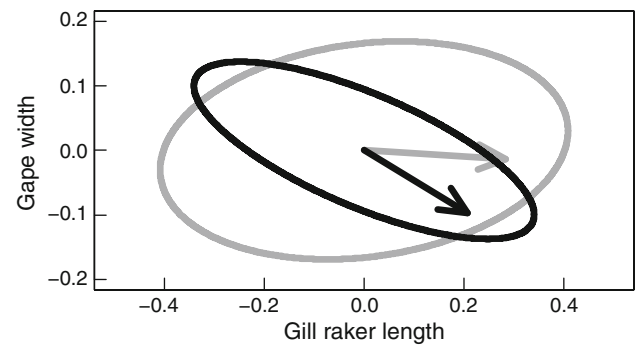


Fig. 2 Trait (co)variances and differences in population means obtained by applying ISMSC (plotted in gray) and PCSC (black) methodology to real threespine stickleback (*Gasterosteus aculeatus*) data (mean-scaled before size correction; see text for details). The resulting patterns are displayed for gill raker length and gape width only, although five total traits were analyzed. Shown are 95% confidence ellipses (centered to the same origin) visualizing the trait (co)variance structure within the Farewell Lake population. The arrows indicate the trajectory of divergence from the Farewell Lake to the Gray Lake bivariate population mean

underestimation of the mean difference in the first trait increased from 24 to 65% over the simulated parameter range (Fig. 3b). Again, bias was strongest with only two traits and decreased in magnitude as trait space dimensionality increased (Fig. 3c).

Not surprisingly, shifts in population means between the Farewell Lake and Gray Lake stickleback population differed substantially when estimated after PCSC as opposed to ISMSC (Table 2). Notably, the former approach indicated a lower population difference in gill raker length, but

Table 1 Size-independent (co)variance and correlation structure among five morphological traits for the Farewell Lake threespine stickleback (*Gasterosteus aculeatus*) population, as estimated after PCSC (BBP against PC1) and ISMSC (regression of each trait against linearized body mass)

	Body length	Body depth	Gape width	Gill raker number	Gill raker length	Loadings on PC1
PCSC						
Body length	2.11	0.70	0.18	0.09	-0.74	-0.225
Body depth	1.59	2.42	0.19	0.11	-0.76	-0.248
Gape width	0.46	0.51	3.11	-0.04	-0.73	-0.262
Gill raker number	0.25	0.35	-0.13	3.93	-0.01	-0.009
Gill raker length	-4.73	-5.20	-5.62	-0.08	19.20	0.905
ISMSC						
Body length	0.67	-0.06	0.06	-0.01	0.04	0.006
Body depth	-0.04	0.58	0.03	0.04	0.01	0.002
Gape width	0.11	0.05	4.70	-0.14	0.18	0.089
Gill raker number	-0.02	0.06	-0.62	3.98	-0.08	-0.035
Gill raker length	0.16	0.06	2.03	-0.79	27.56	0.995

Prior to size correction, all traits were divided by their mean to decouple variances from means and measurement scales. Variances and covariances (multiplied by 1,000) are given in the lower semimatrix, correlations in bold in the upper semimatrix. The last column presents the trait loadings on PC1 extracted from the size-corrected data. Note the difference between the methods in the magnitude of gill raker length variance, and in the (co)variance and correlation between this trait and the other traits. For a bivariate illustration, see Fig. 2

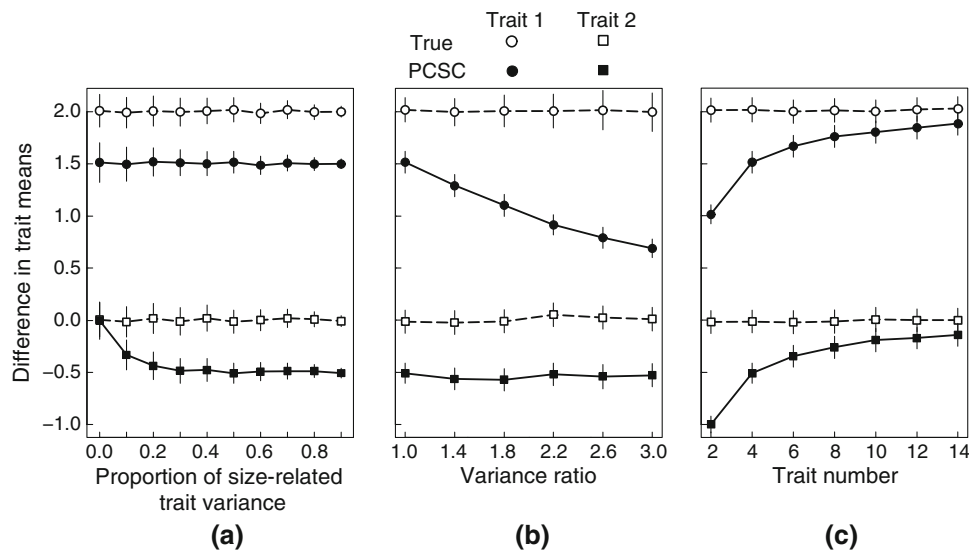


Fig. 3 Size-independent population mean differences in simulated traits. Shown is the mean shift in the first and second trait. *Open symbols* indicate the true size-independent values realized in the simulations, *filled symbols* give their estimates after PCSC (BBP against PC1 extracted from the pooled within-group (co)variance matrix), both averaged over 1,000 replicates (*error bars* are 0.25

standard deviation). The three panels illustrate the effects of **a** a gradual increase in the relative proportion of size-related variance in all traits, **b** an increase in the total variance in the first trait relative to the other traits within both populations (expressed as the expected ratio of the variance in the first and second trait), and **c** variation in the number of traits under investigation

Table 2 Difference in the mean of five morphological traits between the Farewell Lake and the Gray Lake stickleback populations, as estimated after PCSC and ISMSC

	PCSC			ISMSC		
	Mean Farewell	Mean Gray	Difference	Mean Farewell	Mean Gray	Difference
Body length	0.163	0.155	−0.009	0.974	1.026	0.052
Body depth	0.093	0.009	−0.084	1.008	0.992	−0.016
Gape width	−0.106	−0.204	−0.097	1.006	0.994	−0.013
Gill raker number	1.055	1.027	−0.028	1.016	0.984	−0.031
Gill raker length	−0.062	0.145	0.207	0.858	1.142	0.284

Prior to size correction, the traits were divided by the grand mean to decouple variances from means and measurement scales. For PCSC, the shared PC1 among populations was determined by common principal component analysis (CPCA). Population means and their difference (Gray – Farewell) were then calculated from the variables obtained by BBP of the traits against CPC1. Mean differences were tested for significance by subjecting these variables to univariate GLMs with population as factor (ANOVA). For ISMSC, the corresponding population differences were estimated (and tested) by the population coefficients from univariate GLMs with size (linearized body mass) as covariate and population as factor (ANCOVA). Means were calculated using intercept, slope (at grand mean size), and population coefficients. All population differences estimated by both methods are significant (all $P < 0.0198$). Note, however, that the (positive) gill raker length difference is substantially lower when using PCSC as opposed to ISMSC, and that all other differences are constrained to (strongly) negative values. For a bivariate illustration, see Fig. 2

stronger differences of opposed sign in the other traits (visualized for two traits in Fig. 2).

Discussion

Examining patterns of multivariate trait (co)variation within a single population, and differences in trait means among populations, often requires size correction. This is generally done either by using procedures that partial out trait variation associated with a size metric mathematically

independent from the traits under investigation (approaches here subsumed under ISMSC), or by using procedures that partial out trait variation associated with a compound size metric computed from the focal traits themselves, typically the first (common) principal component (approaches subsumed under PCSC). The goal of this study was to examine the validity of the latter family of approaches. As revealed by simulation, PCSC leads to systematically biased estimates of trait (co)variance and correlation within a population, and of mean differences among multiple populations. Surprisingly, the magnitude of bias proves essentially

independent from the strength of correlation between the traits and the latent size variable (and hence between PC1 and the latent size variable as well); artifacts emerge even when all morphological variables are tightly size-related, i.e., where the method is expected to make most sense (e.g., Rohlf and Bookstein 1987). The simulations have also shown that bias is particularly severe when there are substantial differences in the magnitude of variance among traits—a situation expected with real data, and when the number of traits under study is low.

These observed patterns can be attributed to two factors inherent in PCSC. The first is that the orientation of PC1 (or CPC1) may be influenced by idiosyncratic patterns of variance in and (co)variance among traits. Especially with few traits and heterogeneity in variance, one or a few traits can have a disproportionately strong influence on the orientation of PC1, irrespective of how well it reflects latent size. PC1 will then confound size and shape variation (Sprent 1972; Pimentel 1979; Reist 1985; Rohlf and Bookstein 1987). This factor (alone), however, is not the main problem in PCSC, given that artifacts arise even when the proportion of size-related variance in all traits approaches one and (C)PC1 therefore estimates latent size nearly perfectly (Figs. 1a and 2a). The main issue is orthogonalization. (C)PCA identifies vectors sequentially capturing maximal variance in trait space with the constraint that all vectors are orthogonal to each other (rotation of the coordinate system). This necessarily renders the underlying traits non-independent. In other words, even when (C)PC1 captures latent size accurately, PC2–PC n will still be driven by idiosyncratic patterns of (co)variance among the constituent traits. One cannot therefore assume that PCs necessarily capture biologically real (co)variance structure among the underlying variables. The discrepancy between patterns of trait (co)variation identified by (C)PCA and biological patterns of trait (co)variation, and how this discrepancy translates to artifacts in PCSC, is explained in more detail in Fig. 4.

The finding that PCSC produces artifacts has a serious implication: patterns of trait (co)variation within and between populations that have been interpreted biologically in a great number of evolutionary and ecological studies are biased to some extent, or in the worst case, lack a biological basis altogether. This point is illustrated by the analysis of stickleback data with PCSC versus an ISMSC approach. For the Farewell Lake population, strong negative covariance and correlation between gill raker length and body length, body depth, and gape width was indicated after PCSC (Table 1; Fig. 2). At least some of these associations make sense in an ecological context and hence appear to reflect correlational adaptive diversification within the population. In particular, the combination of long gill rakers with a shallow body and a narrow gape

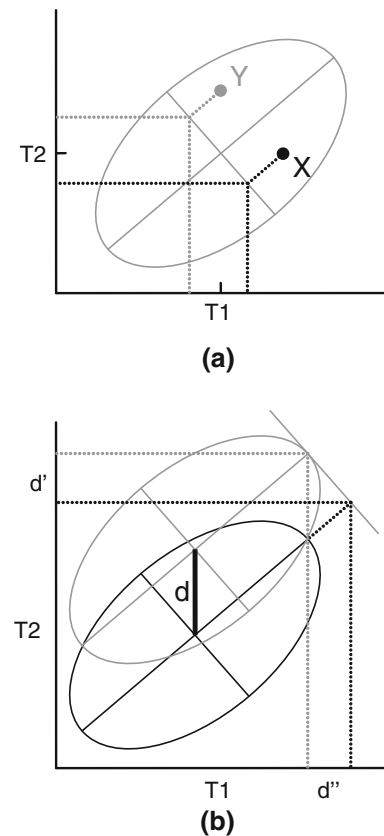


Fig. 4 Artificial trait (co)variation arising from orthogonalization in PCSC, visualized for simple bivariate situations where **a** the size-independent (co)variance structure within a single population is examined and **b** the difference in trait means between two populations is quantified. **a** The two individuals X and Y are members of a population whose (co)variance structure is displayed as gray ellipse. We assume that the traits $T1$ and $T2$ are correlated with body size; hence PC1 (the long axis of the ellipse) would be regarded as size axis, and the orthogonal short axis (PC2) as shape axis. X deviates positively from the population centroid by an arbitrary unit along $T1$, while Y deviates from the centroid by the same magnitude along $T2$. During PCSC, X and Y are projected (dotted lines) on trait space orthogonal to PC1 (in the bivariate case simply a projection on PC2), and back-projected to the original traits. After this procedure, the positive displacement of X from the population mean on $T1$ (indicated by the tick on the axis) is biased downward, while an artificial negative displacement on $T2$ is introduced. An analogous pattern arises for Y . The traits thus covary negatively after PCSC. The key point is that this will *always* be true as long as there is any variation orthogonal to PC1, no matter whether real size-independent covariance between the traits exists. Also, $T1$ and $T2$ will always have loadings of opposed sign on PC2, again indicating negative covariance that may or may not have a biological basis. **b** Two populations, displayed as black and gray ellipses, show identical (co)variance structure but differ by the magnitude d along $T2$ (mean difference along $T1$ is zero). Analogous to **(a)**, PCSC by back-projection (dotted lines) will bias downward the difference between the population means on $T2$ (d'), and introduce an artificial mean difference of opposed sign on $T1$ (d'')

should facilitate foraging on limnetic prey resources, whereas the opposite trait value combination should promote the efficient capture of benthic prey (e.g., Robinson

and Wilson 1994; Robinson and Schluter 2000). This scenario seems even more plausible as competition-driven disruptive selection on limnetic versus benthic resource exploitation within lake-dwelling stickleback populations is frequent (Bolnick and Lau 2008). We might thus hypothesize that traits related to resource use are genetically coupled in stickleback. Indeed, an analysis of the genetic (co)variance structure among the same five morphological traits in a lacustrine stickleback population from the same region (Vancouver Island) (Schluter 1996) estimated a perfect negative genetic correlation ($r = -1$) between gill raker length and gape width (see Table 1 in that paper). Schluter's (1996) study, however, used PCSC (PC1 residuals). As the present analysis of stickleback traits using ISMSC methodology indicates, this negative correlation between gill raker length and gape width (and other traits) is an artifact. If anything, these traits appear weakly positively correlated (see Berner et al. 2010, for a similar conclusion from an analysis across 17 additional stickleback populations). The particularly high size-independent variance in gill raker length, however, is real and strongly drives bias in the estimation of (co)variances and correlations when using PCSC.

PCSC and ISMSC also led to very different conclusions when applied to quantify differences in trait means between the Farewell Lake and the Gray Lake stickleback populations. Relative to ISMSC, PCSC identified weaker divergence in gill raker length, but much stronger divergence of opposed sign in most of the other traits (Table 2; Fig. 2). Combined with the simulation findings, the stickleback examples highlight PCSC's potential to generate biologically misleading patterns of trait (co)variation within and among populations. This strongly questions the value of PCSC. I argue that this family of size correction methods should be abandoned in favor of ISMSC performed as univariate GLMs with an independently determined size metric as covariate.

Indeed, the linear regression and ANCOVA methods described in the "Introduction" (and applied to the stickleback data) *always* produced unbiased results when applied to the simulated datasets (analyses not presented). This is expected, however, given that in the simulations the latent size variable was directly available, and that the morphological traits were constructed such as to be linearly related to latent size. With empirical data, however, one should make sure the assumptions underlying ISMSC are satisfied (details given in Reist 1985, 1986; Darlington and Smulders 2001; Garcia-Berthou 2001). For instance, when dealing with multiple populations, the occurrence of strong body size by population interactions will preclude meaningful size correction, especially when body size distributions differ greatly among populations (Garcia-Berthou 2001; McCoy et al. 2006). Also, the relationship

between a trait and body size might be non-linear. When performing ISMSC, it is therefore recommended to test for interactions and to examine whether a curvilinear regression model fits the data better (Reist 1986). Another critical issue is the identification of a meaningful integral estimator of overall body size. In many studies, such a metric will be easy to find (e.g., linearized body mass, Sneath and Sokal 1973; or landmark-based centroid size, Zelditch et al. 2004). Situations may exist, however, where a collection of linear distance traits (and/or counts) is available only, and where treating one single trait as body size metric is undesirable. I here explore a potential ad hoc solution to this problem, mentioned briefly in Rohlf and Bookstein (1987): splitting the traits into a subset used to compute (C)PC1 as an estimator of latent size, and size-correcting the other subset of traits using that (C)PC1 as covariate.

To explore this approach, I repeated all the above simulations for a single and for multiple populations. The difference was that the set of morphological traits generated in the different simulation series served only to extract (C)PC1 scores, which were then used to size-correct four additional traits generated as in the default simulations (further details on the methods and results are given in the Appendix). To summarize the salient results for a single population, PC1 here indeed provides a reliable size metric allowing essentially unbiased estimation of size-independent within-population trait (co)variance and correlation structure, *provided* the traits used to derive PC1 reflect latent size well (Appendix, Fig. A1a). This approach is largely unaffected by heterogeneity among traits in the magnitude of variance (Appendix, Fig. A1b), but again performs better when PC1 is computed from numerous traits (Appendix, Fig. A1c). The estimation of differences in means among multiple populations using this alternative method proves unbiased across the entire simulated parameter range.

To summarize, removing body size-associated trait (co)variance from multivariate data by defining size as the first (common) principal component of the morphological variables under investigation produces artifacts. This holds for analyses of both trait (co)variation and correlation structure within single populations and differences in trait means among multiple populations, and irrespective of whether one treats PC2– n , PC1-residuals, or (C)PC1-back-projected traits as size-independent morphology. The application of these methods for size correction is poor practice and should be avoided. Instead, size-related trait (co)variation should be partialled out from morphological datasets using general linear models with an independently determined size metric as covariate. Expressing this size metric as (C)PC1 scores from a set of morphological variables can be a valid option as

long as these variables are not subjected to size correction themselves.

Acknowledgments This paper benefited greatly from comments and suggestions by F. James Rohlf, Ben Bolker, Thom DeWitt, Dan Bolnick, Dolph Schluter, Pedro Peres-Neto, and an anonymous reviewer. Dan Bolnick kindly shared his stickleback data, and Philadelphia Airport provided electricity and a rocking chair. Funding was provided by an Ambizione fellowship from the Swiss National Science Foundation (grant PZ00P3_126391/1), and by the Research Fund of the University of Basel.

References

- Albrecht GH, Gelvin BR, Hartman SE (1993) Ratios as a size adjustment in morphometrics. *Am J Phys Anthropol* 91:441–468
- Badyaev AV, Hill GE (2000) The evolution of sexual dimorphism in the house finch. I. Population divergence in morphological covariance structure. *Evolution* 54:1784–1794
- Berner D, Adams DC, Grandchamp A-C, Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *J Evol Biol* 21:1653–1665
- Berner D, Stutz WE, Bolnick DI (2010) Foraging trait (co)variances in stickleback evolve deterministically and do not predict trajectories of adaptive diversification. *Evolution* 64:2265–2277
- Bolnick DI (2004) Can intraspecific competition drive disruptive selection? An experimental test in natural populations of sticklebacks. *Evolution* 58:608–618
- Bolnick DI, Lau OL (2008) Predictable patterns of disruptive selection in stickleback in postglacial lakes. *Am Nat* 172:1–11
- Burnaby TP (1966) Growth-invariant discriminant functions and generalized distances. *Biometrics* 22:96–110
- Cotton S, Fowler K, Pomiankowski A (2004) Condition dependence of sexual ornament size and variation in the stalk-eyed fly *Cyrtodiopsis dalmanni* (Diptera : Diopsidae). *Evolution* 58:1038–1046
- Darlington RB, Smulders TV (2001) Problems with residual analysis. *Anim Behav* 62:599–602
- Flury BD (1988) Common principal components and related multivariate models. Wiley, New York
- Fox CW, Wolf JB (2006) Evolutionary genetics: concepts and case studies. Oxford University Press, New York
- Garcia-Berthou E (2001) On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance. *J Anim Ecol* 70:708–711
- Hansen TF, Houle D (2008) Measuring and comparing evolvability and constraint in multivariate characters. *J Evol Biol* 21:1201–1219
- Holtmeier CL (2001) Heterochrony, maternal effects, and phenotypic variation among sympatric pupfishes. *Evolution* 55:330–338
- Houle D (1992) Comparing evolvability and variability of quantitative traits. *Genetics* 130:195–204
- Hoverman JT, Relyea RA (2008) Temporal environmental variation and phenotypic plasticity: a mechanism underlying priority effects. *Oikos* 117:23–32
- Hulseley CD, Mims MC, Streelman JT (2007) Do constructional constraints influence cichlid craniofacial diversification? *Proc R Soc Lond B* 274:1867–1875
- Humphries JM, Bookstein FL, Chernoff B, Smith GR, Elder RL, Poss SG (1981) Multivariate discrimination by shape in relation to size. *Syst Zool* 30:291–308
- Jockusch EL (1997) Geographic variation and phenotypic plasticity of number of trunk vertebrae in slender salamanders, *Batrachoseps* (Caudata : Plethodontidae). *Evolution* 51:1966–1982
- Jolicoeur P (1963) The multivariate generalization of the allometry equation. *Biometrics* 19:497–499
- Jungers WL, Falsetti AB, Wall CE (1995) Shape, relative size, and size-adjustments in morphometrics. *Am J Phys Anthropol* 38:137–161
- Klingenberg CP (1996) Multivariate allometry. In: Marcus LF, Corti M, Loy A, Naylor GJP, Slice DE (eds) *Advances in morphometrics*. Plenum, New York, pp 23–49
- Lande R, Arnold SJ (1983) The measurement of selection on correlated characters. *Evolution* 37:1210–1226
- Leon-Garcia A (2008) Probability, statistics, and random processes for electrical engineering, 3rd edn. Pearson Education, Upper Saddle River
- Losos JB, Jackman TR, Larson A, de Queiroz K, Rodriguez-Schettino L (1998) Contingency and determinism in replicated adaptive radiations of island lizards. *Science* 279:2115–2118
- Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer, Sunderland
- McCoy MW, Bolker BM (2008) Trait-mediated interactions: influence of prey size, density and experience. *J Anim Ecol* 77:478–486
- McCoy MW, Bolker BM, Osenberg CW, Miner BG, Vonesh JR (2006) Size correction: comparing morphological traits among populations and environments. *Oecologia* 148:547–554
- McGuigan K, Franklin CE, Moritz C, Blows MW (2003) Adaptation of rainbow fish to lake and stream habitats. *Evolution* 57:104–118
- Merilä J, Björklund M (1999) Population divergence and morphometric integration in the greenfinch (*Carduelis chloris*)—evolution against the trajectory of least resistance? *J Evol Biol* 12:103–112
- Mosimann JE, James FC (1979) New statistical methods for allometry with applications to Florida red-winged blackbirds. *Evolution* 33:444–459
- Phillips PC, Arnold SJ (1999) Hierarchical comparisons of genetic variance-covariance matrices. I. Using the Flury hierarchy. *Evolution* 53:1506–1515
- Pimentel RA (1979) Morphometrics. Kendall/Hunt, Dubuque
- R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Austria
- Reist JD (1985) An empirical evaluation of several univariate methods that adjust for size variation in morphometric data. *Can J Zool* 63:1429–1439
- Reist JD (1986) An empirical evaluation of coefficients used in residual and allometric adjustment of size covariation. *Can J Zool* 64:1363–1368
- Revell LJ, Johnson MA, Schulte JA, Kolbe JJ, Losos JB (2007) A phylogenetic test for adaptive convergence in rock-dwelling lizards. *Evolution* 61:2898–2912
- Robinson BW, Schluter D (2000) Natural selection and the evolution of adaptive genetic variation in northern freshwater fishes. In: Mousseau A, Sinervo B, Endler JA (eds) *Adaptive genetic variation in the wild*. Oxford University Press, New York, pp 65–94
- Robinson BW, Wilson DS (1994) Character release and displacement in fishes—a neglected literature. *Am Nat* 144:596–627
- Rohlf FJ, Bookstein FL (1987) A comment on shearing as a method for size correction. *Syst Zool* 36:356–367
- Schluter D (1996) Adaptive radiation along genetic lines of least resistance. *Evolution* 50:1766–1774
- Schluter D, McPhail JD (1992) Ecological character displacement and speciation in sticklebacks. *Am Nat* 140:85–108

- Sneath RR, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco
- Sprent P (1972) The mathematics of size and shape. Biometrics 28:23–37
- Touchon JC, Warkentin KM (2010) Short- and long-term effects of the abiotic egg environment on viability, development and vulnerability to predators of a neotropical anuran. Funct Ecol 24:566–575
- Zelditch ML, Swiderski DL, Sheets HD, Fink WL (2004) Geometric morphometrics for biologists. Elsevier, London